applied sciences

MDPI

Article

High-Quality Data from Crowdsourcing towards the Creation of a Mexican Anti-Immigrant Speech Corpus

Alejandro Molina-Villegas [1,2], Thomas Cattin [2,3], Karina Gazca-Hernandez [4] and Edwin Aldana-Bobadilla [1,4,*]

1 CONAHCYT, Mexico City 03940, Mexico; amolina@centrogeo.edu.mx
2 Centro de Investigación en Ciencias de Información Geoespacial, Mexico City 14240, Mexico; tcattin@centrogeo.edu.mx
3 IFG Lab Centre de recherches et d'analyses géopolitiques, Université Paris 8, 93526 Saint-Denis, France
4 Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional—Unidad Tamaulipas, Ciudad Victoria, Tamaulipas 87130, Mexico; karina.gazca@cinvestav.mx
* Correspondence: edwyn.aldana@cinvestav.mx

Abstract: Currently, a significant portion of published research on online hate speech relies on existing textual corpora. However, when examining a specific context, there is a lack of preexisting datasets that include the particularities associated with various conditions (e.g., geographic and cultural). This issue is evident in the case of online anti-immigrant speech in Mexico, where available data to study this emergent and often overlooked phenomenon are scarce. In light of this situation, we propose a novel methodology wherein three domain experts annotate a certain number of texts related to the subject. We establish a precise control mechanism based on these annotations to evaluate non-expert annotators. The evaluation of the contributors is implemented in a custom annotation platform, enabling us to conduct a controlled crowdsourcing campaign and assess the reliability of the obtained data. Our results demonstrate that a combination of crowdsourced and expert data leads to iterative improvements, not only in the accuracy achieved by various machine learning classification models (reaching 0.8828) but also in the model's adaptation to the specific characteristics of hate speech in the Mexican Twittersphere context. In addition to these methodological innovations, the most significant contribution of our work is the creation of the first online Mexican anti-immigrant training corpus for machine-learning-based detection tasks.

Keywords: anti-immigrant speech; Mexican Spanish hate speech; anti-immigrant corpus

Citation: Molina-Villegas, A.; Cattin, T.; Gazca-Hernandez, K.; Aldana-Bobadilla, E. High-Quality Data from Crowdsourcing towards the Creation of a Mexican Anti-Immigrant Speech Corpus. *Appl. Sci.* **2023**, *13*, 8417. https://doi.org/10.3390/app13148417

Academic Editors: Ahmed Rafea and Julian Szymanski

Received: 20 May 2023
Revised: 16 June 2023
Accepted: 18 June 2023
Published: 21 July 2023

1. Introduction

Nowadays, high-quality data are a valuable resource because modern artificial intelligence (AI) depends entirely on them. Numerous AI applications use machine learning methods that must be adjusted by processing large amounts of good-quality data. However, accessing high-quality data is one of the greatest challenges in using machine learning algorithms, in cases such as the study of online anti-immigrant speech in Mexico, a growing issue repeatedly pointed out by the Mexican government and international organizations [1,2]. Despite having a vast amount of HS resources available from public data repositories, there are no Mexican anti-immigrant speech resources for natural language processing. The closest works that we have found are not related to Mexican Spanish anti-immigrant speech (see Section 1.2).

In this scenario, crowdsourcing is an alternative when there are no data to use in specific machine learning tasks. However, although crowdsourcing is an interesting solution, it presents serious challenges. It is difficult to control the quality of the data collected and annotated by a large group of non-experts. The algorithm's accuracy can be significantly affected by human errors and spammers, which can lead to poor data quality. The latter can ultimately cause useless models for in-production technology. According to the Harvard

Business Review, "*poor data quality is enemy number one to the widespread, profitable use of machine learning*" [3].

In this article, we construct a high-quality training corpus using a mixed approach combining expert annotation and non-expert annotation obtained with crowdsourcing to study Mexican anti-immigrant speech on Twitter. We implement crowdsourcing through a custom annotation platform where we can control many aspects of the annotation, including a quality control mechanism and the continuous improvement of a classification model.

Our main contributions are as follows: (1) the first labeled corpus with a total of 11,582 labeled texts around Mexican anti-immigrant speech (3326 classified as positive instances and 8256 classified as negative); (2) the proposal of a methodology to control the quality from non-expert annotators; (3) a set of refined guidelines crafted from the observed phenomena in real texts; (4) a better understanding of Mexican anti-immigrant speech; (5) a baseline machine learning model to automatically obtain high-volume data for further investigation in a specific geographical, temporal, and cultural context.

We start by describing the context of the study as well as the principal objectives (Section 1.1) and the relevant related work (Section 1.2). We point out the research purpose and research questions in Section 2 Then, we detail the novel methodology to guarantee high-quality data despite the source of the data in Section 3. In Section 4, we present and discuss the study's key results. Lastly, we present some final remarks in Section 5 and present the conclusions and future work in Section 6.

1.1. An Urgent Need to Build Resources for Mexican Anti-Immigrant Speech Analysis

Mexico is a country of emigration, return migration, transit migration, and, to a lesser extent, immigration. This specificity is due to its geographical position as a "bridge" between the United States, the world's largest economy, and Latin America. The emigration experience, the sheer size of the Mexican diaspora living in the United States, and the remittances that they send home have shaped a relatively empathetic public opinion and political discourse towards migrants [4].

However, in the context of increasing restrictions on legal immigration to the United States and the worsening of socioeconomic, political, and security conditions in some Latin American countries, Mexico is becoming the main gateway for increasing clandestine migratory flows. Furthermore, since the beginning of 2018, the media coverage of clandestine transit migration in Mexico has increased considerably through the "migrant caravan" phenomenon, a form of mass mobility based on visibility.

In the United States, illegal immigration has been a major national issue since the late 1990s and is one of the most polarizing factors in the political landscape [5]. In the United States, in 2017, Donald Trump started his presidential term with an ambitious program to address illegal immigration [6]. Faced with the impossibility of initiating a reform of the American immigration system, Trump's administration pressured Mexico, its primary trading partner, to stop the migratory flows passing through Mexican territory. Against all expectations, the Mexican government of Andrés Manuel López Obrador executed with great zeal the injunctions of the American government, transforming Mexico into a vast migratory buffer zone [7].

American pressure, the tightening of migration controls in Mexico, and the "migrant caravans" have made migration a "spectacle" that the Mexican public is forced to witness. According to the National Discrimination Survey (ENADIS) in 2017, only 13.0% of Mexicans surveyed were in favor of closing the border with Guatemala and deporting migrants to their origin countries [8]. Two years later, a survey by the Reforma newspaper and the Washington Post indicated that this figure was over 50.0% [9]. The increasing hostility against migrants in Mexican public opinion is an emerging phenomenon that should be studied in the Mexican media, where discourse on migration is relayed, produced, and shared.

The emergence of digital social networks and, more broadly, the growing digitization of media have accelerated and extended the dissemination of information. Moreover,

anyone with access to Web 2.0 can now generate content, express their opinions, and react to content produced by others. The inherent distance in online interactions and the varying degrees of the echo-chamber effect and competition tend to create an online debate that is more polarized than its traditional counterparts, especially on controversial and widely discussed topics such as migration [10]. Recently, several scholars have studied the emergence of anti-immigrant speech in Mexican social media [2,11–13]. As relevant as they are, these works only consider a small portion of the anti-immigrant discourse published on Mexican social networks, both in terms of the amount of data considered and the temporal coverage of the analysis. We still have very little visibility on this emerging phenomenon. We argue that it is paramount to build resources to systematically detect Mexican online anti-immigrant speech—specifically, to build a large corpus of such discourses for future analysis.

1.2. Related Work

There is growing interest in the scientific community regarding the automatic detection of online hate speech (HS), online content that spreads hate against a particular group or individual based on characteristics such as gender, origin, culture, or religious belief. Poletto et al. recently published an overview of existing research and resources on such topics [14]. Until recently, most research has focused on HS as a generic phenomenon and has been published in English, the most spoken language. However, HS is a complex object that varies in intensity, directness, targeted groups, and cultural and geographical context. In other words, HS is related to a particular temporal and territorial context. Quoting Arcila-Calderon et al. [15], undertaking a generic approach "could be a limitation because the resulting models may not be as effective, reliable, and, paradoxically, generalizable as those trained with real examples of a specific context, a specific type of hate, and a specific discriminatory category, separating and differentiating concepts, characteristics, and linguistic nuances".

Several recent analyses have been conducted on anti-immigrant or xenophobic content detection in specific cultural, temporal, and geographical areas. In English, Pitropakis et al. collected anti-immigrant tweets published in the UK, the U.S., and Canada following specific events related to immigration in these three nations [16]. Siegel et al. studied white nationalist rhetoric on Twitter during the 2016 United States presidential election and relied on a mixed approach combining a dictionary and supervised machine learning to detect racist and xenophobic speech in a dataset of more than one billion tweets [17]. Another research study was conducted in European countries at the forefront of the unfolding migratory crisis in 2015. In Italy, researchers created a corpus in Italian of offensive tweets against immigrants, Muslims, and Rom [18,19]. They used a mixed annotation process that included both experts and paid crowd workers to follow the same novel multilayered classification scheme that considered the presence of HS but also five other categories: aggressiveness, offensiveness, irony, stereotypes, and intensity [20].

Anti-immigrant speech detection in Spanish was the object of a collective event at SemEval 2019. In Task 5, participants were provided with a publicly available dataset, HatEval, that contained 1991 labeled tweets regarding immigrants [21]. To build the dataset, they relied on expert annotation using a combination of techniques to collect data from Twitter, including hateful keywords related to Spanish and Latin American contexts. Plaza-Del-Arco et al. and Hasan et al. used HatEval to train and compare several machine learning models to detect HS against immigrants [22,23]. Arcila-Calderón et al. manually created an ad hoc dataset to train deep and shallow learning models to detect anti-immigrant speech in European Spanish [15]. Based on a broad definition of HS against migrants, considering xenophobia and racism, they downloaded and filtered Twitter data with keywords assembled during an exploratory qualitative stage conducted in the Spanish Twittersphere.

Regarding specific aggressiveness detection for Latin American Spanish, the most relevant works are the MEX-A3T track at IberLEF 2019 [24], where the organizers considered

two tasks focused on authorship and aggressiveness in Mexican tweets, and the Language Model for Misogyny Detection in Latin American Spanish [25].

Inspired by the investigations described in this section and building on previous work [26], this paper presents new methodological elements to build a corpus for HS detection under certain constraints. We worked with a group of experts for the manual annotation process, had no financial resources to use crowdsourcing services such as those provided by Crowdflower, and had to build our own annotation platform. Lastly, our work stands out because we built an anti-immigrant speech corpus in a very specific temporal and geographical context not yet considered in Mexican Spanish. Moreover, our work expands the horizon of online anti-immigrant speech detection in specific linguistic, geographical, and temporal contexts by building a Mexican Spanish corpus, which is particularly relevant as migration is becoming an increasingly polarizing and politically instrumentalized topic. All these contributions together make this a novel methodology to obtain quality data for machine learning.

2. Research Purpose and Research Questions

The general objective of this work is to propose a sophisticated methodology to create a dataset of human-annotated tweets that can then be used to train machine learning models to detect online anti-immigrant speech in Mexican Spanish. This can be seen as the first stepping stone in the research project briefly presented in Section 1.1 and described in a previous communication [26]. Thus, the dataset must not only be adequate for machine learning purposes to achieve internal statistical accuracy, but it also must represent "real" data that can be used to conduct an extensive analysis of Mexican anti-immigrant online speech produced on Twitter between 2018 and 2022.

More specifically, this work aims to overcome one of the limitations of a human-labeled corpus: the significant amount of time required by expert annotators to label a large corpus of examples by hand. Since our solution involves supplementing expert annotations with non-expert annotations, we raise the following question: How does implementing a control mechanism based on the kappa coefficient value affect the quality and temporal efficiency of annotations for data compared to relying solely on expert annotators? (RQ1)

Another way to assess non-expert labeled data's reliability would be to look at their performance in the task by revising the data that they produce. Given that each of them had the same guidelines used by the experts and that they had to label a small sample containing tweets previously classified by the experts, we asked the following: Would the guidelines be enough for non-experts to recognize all categories of anti-immigrant speech identified by the experts? (RQ2) If not, what are the mislabeled categories? (RQ2A).

It is also relevant to inquire as to whether the data obtained from a mixture of experts and non-experts represent a valid input for any learning algorithm so it can serve any machine learning practitioner to train their model (RQ3).

Lastly, we expect that the annotation process can highlight those semantic aspects of the corpus that allude to Mexican anti-immigrant HS. In this regard, an arising question is as follows: How can we determine the existence of these aspects, and to what extent are they related to the context of our study? (RQ4).

3. Methodology

In Figure 1, we present the overall strategy. Four stages were designed to achieve the objectives of the research. In Stage I, we collected a large volume of raw data from Twitter. In Stage II, we focused on defining criteria to distinguish anti-immigrant discourse and exploring the kappa coefficient to discriminate annotations of quality. In Stage III, we created three data pools with helpful characteristics to integrate the data annotated in the next stage. Lastly, in Stage IV, we developed an annotation platform that integrates mechanisms for quality control and the continuous improvement of the annotated data. In the rest of this section, we detail all the stages.

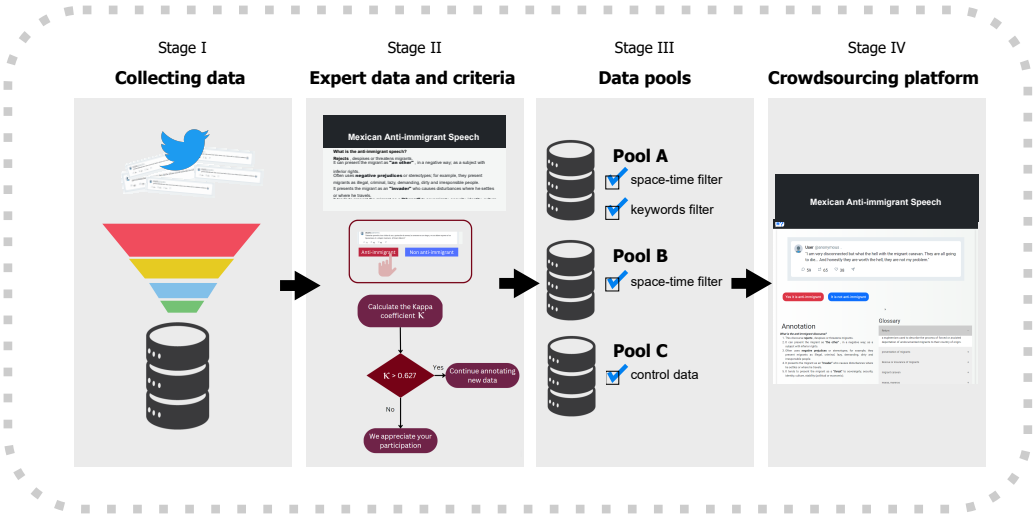

Figure 1. Overall methodology. Our proposal is a pipeline made up of a stage sequence focused on (1) collecting and filtering a comprehensive dataset containing tweets generated in Mexico from 1 September 2017 to 1 January 2021; (2) carrying out a preliminary expert annotation process on a preset set of tweets based on a sophisticated guideline; (3) creating data pools for control data and data subject to annotation, and (4) implementing a crowdsourcing annotation platform for these latest data.

3.1. Stage I: Data Collection

This stage focuses on the methods used to collect Twitter data. We start by describing the process by which the raw data were collected and filtered under the purpose and context of our research.

3.1.1. Raw Twitter Data

We first collected all the tweets available on the Autómata Geointeligente en Internet (AGEI) developed by Centrogeo from 1 September 2017 to 1 January 2021. AGEI collects georeferenced tweets published in Mexico, part of the United States, and Central America. The information includes text and metadata such as the ID, username, date, language, sometimes geolocalization, and information about the interaction. This collection has been previously used for other research related to human activity and geographical patterns [27], misogyny [28], and discrimination [13]. To obtain additional data, we retrieved tweets from the INGEOTEC text models [29]. In the end, we managed to obtain hundreds of thousands of georeferenced tweets.

3.1.2. Data Filtering

In this stage, represented in Figure 1 as data filtering, the main goal was to keep only relevant tweets for annotation. The raw data obtained from AGEI and INGEOTEC (Section 3.1.1) were processed using a two-fold filter based on spatiotemporal criteria and a preliminary set of keywords.

The georeferenced AGEI data were mapped in a Geographical Information System (GIS) and regrouped by year, three-month periods, and publication region: northern Mexican border states, southern border states, and the rest of the Mexican states. To address possible imprecision in tweet georeferencing, we applied a 20 km buffer to the regions. We discarded all data from outside these regions. The INGEOTEC data were filtered using the attribute *place* of the tweet object. Quoting Twitter's official documentation, *Place* "*indicates that the tweet is associated (but not necessarily originating from) a place*". We only kept posts associated with "Mexico". We acknowledge that some filtered data might have been produced outside Mexico.

Although the timeframe of interest was vast (almost four years), we chose to apply a temporal filter to maximize the probability of finding posts related to migration and migrants. This choice was consistent with the variations in media coverage and public discussion on migration, which tend to crystallize during episodes of "crises" related to specific events. Moments of significance were selected by the experts, who identified special events related to migration in Mexico, represented graphically in Figure 2. We focused primarily on the 2018–2019 period, the "migrant caravan" crisis, and then extended the selection to December 2022.

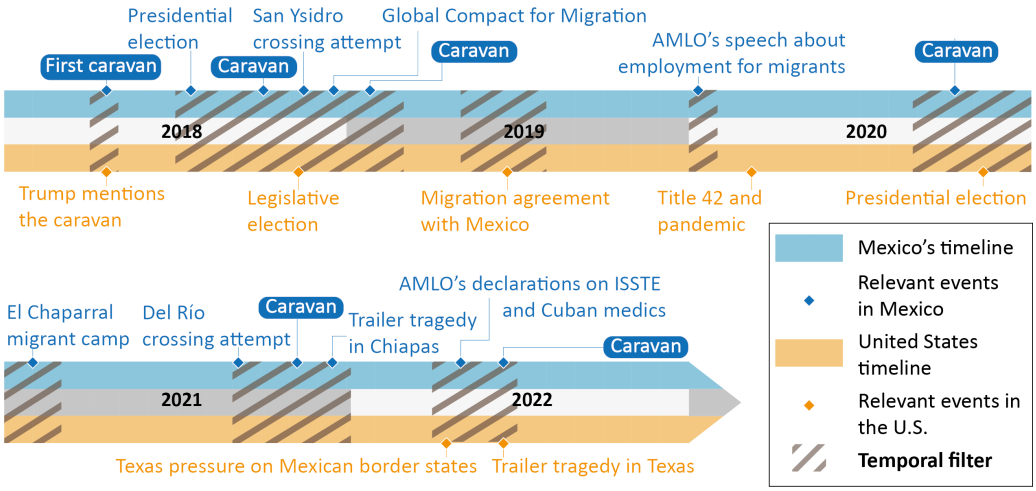

Figure 2. We divided the Twitter timeline into moments (month) around specific events occurring in Mexico and in the United States that potentially triggered online discussion around migration.

A second linguistic filter split the result of the first filter based on an ad hoc set of keywords related to migration and anti-immigrant speech in Mexico. We are aware that using keywords to retrieve data can lead to topic bias, as shown by Wiegand et al. in [30], but it also makes the annotation process much more efficient. We agreed to use neutral and explicit anti-immigrant words and hashtags. The former allowed us to collect both positive and negative instances of anti-immigrant speech but also to retrieve posts that were not, at first glance, particularly hateful towards migrants. We used the latter to maximize the probability of obtaining anti-immigrant speech. Building the dictionary was not a linear process. It required a thorough qualitative exploration of the Twitter data. During this process, we could identify content and users of interest who were especially likely to generate anti-immigrant content. This "exploratory" stage was essential to understand the specificity of Mexican online anti-immigrant speech regarding the targeted groups, mainly Central American, Caribbean, and South American migrants, and the most commonly used words and hashtags. We discarded some keywords (due to off-topic content) and added others to then obtain the selection described in Table 1.

After filtering the data collected from AGEI and INGEOTEC, we obtained a dataset of 47,343 tweets to be annotated.

3.1.3. Assembling the First Training Corpus

The qualitative exploratory process in Stage I was an opportunity to build the first instance of an anti-immigrant dataset. While exploring the Mexican online conversation on Twitter, one expert collected positive and negative instances of anti-immigrant speech based on the definition described in Section 3.2.1. We included a corpus of misogynistic posts used in previous work [28] to bolster negative instances of anti-immigrant speech. After this, we retrieved the first instance of the corpus, which was composed of 1073 anti-immigrant tweets and 4548 negative instances. Using this first dataset, we trained a convolutional neural network (CNN) for binary classification and achieved an initial accuracy rate of 0.7675 (see Table 2), as detailed in [26].

Table 1. Dates, keywords, and hashtags of interest for filtering criteria.

Filtering Criteria	Year	Description
Dates	2018	First migrant caravan, presidential election in Mexico, caravan, San Ysidro crossing attempt, global compact for migration, Trump mentions the caravan, legislative election
	2019	Migration agreement with Mexico, AMLO's speech about work for migrants, Title 42 and pandemic
	2020	Caravan, presidential election in the USA, El Chaparral camp
	2021	Del Rio crossing attempt, caravan, trailer tragedy in Chiapas
	2022	AMLO's declaration on ISSSTE and Cuban medics, caravan, trailer tragedy in Texas
Keywords	2018–2022	Migr, caravana, hondureñ, frontera, centroamerican, haitian, delincuent, mara, extranjer, refugiad, invas, mugros, indocumentad, ilegales, soberanía, nación, Trump, negro, african, venezolano, cubano, nicaraguense, guatemaltec, chapín, onu, catracho, nica, Nicoya, pinolero, sudaca, cachucos
Hashtags	2018–2022	#caravana #caravanamigrante, #mexicoprimero, #mexicoparalosmexicanos, #fueracaravanamigrante

Table 2. Accuracy of Mexican anti-immigrant classification models in the continuous improvement loop.

Training Date	Accuracy	Support
1 October 2021	0.7675	200
5 December 2022	0.8291	200
25 January 2023	0.8801	240
21 March 2023	0.8828	250

3.2. Stage II: Expert Data and Criteria

This stage revolved around the design and validation of a set of criteria to use as a guideline for the annotation process for both expert and non-expert contributors. We first describe the definition of five criteria and then present the validation process, which involved agreement measurement between three experts asked to annotate 76 selected tweets using the criteria.

3.2.1. Defining Specific Criteria for Mexican Anti-Immigrant Speech

Considering a strict definition, anti-immigrant speech conveys a hostile attitude towards people, or their descendants, who reside or settle in a territory other than their native country. Usually, individuals take an anti-immigrant stance "on behalf of the 'natives' of a country, their interests, their culture, and even their origins (or 'race')" [31]. Even if racism often fuels anti-immigrant speech, they are not equivalent. Indeed, ethnocentric political movements can advocate for immigration if migrants fit into supposed specific social, cultural, or biological criteria.

In Mexico, it is unusual to employ the term "anti-immigrant". Mexico is not an immigration country. In a report on HS against migrants [1], the Mexican government prefers to utilize the more generic expression "xenophobic speech". However, in the specific context described in Section 1.1, HS against migrants in Mexico is increasingly similar to that present in traditional countries of immigration.

Published research on online xenophobic or anti-immigrant speech tends to use a strict definition of HS, where posts must be addressed explicitly to the vulnerable category of interest and must be intentionally harmful to this category; see, for example, [18]. Because intentionality can be quite difficult to detect and we wished to consider even the most implicit types of anti-immigrant speech, we took a broader approach, arguing that

in order to be anti-immigrant, a post must (a) reference migrants, explicitly or implicitly, and (b) incite, spread, or promote negative representations or behaviors against them.

According to Ross et al., providing non-expert annotators with a general definition such as the one mentioned above does not improve the reliability of their annotation [32]. This is why we decided to provide a more sophisticated definition, more akin to a guideline, that could help the annotators recognize certain characteristics of Mexican online anti-immigrant speech when given new data. The exploratory analysis of Mexican online conversations on migration described in Section 3.1.2 proved to be essential in confronting our general, theoretical considerations with empirical data, thus grounding our work in the specific context of interest. It proved to be a necessary step, as we found new characteristics of Mexican online anti-immigrant speech previously not considered in the guideline. For example, we found many tweets that could have been categorized as "conspirational", structured around the idea that the migrant caravans were organized by individuals to politically, economically, or socially destabilize Mexico and the United States (see examples below).

We agreed upon five criteria, corresponding to five categories that accurately but concisely describe the heterogeneity of anti-immigrant speech observed in the Mexican Twittersphere. It is not uncommon for a tweet to fall under several criteria. The five elected criteria are given below, each with a tweet from our dataset to illustrate it.

1. Anti-immigrant speech attacks, despises, or threatens migrant individuals. As illustrated by the tweet below, this type of speech often uses anti-immigrant and racist slurs:
   "*Deja de ladrar honduchango*".
   Translation: "Stop barking Honduran monkey".
2. Anti-immigrant speech revolves around a negative otherness based on assumed differences, which, in turn, provokes suspicion and justifies excluding or diminishing liberties and rights. In the next tweet, the difference between natives and foreigners justifies not helping migrants:
   "*Ok, ojo no tengo nada en contra de los migrantes, pero y los mexicanos? Ella debería de estar primero revisando de que los mexicanos tengan empleo, esa debe de ser su prioridad!!!*"
   Translation: "Ok, look, I'm not against migrants, but what about Mexicans? She should first think about Mexicans being employed. That should be her priority!".
3. Anti-immigrant speech uses negative stereotypes to qualify migrants. In our dataset, the "qualities" attributed most often to migrants were "illegals", "lazy", "demanding", "poor", "dirty", "irresponsible", and "criminals", as illustrated in the tweet below:
   "*La redacción completa es: refugio para migrantes maras salva truchas y todos los delincuentes que uyen de otros países. unos genios!*"
   Translation: "The entire story is to give refuge to Mara Salvatrucha migrants and all criminals fleeing from other countries. Such geniuses!".
4. Anti-immigrant speech often presents migrants as invaders. This idea is built on the representation of a geographical space being violently disrupted:
   "*su ayuda solo favorece y promueve que nuevas caravanas sigan llegando trastornando enormemente la vida cotidiana de las ciudades fronterizas de México*"
   Translation: "Your help only encourages and fosters new caravans to arrive while causing havoc in the daily life of Mexican border towns".
5. Anti-immigrant speech also portrays migrants as tangible or symbolic threats to larger issues such as the economy, security, identity, or even sovereignty:
   "*#CaravanaMigrantes "el derecho al respeto ajeno es la paz" respenten nuestra soberania. #MexicoparalosMexicanos*".
   Translation: "#Migrantscaravan "Peace is the right to foreign respect" respect our sovereignty #MexicoforMexicans".

3.2.2. Agreement Metrics

To evaluate the quality and operability of the five criteria, we asked three experts to classify a small dataset of tweets retrieved during the previous stage. The group of experts was integrated with two Ph.D. candidates who specialized in migration issues and one professor who specialized in border studies and migration in the Mexican southern region. Since our goal was to test the limits of the guideline's usefulness to classify potentially problematic and ambiguous tweets, we did not build the dataset randomly. We manually selected a small amount of clearly pro-migrant and neutral tweets, a large amount of anti-immigrant tweets from the sample at hand, and some "problematic" examples. The latter were difficult to classify because of a lack of context (irony, sarcasm), a mixture of forms and ideas characteristic of both pro- and anti-immigrant speech, or the group or individual targeted was not clearly identifiable. Lastly, we selected 76 tweets. The experts had to independently classify the dataset using the criteria.

We then obtained the experts' feedback regarding the criteria and specific tweets where no agreement was reached. As expected, some anti-immigrant tweets were not easily identifiable. We agreed on 73 tweets, three unclassified. Inter-rater agreement was needed to assess the reliability of the classifications made by the aforementioned experts. This was important to ensure that we were working with high-quality control data. We established a baseline rate to evaluate the agreement between expert annotators with a statistical metric named the kappa coefficient value. Its formula is presented in Equation (1).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

In Equation (1), we see the kappa coefficient, where P_o is the observed proportion of agreement between annotators and P_e is the expected proportion of agreement between annotators due to luck.

As a first result, we obtained a global inter-annotator agreement coefficient of 0.726. Moreover, we calculated the peer-to-peer agreement to have a better idea of the boundaries in the inter-agreement values among experts. The peer-to-peer agreement among the three experts is shown in Figure 3. This result proves an important aspect of anti-immigrant HS data: we cannot expect perfect agreement from individuals, not even experts, during manual annotation. For this reason, we lowered our expectations about the possible coefficient values that non-experts could obtain, which will be described in the following sections.

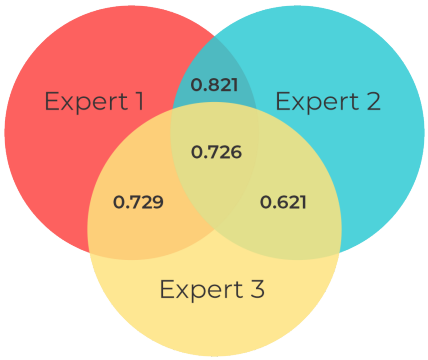

Figure 3. Expert inter-annotator agreement for control data pool.

3.3. Stage III: Data Pools

This stage revolved around creating three data pools: a first pool (A type) containing texts with a high chance of belonging to the positive class; a second pool (B type) representing the negative class and some noise; a third pool (C type) composed of a modest quantity of experts' labeled data, described in Section 3.2.2. Then, we discuss the trade-offs in the annotation process, including time observation, batch size determination, and the kappa

coefficient adjustment. We aim to highlight the need to balance practicality and agreement while leveraging the insights of competent non-experts.

3.3.1. Splitting into Annotation Data Pools

The main process to compose Pool A was data filtering. In this pool, we tried to concentrate on relevant tweets for annotation. The raw data obtained from AGEI (Section 3.1.1) were filtered using the spatiotemporal criteria defined by dates of interest (Table 1). Note that experts selected the days of interest. They recognized special events related to migrants traversing Mexico. Thus, the spatiotemporal filtering retained only tweets published on such days in the Mexican territory and part of its northern and southern borders. A second filter split the result of the first filter based on a set of keywords related to the subject. The split created the first two pools: Pool A was composed of alleged anti-immigrant data, and Pool B was composed of alleged non-anti-immigrant data. However, data from Pool A and Pool B had to be manually annotated, but the odds of obtaining data relevant to the objective of research from Pool A were greater than for Pool B. With this in mind, we later combined both Pools A and B conveniently to obtain sufficient data for machine learning.

Pool C was a qualitative representation of the 76 tweets that the experts annotated. Pool C was particularly significant for our approach because we used it to identify the ability of non-expert annotators to reproduce the criteria used by the experts. Pool C was composed of a sample of expert-labeled text. Using the agreement score and this pool, we could estimate whether a non-expert was applying the annotation guideline well. However, a relevant question arose: How many control texts do we need to have confidence in the criteria for assessment? A simple answer to this question would be as many as possible (76). However, adding more control texts implies more time than that which annotators use to label data not used for machine learning. Thus, we sought to obtain the best trade-off between time and high-quality annotated data. In the following section, we describe the experiments to determine the best trade-off between time and quality.

3.3.2. Annotation Batches with Trade-Off in Time and Quality

To consider the time and quality trade-off, we asked 32 users to annotate the control data. Our objective was to observe the time required by the users to complete the task. They had to determine whether a tweet contained anti-immigrant speech or not. The average time to classify a tweet was 12.583 s, as shown in Figure 4.

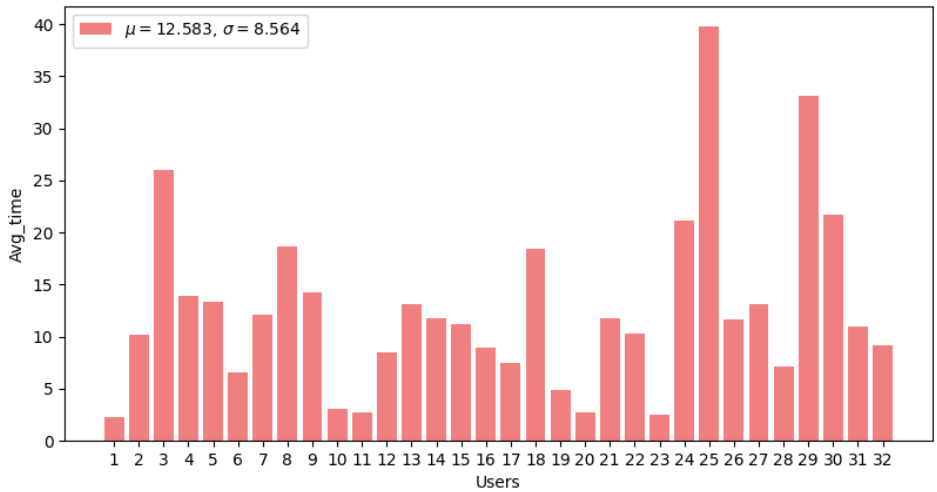

Figure 4. Average time that the 32 users took to categorize tweets as anti-immigrant or not, where the x-axis represents the users and the y-axis represents the average time that each user took to choose the category.

Based on the average time obtained, we conducted a simulation experiment to determine the appropriate batch size for annotation. We wanted to ensure that the non-expert

users did not spend more than 20 min annotating. Considering that the average time spent annotating the control data was 6.29 min, we selected a real annotation batch of 50 tweets, since this would require approximately 10.49 min for the user to annotate, resulting in a total of 17.18 min. This fell within the desired maximum timeframe of 20 min per annotation batch.

For users to annotate the 50 "real" unlabeled tweets, we first needed to evaluate the user agreement coefficient. However, before proceeding, we conducted an experiment to establish the kappa coefficient for comparison with the non-expert annotators. Our experiment aimed to assess the impact of incorrect categorizations on control data. We allowed the annotators to make up to 9 errors in categorizing the tweets. However, the resulting kappa coefficient was 0.448, indicating only moderate agreement according to the interpretation of kappa. This agreement level is not highly recommended.

Considering this, we turned to the kappa coefficient among the experts in the field. Even the experts did not achieve a perfect kappa coefficient, with the highest agreement value being 0.821. Therefore, we decided to adjust our kappa coefficient base. Instead of using 1 as the perfect score, we used 0.821. Mathematically, we subtracted 0.821 from 1, resulting in a difference of 0.179. We added this difference to the kappa coefficient obtained by the users, creating what we now refer to as the adjusted kappa coefficient.

Allowing users to make 9 errors produced an adjusted kappa coefficient of 0.627. In terms of kappa interpretation, this indicated substantial agreement.

By adjusting the kappa coefficient threshold, we acknowledged the unrealistic expectation of achieving perfect agreement, even among experts. This adjustment allowed for a more practical and feasible assessment of agreement while still maintaining a reasonable level of quality in the annotations. It recognized that a high level of agreement, although not perfect, could still provide valuable and meaningful insights into the task at hand.

Lowering the kappa coefficient threshold in this context ensured that we did not exclude annotations from competent and knowledgeable non-experts who could provide valuable perspectives and insights, even if their judgments may differ slightly from the experts. It allowed us to leverage the collective wisdom of a diverse group of annotators while maintaining a level of agreement that was substantial and reliable for the specific task or domain.

In conclusion, our annotation process involved trade-offs regarding time efficiency and agreement among annotators. We determined an appropriate batch size to ensure that users spent no more than 20 min annotating based on observations and simulation experiments. While initial agreement levels were moderate, we adjusted the kappa coefficient to account for realistic limitations and achieve substantial agreement. This approach allowed us to incorporate valuable perspectives from competent non-expert users while maintaining a practical assessment of annotation quality.

3.4. Stage IV: Crowdsourcing Platform

This section focuses on the need for a crowdsourcing annotation platform. We explore the admin and user annotator panels that comprise the platform and highlight their functionalities. Additionally, we describe the strategy used to continuously improve the model accuracy, mentioning how the model utilizes the annotated data collected from the crowdsourcing annotation platform.

3.4.1. Designing the Crowdsourcing Platform

Traditionally, data annotation is one of the most time-consuming tasks to obtain a language model due to its manual process. A set of annotators must observe texts and classify them manually, usually in a spreadsheet [26,33]. Moreover, as the number of documents grows, this traditional method becomes impractical and inefficient.

An annotation platform becomes necessary to address these challenges. It provides a user-friendly interface to annotate data by following a well-defined methodology with quality control mechanisms to improve the overall performance of the language model.

Moreover, it allows many annotators to work simultaneously on different parts of the dataset, regardless of the location of the annotators, translating into greater productivity and a significant reduction in the time required.

At the outset, we explored an annotation platform called Doccano, which allows users to annotate data at no cost, as the authors in [15] did. However, our specific requirements necessitated certain features not readily available on the platform. For instance, we aimed to ensure the quality of the annotated data by evaluating the users' performance in the control pool and comparing it to the inter-annotator agreement obtained in Section 3.3.2 (0.627). Additionally, we needed a mechanism where each user could annotate at the same time a different portion of the dataset by establishing the number of documents allocated for annotation and the inter-annotator coefficient in the platform, ensuring that only if the user passed the set kappa coefficient would they be able to annotate the following 50 unlabeled documents from the dataset.

As a result of these requirements and considerations, we decided to develop a custom annotation platform from scratch. This approach allowed us to tailor the platform to meet the needs of our research project. Furthermore, we recognized the value of contributing to the research community by making the platform open source—it can be found in the GitHub repository—enabling other researchers to benefit from and customize it according to their specific annotation requirements, as well as obtaining the data (on demand): https://github.com/Kgazcah/Annotation-platform, accessed on 19 June 2023 .

3.4.2. User Interface

The platform offers users the possibility to upload the dataset that is to be annotated. Additionally, the control data, previously annotated by experts, can be uploaded. It is also possible to configure the inter-annotator kappa agreement coefficient, which is a measure to quantify the annotated data's quality.

The platform allows the configuration of more parameters, such as the size of the data chunks that users will annotate after passing the predefined threshold. The admin panel displays all the set configurations, including the total number of documents in the dataset and the progress of annotations completed.

Our platform also offers the flexibility to edit these parameters, delete datasets if necessary, and obtain a shareable link to invite multiple users to participate. Moreover, administrators can view the labeled documents without downloading them. However, it is possible, as shown in Figure 5.

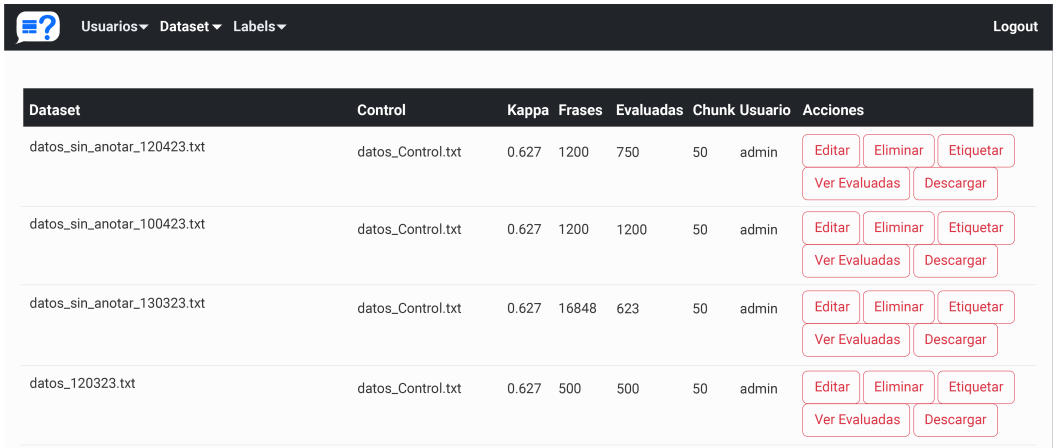

Figure 5. The admin panel of the annotation platform: a comprehensive interface showcasing configurable parameters and powerful functionalities, including editing, deleting, obtaining links, accessing evaluated documents, and downloading them for seamless annotation management.

Once all the necessary configurations are implemented, the link is shared with the users, who will then be presented with the annotation interface, as shown in Figure 6.

In this interface, annotators have to classify a tweet as either anti-immigrant or not by using the corresponding buttons provided. Moreover, our platform allows the customization of shortcut keys for each class. In our project, we utilized the letter "a" as a shortcut to classify a tweet as anti-immigrant and the letter "n" for non-anti-immigrant tweets.

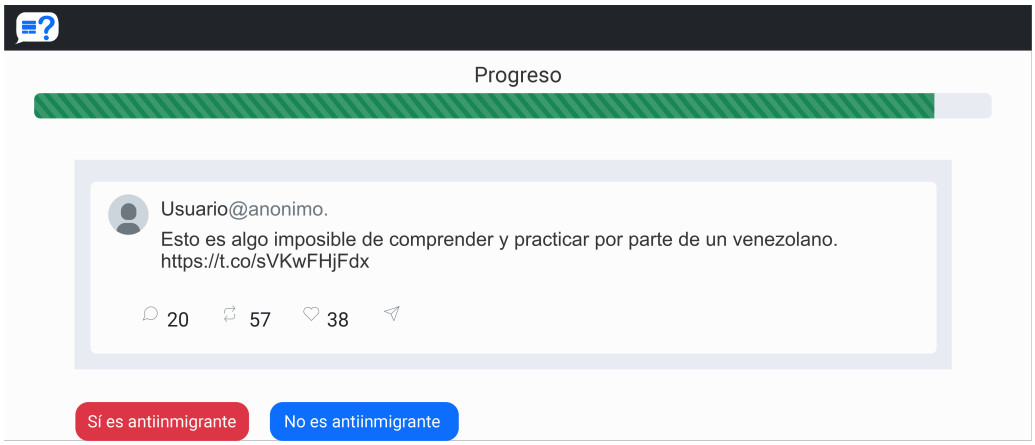

Figure 6. A snapshot of the platform's intuitive interface, displaying a tweet in Spanish and the buttons to determine whether the tweet exhibits anti-immigrant speech. Translation: "*This is something impossible for a Venezuelan to understand and practice*".

Initially, the user must read a synthesized version of the guideline provided by the experts, thus ensuring that the annotation process is based on a definition of Mexican online anti-immigrant speech. The user will then annotate the control data pool, with the 30 documents pre-annotated by the experts. If the annotator passes the predetermined kappa coefficient, they will annotate 50 unlabeled documents from the dataset. The progress bar will indicate the completion status as the annotator progresses. Finally, the button "send" will appear to complete the process in less than 20 min, as analyzed in Section 3.3.2.

3.4.3. Continuous Improvement

During the annotation, we used a human-in-the-loop strategy to achieve better results. First, we sorted all the available data using the hate score produced from our first classification model (described in previous work [26]). After some time, we retrained the classification model using more data obtained from the annotation platform and increasing the model accuracy constantly, as presented in Table 2. However, continuous improvement is not only relevant in the context of classification models; it has also proven highly important for research. From this strategy, we learned that some words were biasing the results. For instance, the model training of 25 January 2023 produced many false positives, including the word "negro" (*black*). In consequence, we explored the data to observe which other particular words were confusing the model. We used the identified confusing words (e.g., "delincuente", "negr") to search and include examples of texts using these words but having the negative class. As a result, after retraining a classification model with better data, the produced model increased its accuracy and achieved the better sorting of the remaining data and a better understanding of the pragmatics in anti-immigrant speech. It is worth mentioning that the bias detected by words such as "delincuente" (criminal) or "negro" (black) is not a bias associated with the quality of the data but rather a problem inherent to the classifier itself. This is important because we discovered that the final models could conveniently generalize the differences in the anti-immigrant speech, which had a good impression on the experts.

4. Results and Discussion

This section focuses on the research's most relevant results, oriented towards assessing and discussing the quality of the annotation protocol and the labeled data. To answer the

research questions previously posed and discuss the results, we explore metrics and techniques to measure annotation reliability through machine learning algorithms. Additionally, we describe an interpretability analysis method that enhances our understanding of model predictions and decision-making processes, thus underlining some relevant aspects of the training corpus.

4.1. Result 1: Assessing the Performance of the Crowdsourced Annotation Protocol

To address RQ1, we conducted two experiments to evaluate the performance of the crowdsourced annotation protocol. The first experiment focused on obtaining high-quality data annotations by following the annotation protocol outlined by experts in the field, mentioned in Section 3.2, and involving as many non-expert annotators as possible. The second experiment aimed to analyze the time that non-expert annotators took compared to expert annotators, and the role of the crowdsourced annotation platform in facilitating the process.

For the first experiment, 47 contributors participated in the crowdsourced annotation process, as described in Section 3.4. Not all contributors were able to pass the control mechanism described in 3.3.2. Upon presenting the contributors with control data (from Pool C), it was found that 15 of them did not meet the established threshold of the coefficient of agreement (0.627), as shown in Figure 7, which illustrates the distribution of kappa coefficients obtained from the non-expert users. The average kappa coefficient for these users was 0.444.

Figure 8 displays the kappa coefficient distribution for non-expert users who did pass the threshold. These users achieved an average kappa value of 0.796 with a standard deviation of 0.105, indicating good-quality annotations for the "real" data.

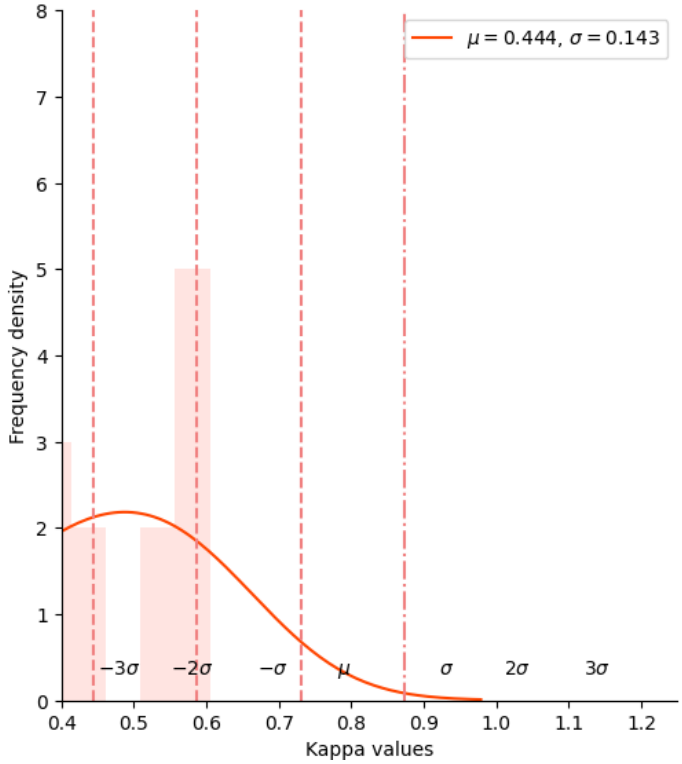

Figure 7. Kappa coefficient distribution for users that did not pass the set threshold, obtained from m non-experts (where $m = 15$) through the platform with data quality control protocols.

The contrast in the kappa coefficients distribution between the non-expert users who passed the threshold and those who did not pass highlights an aspect of crowdsourced annotation: not all individuals possess the qualifications to adhere to established annotation protocols guided by expert knowledge. Furthermore, upon examining the distributions of

the kappa coefficients between experts and non-experts, a noticeable disparity emerged, highlighting the significant difference in data quality that would have resulted if we had allowed all users to annotate "real" data without quality control. Therefore, we strongly recommend that other research teams implement annotation protocols, including a control mechanism, such as the minimum acceptable agreement coefficient discussed in this paper.

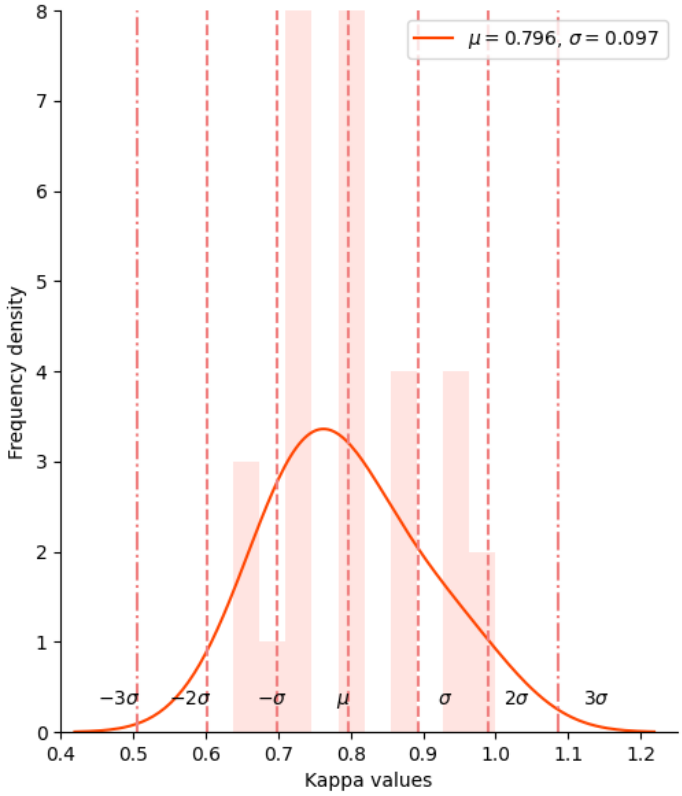

Figure 8. Kappa coefficient distribution for users that passed the set threshold, obtained from n non-experts (where $n = 32$) through the platform with data quality control protocols.

An interesting pattern emerged during the analysis of the second experiment, which involved comparing the time that the non-expert annotators spent annotating the data to the time taken by expert annotators. We observed that the expert annotators completed the annotations in less time than the non-experts. Ideally, all unlabeled data should be annotated by experts, as this would likely result in faster annotation overall.

However, it became clear that annotating a substantial amount of documents solely with the help of experts would require a significant amount of time. In light of this, we decided to leverage the annotations provided by non-expert contributors that adhered to the expert criteria that we were looking for. Although it was evident that these non-expert annotators took nearly three times longer than the experts, shown in Figure 9, their contributions were valuable.

Here, the crowdsourced annotation platform proved to be advantageous. By allowing multiple non-experts, who had passed the threshold, to annotate different parts of the dataset simultaneously or independently, we optimized the annotation process and significantly reduced the overall time that it would have taken otherwise. This approach proved to be a convenient and efficient way to handle the annotation of a large volume of data, responding to RQ1.

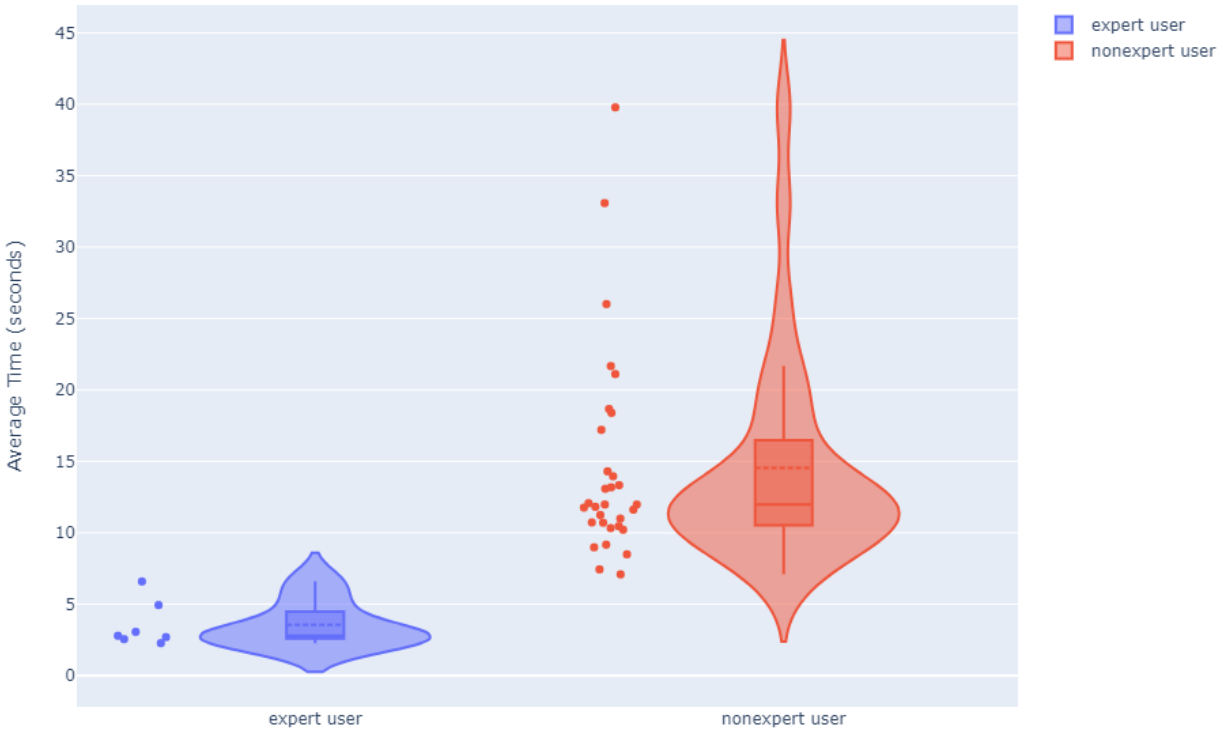

Figure 9. Annotation time distribution per text between experts and non-experts.

4.2. Result 2: Evaluating Non-Experts' Annotations

To answer RQ2, we used Pool C annotation statistics, manually annotated by experts and non-experts, as a proxy to estimate the quality of the 1650 new tweets annotated by non-expert contributors. A reduced number of six tweets accounted for 81.63% of a total of 98 erroneous annotations made by non-experts who reached the kappa coefficient threshold. Errors were equally distributed between the positive (52.04%) and the negative (47.96%) classes. Specifically for the anti-immigrant class, three tweets accounted for 82.35% of errors. Responding to RQ2, while non-experts recognized almost all anti-immigrant tweets in Pool C, providing them with a guideline was insufficient to replicate expert annotation. In some instances, the guideline and the overall task framework reached some limits, which will be discussed here.

1. *"la caravana de migrantes se detuvo aki en Infonavit este es un hondureño"*.
   Translation: "The migrant caravan stopped here at Infonavit, this guy is Honduran".

Fourteen contributors mislabeled tweet number 1 as anti-immigrant, and the average annotation time was high (21 s). Although the tweet did not contain offensive or disrespectful comments about the individual identified as Honduran, the last words of the sentence clearly objectified the person mentioned. Following the criteria established in Section 3.2.1, it was not suffcient to classify it as anti-immigrant, as the tweet fell within a "border zone" between the two classes. When applied to ambiguous tweets or low-intensity anti-immigrant speech, binary classification sometimes is not sufficient to meet the required level of complexity of HS. Other works have attempted to overcome this issue by introducing more classes or sub-classes—for example, based on the intensity of HS [20]—although it seems to add higher complexity to the annotation process and results in lower agreement between contributors.

2. *"En Tijuana la caravana migrante LGBT llegó a una zona residencial rentando un AirBnB. Los vecinos se quejaron, pero en lugar de cuestionar la logística gubernamental, mandaron mensajes de odio hacia los migrantes"*.

Translation: "In Tijuana, the LGBT caravan arrived in a residential neighborhood and rented an Airbnb. The neighbors complained, but instead of questioning the government's logistics, hate messages were sent to migrants".

Tweet number 2 was mislabeled as anti-immigrant by 11 contributors. Although the tweet mentions migrants, hate, and a specific vulnerable group, the author is denouncing the attitudes of the neighbors. In this specific case, contributors either misread the tweet or did not understand the guideline. To keep the annotation process as fast as possible for non-experts, we did not specify in the guideline that a tweet containing anti-immigrant speech or action was not necessarily anti-immigrant. Ideally, a very detailed guideline should be available for contributors to avoid such errors, but it would make the reading process longer, and "time is of the essence" when working with non-experts. For future works, an alternative solution would be to include an optional panel for contributors who need additional assistance in the annotation platform.

It is quite interesting that the three positive instances of anti-immigrant speech with the highest number of erroneous classifications, presented below, all fall under criterion 3 described in Section 3.2.1.

3. *"Denuncian a migrante centroamericano por abusar sexualmente de una niña en Tijuana".*
   Translation: "Central American migrant is reported for sexually assaulting a girl in Tijuana".
4. *"Detectan primer caso de varicela en Caravana Migrante en México. México, debe declarar alerta sanitaria, ante alto riesgo de contagio a la población. Autoridades omiten casos de enfermedades infectocontagiosas y degenerativas de migrantes".*
   Translation: "-First chicken pox case detected in the migrant caravan in Mexico. Mexico must declare a health alert, given the high contagion risk to the population. Authorities omit cases of infectious and degenerative diseases of migrants".
5. *"DETECTAN Y DEPORTAN MARAS SALVATRUCHAS DE LA CARAVANA MIGRANTE EN PIEDRAS NEGRAS".*
   Translation: "MARAS SALVATRUCHAS FROM THE MIGRANT CARAVAN ARE DETECTED AND DEPORTED IN PIEDRAS NEGRAS". (The use of uppercase is preserved.)

All three contain negative stereotypes, portraying migrants as criminals and vectors of dangerous diseases. Beyond knowing whether the information relayed is true or false, tweets 3 and 4 establish an explicit link between criminal behavior and migrant status. In Mexico, this is a process commonly used in the press [34], intentionally or not, that plays an important role in portraying the migrant population as a threat that should be acted upon as such. Responding to RQ2A, the "negative stereotype" tweets were by far the most mislabeled type of anti-immigrant speech, even though these specific stereotypes were explicitly cited in the guideline. For tweet number 3, 19 contributors stated that it was not anti-immigrant, 14 for tweet number 5, and 9 for tweet number 4. Moreover, contributors only spent 10 s on tweet number 3 and 13 s on tweet number 5. The ease with which annotators classified the two tweets as non-anti-immigrant reflects the banality and insidiousness of these negative stereotypes. It is unsettling that a significant portion of the "reliable" contributors did not detect such stereotypes, as they are a common tool by which users spread, consciously or not, HS against migrants on Mexican social media. It also underlines that any initiative to fight anti-immigrant speech online should focus on deconstructing the most commonly used stereotypes.

4.3. Result 3: Any Model Will Learn

To answer RQ3, we designed an experiment to prove that after using the proposed methodology, we obtained good-quality labeled data ready for machine learning algorithms. For the experiment, we used a set of 5272 manually labeled tweets, through the platform described in Section 3.4, half of which were positive (*anti-immigrant*) and half were negative (*non-anti-immigrant*). The texts were transformed into vectors so that they could function as input for three different types of models, so that we could answer RQ3. The vectorization

consisted of using a previously trained convolutional neural network (described in a previous work [26]) as if it were an encoder. In other words, we use the text of the tweets as input and obtained their 250-dimensional vector representation from the Keras API Embedding layer. In the end, we generated a matrix of word embeddings $X_{[5272,250]}$, used throughout the experiment. To verify the learning process, we used two different instances of the labels: labels obtained via our methodology, correct ones (y), and randomly generated labels (y_{rand}). The main idea was to verify that the data collected contained the necessary patterns to distinguish the two classes and that, because of these patterns, we could train any classifier to be used in large-scale production. We used three different types of machine learning models: decision tree, support vector machine, and neural networks. The default configurations for each type of model from the scikit-learn library were used. The training parameters were the same for all model types: {input: X; amount of training examples used to generate the learning curves: (10.0% , 32.5%, 55.0% , 77.5%, 100.0% relative to X); number of shuffle splits for cross-validation: 20; score type: accuracy; test size: 10.0%}. Each training session was performed twice for each type of model; in the first training session, we used the labels obtained with our methodology (y), while, in the second, we used the randomly generated labels (y_{rand}).

In Figures 10–12, we show the learning curves obtained for both the models trained with correct (y) and random (y_{rand}) labels for the three different types of machine learning models: decision tree, support vector machine, and neural networks. In all plots, the x-axis corresponds to the number of samples, the y-axis corresponds to the accuracy value, the blue line refers to the training data, the orange line refers to the test data, and the translucent shade of the curves corresponds to the standard deviation. In the figures, it can be seen that when the platform labels y were used, the three types of models obtained a high accuracy score for the test data (above 0.8 in all cases). In contrast, for y_{rand} labels, the accuracy was always below 0.5. It is also very noticeable in the learning curves that in the case of y labels, there was an increasing accuracy trend for the training data, while, for the y_{rand}, the trend was static or decreasing.

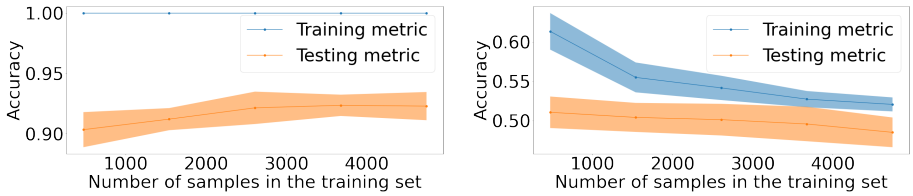

Figure 10. Learning curves for decision tree models. The left side is the training accuracy using our methodology (y); the right side is the training accuracy using random labels (y_{rand}).

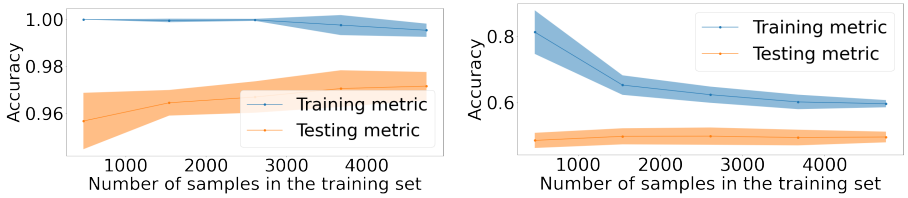

Figure 11. Learning curves for neural network models. The left side is the training accuracy using our methodology (y); the right side is the training accuracy using random labels (y_{rand}).

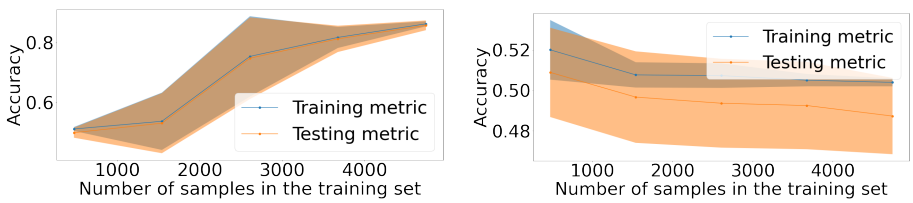

Figure 12. Learning curves for support vector machine models. The left side is the training accuracy using our methodology (y); the right side is the training accuracy using random labels (y_{rand}).

We conclude that the data obtained from the platform are far from being simply noisy data. On the contrary, exploring different machine learning algorithms revealed good-quality data ready to be used for any learning technique (RQ3), thanks to the proposed methodology to control the annotation. It is important to note that, in this stage of the research, it is not our intention to obtain the best classifier but to ensure that the proposed control for data collection guarantees class separability. In the following sections, we will reaffirm this claim by exploring other qualities of the obtained data.

4.4. Result 4: Interpretability Analysis

To complete the analysis of our corpus, we attempted to establish more effective terms associated with anti-immigrant speech in this context. We used a representative sample of the collected data to obtain a classification model via a white box learner that gives a classification result and also allows us to infer the relationships between the results and the terms or words given as explanatory variables. We used a bag-of-words vectorizer and a linear learner (logistic regression) to map directly between individual words and the model coefficients, and also to describe the mathematical relationship between each word and the dependent variable associated with the class to be predicted (positive and negative anti-immigrant speech). The coefficient values indicate how much the class likelihood changes if there is a one-unit shift in an independent word while holding the remaining words in the model constant. The coefficient sign dictates the direction of the change. We resort to this coefficient interpretation to quantify each word's contribution to the prediction of classes. These contributions are subsequently ranked in decreasing order to create a plot such as the one illustrated in Figure 13. The results show that the terms that contribute the most to predicting anti-immigrant speech correspond to elements specific to the migration phenomenon in Mexico. Furthermore, we can see that the terms that contribute the least allude to hate speech, even to particular groups usually targeted by anti-immigrant speech elsewhere, thus confirming the context-specific nature of our corpus. For illustration, we have chosen a sample of tweets containing both high and low contributory terms (highlighted in blue and pink, respectively), as shown below.

1. *"No los dejen ingresar al país Mexicano Señor pdte con todo respeto regréselos por donde vinieron si ellos no nos respetan ni lo haremos nosotros nada más por que uds dice".*
   Translation: "Do not let them enter the country (Mexico), Mr. President. Deport them where they came from. If they do not respect us, we won't either, just because you say so".
2. *" Fuera Hondureños, aquí no los queremos! Gracias señor alcalde, siga con su postura, le aseguro que los Tijuanenses se lo van a agradecer de sobremanera! Ustedes no tienen por que soportar semejantes transgresiones por parte de los Hondureños. @INAMI_mx debería deportarlos!"*
   Translation: "Get out, Hondurans. We don't want you here! Thank you, Mr. Mayor, keep your stance. I assure you that the people of Tijuana will thank you greatly! You do not have to put up with the Honduran's transgressions. @INAMI_mx should deport them".
3. *"Así es! Confirmado por ellos mismos de que en la caravana migrante vienen migrantes maras".*
   Translation: "Yes! Mara migrants confirm they come in the migrant caravan".
4. *"Estoy escuchando el tema arabe y la verdad que mi culo se mueve solo".*
   Translation: "I'm listening to an Arabic song and the truth is that my ass moves by itself".
5. *"@cristobalsoria Retrasado callate la puta boca asqueroso , encima el puto arbitro no nos pita un penalti legal".*
   Translation: "@cristobalsoria disgusting ass, shut the fuck up. On top of it all, the fucking referee doesn't give us a legal penalty".

6. *"Me encantaria darte unas nalgadas ricas. A mí me encantaría pegarte una hostia en la cara".*
Translation: "I would love to spank you. I would love to punch you in the face".

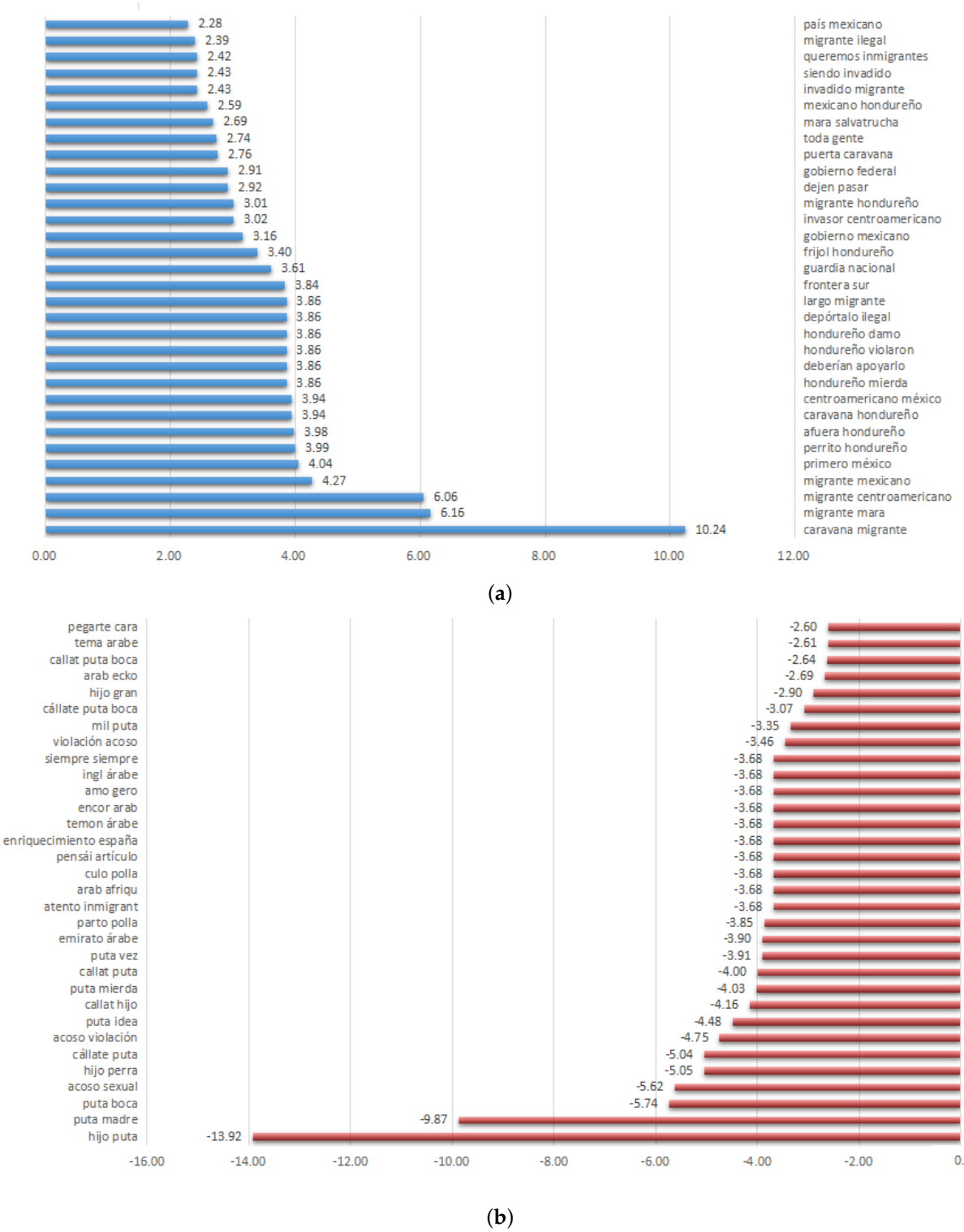

Figure 13. Sample of the terms that contribute more in predicting the speech category of a text (positive or negative anti-immigrant speech). (**a**) Positive values are biased towards anti-immigrant speech. (**b**) Negative values are biased towards negative cases of anti-immigrant speech.

The first three tweets correspond to positive anti-immigrant speech cases, whereas the remaining tweets are negative. As can be observed, the highly contributory terms are coherent in predicting positive anti-immigrant speech cases and vice versa. Negative contributions are not considered determinant in predicting anti-immigrant speech, although they could be pejorative or tied to anti-immigrant speech in some contexts, such as Western Europe or the United States (e.g., *Arab*). Instead, the model gives more relevance to pejorative terms contextualized within the geographical context of the problem—for instance, *perrito hondureño (Honduran dog)*, *(Mara migrant)*, *migrante centroamericano (Central American migrant)*. The above suggests, regarding anti-immigrant speech, that the annotated corpus is constituted by contributory terms and the positive class, which is remarkably consistent with the domain or context of our study.

5. Remarks and Limitations

The research project described in this paper contributes to the field of HS detection by providing several resources for other machine-learning-focused works. Firstly, we created the first corpus of anti-immigrant tweets written in Mexican Spanish and, to our knowledge, the first dataset in Latin American Spanish on this specific type of HS. At the time of writing, 3326 positive and 8256 negative instances compose the "Mexican Anti-Immigrant Speech Training Corpus" (MAISTC). Although the number of negative and positive examples might seem low, we focused on the quality of the data over quantity, while still retaining sufficient instances to train an automatic detection model with acceptable performance. Our dataset has "real" and concrete examples of HS and non-HS produced in the Mexican Twittersphere and collected through a rigorous annotation protocol based on empirical knowledge and solid definitions. In addition, we can always include new examples in the corpus. The unanticipated changes in the Twitter API politics regarding academic research were a setback for this project and any research on social media. Despite this situation, with our data collection protocol proposal, we could expand the analysis to other social media platforms, such as Facebook, Instagram, or YouTube.

Including quality control mechanisms in text annotation is a noteworthy contribution, as it remains independent of the specific investigation phenomenon and language. Consequently, these mechanisms can be readily applied to other research projects, enhancing their reliability in corpus annotation. However, the choice of including non-experts is not without drawbacks. Finding non-experts is not as simple as advertising the project on social media. The low interest from the online community users that we targeted to partake in the annotation task meant that we had to rely on more traditional social networks, mainly professional and personal. Overall, the participation of non-experts and the number of data annotated through this process was lower than expected, but this does not diminish the validity of the methodology. A workaround solution would involve non-experts with existing interest in research, such as students.

Furthermore, this research proposes a sophisticated definition of online Mexican anti-immigrant speech based on the empirical observation of the phenomenon. This definition is functional in the Mexican context but is also general enough to be used in other contexts. The greatest challenge that we had to overcome during the definition work was the rapid evolution of migration and migration policies, which transformed the terms of the online debate throughout the period of interest (2018–2022). Not identifying these moments of change and not adapting the data collection strategy accordingly meant taking the risk that the final corpus would be particularly sensitive to certain events. In this work, such a risk was controlled by including "temporal filters" over the entire period, defined based on detailed knowledge of the context and continuous monitoring of migration-related topics in Mexico, which corresponded to the various significant elements that detonated the discussion on migration.

Thanks to our baseline automatic classification model, we have achieved the ability to automate classification processes on large volumes of data. However, our experimentation has highlighted the criticality of being mindful of potential biases induced by the data prepa-

ration methods. Furthermore, it is crucial to acknowledge the limitation of interpretability in our best-performing baseline model, which relies on a neural network. To address this limitation comprehensively, we have devised a strategy to leverage white box learners that enhance interpretability and transparency. We will implement a continuous monitoring mechanism to assess the frequency of incorrect predictions made by our classification models, enabling us to identify whether such mistakes are due to algorithmic factors or the introduction of new data in the corpus. With the same method, we can scrutinize the words associated with the highest correlated class and determine their semantic relevance to Mexican anti-immigrant HS, thus enabling our team to conduct thorough verification.

6. Conclusions and Future Work

In conclusion, we describe in this paper the methodology used to create an ad-hoc anti-immigrant dataset to train automatic detectors of online anti-immigrant speech published in the Mexican context. We elaborate in detail on the methodology, which introduces elements not yet considered in the literature, and discuss the quality of the obtained data. The experiments conducted in this study provide valuable insights and answer the research questions.

We have verified that the kappa coefficient value can be an efficient mechanism for quality annotations from non-experts. It also enhances the temporal efficiency by involving and utilizing contributions from non-expert annotators who meet the expert criteria, rather than relying solely on expert annotators (RQ1).

We have confirmed that, given a sophisticated but concise guideline, non-expert annotators can detect almost all categories of anti-immigrant speech (RQ2), although some deeply rooted stereotypes remain undetected (RQ2A).

We have verified that the data obtained from experts and non-experts represent a valid input for any learning algorithm whenever there is a quality control mechanism in the annotation process (RQ3).

An analysis based on a white box model proved that the annotation process induces semantic aspects directly associated with our study object, which allows us to assert that our annotation approach is somewhat context-sensitive (RQ4).

Overall, we have demonstrated in this paper that it is possible to build a good-quality, context-sensitive training dataset for machine learning purposes by combining experts' and non-experts' annotations supervised by a control mechanism. As we move towards the creation and analysis of the Mexican anti-immigrant speech corpus, we must conduct a formal external validation of our classifier to ensure its effectiveness when dealing with real new data. We will use the validated model to classify the rest of the data that we obtained through AGEI and INGEOTEC, several million tweets, thus creating the most complete database for the study of online Mexican anti-immigrant speech in a new context of intense migration and political turmoil. With all the necessary resources at hand, we will focus on geographic, temporal, semantic, and topological dimension analysis of online anti-immigrant speech in Mexico.

Author Contributions: Conceptualization, T.C., A.M.-V.; Methodology, A.M.-V., T.C.; Software, K.G.-H., A.M.-V.; Validation, A.M.-V., T.C., K.G.-H., E.A.-B.; Investigation, A.M.-V., T.C., K.G.-H., E.A.-B.; Resources, A.M.-V., E.A.-B.; Data curation, T.C., K.G.-H.; Writing—original draft preparation, T.C., A.M.-V., K.G.-H., E.A.-B.; Writing—review and editing, A.M.-V., T.C., K.G.-H., E.A.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Due to privacy and ethical restrictions, the anonymized datasets are shared only on demand for research purposes. Please visit https://github.com/Kgazcah/Annotation-platform (accessed on 19 June 2023).

Acknowledgments: The authors would like to express their gratitude to Oscar Gerardo Sánchez and Daniela Alejandra Moctezuma Ochoa for providing the raw data. Special thanks to Beatriz Zepeda Rivera and Loraine Morales Pino for their expert annotations, Isabel Sofía de la Cruz Abrín for English revision, and all the non-expert annotators, who made a significant contribution to the development of this research work.

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

References

1. Leite, P.; Correa-Lazzarini, A.; Suárez, M.; Flores-Rodríguez, P.;Ramírez-Rojas, A.; Méndez-Cadena, E.; DelPino-Pacheco, M. Guía para la Acción Pública. Comunicación sin Xenofobia. Recomendaciones Para Medios y Redes Sociales. 2022. Available online: http://www.conapred.org.mx/index.php?contenido=documento&id=411&id_opcion=147 (accessed on 19 June 2023)
2. Xenofobiacero Reporte de Conversación de Migración y Xenofobia México. (OIM, 2021). Available online: https://xenofobiacero.org/blog/datos-clave-sobre-los-comentarios-de-odio-hacia-los-migrantes-en-las-redes-sociales-en-mexico (accessed on 4 May 2023).
3. Redman, T. If Your Data Is Bad, Your Machine Learning Tools Are Useless. *Harvard Business Review* 2018. Available online: https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless?utm_medium=social&utm_campaign=hbr&utm_source=twitter (accessed on 15 March 2023).
4. Caicedo, M.; Mena, A.M. Imaginarios de la Migración Internacional en México: Una Mirada a los que se van y a los Que Llegan: Encuesta Nacional de Migración. (Universidad Nacional Autónoma de México. Instituto de Investigaciones Jurídicas, 2015). Available online: http://ru.juridicas.unam.mx:80/xmlui/handle/123456789/58480 (accessed on 28 September 2022).
5. Wong, T. *The Politics of Immigration: Demographic Change, and American National Identity*; Oxford University Press: Oxford, UK, 2016.
6. Cohen, J. Zero Tolerance: The Trump Administration's Permanent Anti-Immigrant Offensive and its Repercussions in the Americas. *Polit. Am.* **2021**, 37, 39–60. Available online: https://www.cairn.info/revue-politique-americaine-2021-2-page-39.htm (accessed on 25 February 2023). [CrossRef]
7. París Pombo, M. Régime de frontières et politiques migratoires dans le nord du Mexique (2018–2020). *Am. Lat.* **2022**, *1*, 31–60. Available online: https://www.cairn.info/revue-amerique-latine-2022-1-page-31.htm (accessed on 11 February 2023).
8. Conapred & INEGI Enadis 2017. Prontuario de Resultados. (Consejo Nacional para Prevenir la Discriminación, 2018). Available online: https://www.inegi.org.mx/programas/enadis/2017/ (accessed on 19 June 2023).
9. Sieff, K.; Clement, S. Inmigrantes indocumentados vistos de forma desfavorable en México, de acuerdo a encuesta. *Washington Post*, 17 July 2019. Available online: https://www.washingtonpost.com/world/the_americas/inmigrantes-indocumentados-vistos-de-forma-desfavorable-en-mexico-de-acuerdo-a-encuesta/2019/07/16/251acc72-a749-11e9-8733-48c87235f396_story.html (accessed on 4 May 2023).
10. Ferra, I.; Nguyen, D. #Migrantcrisis: Tagging the European migration crisis on Twitter. *J. Commun. Manag.* **2017**, *21*, 411–426. Available online: https://www.emerald.com/insight/content/doi/10.1108/JCOM-02-2017-0026/full/html (accessed on 25 March 2023).
11. Torre Cantalapiedra, E. Migración, racismo y xenofobia en internet: Análisis del discurso de usuarios contra los migrantes haitianos en prensa digital mexicana. *Rev. Pueblos Front. Digit.* **2019**, *14*. Available online: http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S1870-41152019000100106&lng=es&nrm=iso&tlng=es (accessed on 27 December 2022). [CrossRef]
12. Toudert, D. Crisis de la caravana de migrantes: Algunas realidades del discurso público en Twitter. *Migr. Int.* **2021**, *12*. Available online: https://migracionesinternacionales.colef.mx/index.php/migracionesinternacionales/article/view/2172 (accessed on 15 January 2023). [CrossRef]
13. Pérez Díaz, M.; Aguilar Pérez, M.; Pérez Díaz, M.; Aguilar Pérez, M. #LadyFrijoles: Señalamiento, discriminación y estigma de migrantes centroamericanos a través de redes sociales en México. *Andamios* **2021**, *18*, 223–243. Available online: http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S1870-00632021000100223&lng=es&nrm=iso&tlng=es (accessed on 27 December 2022).
14. Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; Patti, V. Resources and benchmark corpora for hate speech detection: A systematic review. *Lang. Resour. Eval.* **2021**, *55*, 477–523. Available online: https://link.springer.com/10.1007/s10579-020-09502-8 (accessed on 27 December 2022). [CrossRef]
15. Arcila-Calderón, C.; Amores, J.J.; Sánchez-Holgado, P.; Vrysis, L.; Vryzas, N.; Alonso, M.O. How to Detect Online Hate towards Migrants and Refugees? Developing and Evaluating a Classifier of Racist and Xenophobic Hate Speech Using Shallow and Deep Learning. *Sustainability* **2022**, *14*, 13094. Available online: https://www.mdpi.com/2071-1050/14/20/13094 (accessed on 10 May 2023). [CrossRef]
16. Pitropakis, N.; Kokot, K.; Gkatzia, D.; Ludwiniak, R.; Mylonas, A.; Kandias, M. Monitoring Users' Behavior: Anti-Immigration Speech Detection on Twitter. *Mach. Learn. Knowl. Extr.* **2020**, *2*, 192–215. Available online: https://www.mdpi.com/2504-4990/2/3/11 (accessed on 19 September 2022). [CrossRef]
17. Siegel, A.; Nikitin, E.; Barberá, P.; Sterling, J.; Pullen, B.; Bonneau, R.; Nagler, J.; Tucker, J. Trumping Hate on Twitter? Online Hate Speech in the 2016 U.S. Election Campaign and its Aftermath. *Q. J. Political Sci.* **2021**, *16*, 71–104. Available online: http://www.nowpublishers.com/article/Details/QJPS-19045 (accessed on 19 September 2022). [CrossRef]

18. Capozzi, A.; Lai, M.; Basile, V.; Poletto, F.; Sanguinetti, M.; Bosco, C.; Patti, V.; Ruffo, G.; Musto, C.; Polignano, M.; et al. "Contro L'Odio": A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media. *IJCoL Ital. J. Comput. Linguist.* **2020**, *6*, 77–97. Available online: https://journals.openedition.org/ijcol/659 (accessed on 19 September 2022). [CrossRef]
19. Florio, K.; Basile, V.; Lai, M.; Patti, V. Leveraging Hate Speech Detection to Investigate Immigration-related Phenomena in Italy. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, UK, 3–6 September 2019; pp. 1–7.
20. 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. Available online: https://aclanthology.org/L18-1443 (accessed on 19 April 2023).
21. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Pardo, F.M.R.; Rosso, P.; Sanguinetti, M. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In Proceedings of the 13th International Workshop On Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 54–63. Available online: https://www.aclweb.org/anthology/S19-2007 (accessed on 23 August 2021).
22. Plaza-Del-Arco, F.; Molina-González, M.; Ureña-López, L.; Martín-Valdivia, M. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *ACM Trans. Internet Technol.* **2020**, *20*, 12:1–12:19. [CrossRef]
23. Hasan, A.; Sharma, T.; Khan, A.; Al-Abyadh, M.H.A. Analysing Hate Speech against Migrants and Women through Tweets Using Ensembled Deep Learning Model. *Comput. Intell. Neurosci.* **2022**, *2022*, e8153791. Available online: https://www.hindawi.com/journals/cin/2022/8153791/ (accessed on 24 April 2023). [CrossRef]
24. Aragon, M.; Carmona, M.; Montes, M.; Escalante, H.; Villaseñor-Pineda, L.; Moctezuma, D. Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. In Proceedings of the 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Negation, Spanish, 24 September 2019.
25. Aldana-Bobadilla, E.; Molina-Villegas, A.; Montelongo-Padilla, Y.; Lopez-Arevalo, I.; SSordia, O. A language model for misogyny detection in latin american spanish driven by multisource feature extraction and transformers. *Appl. Sci.* **2021**, *11*, 10467. [CrossRef]
26. Cattin, T.; Molina-Villegas, A.; Fuentes, J.; Siordia, O. The Geopolitical Repercussions of US Anti-immigrant Rhetoric on Mexican Online Speech About Migration: A Transdisciplinary Approach. *Adv. Geospat. Data Sci.* **2022**, *1*, 41–51. [CrossRef]
27. López-Ramírez, P.; Molina-Villegas, A.; Siordia, G.S. Geographical aggregation of microblog posts for LDA topic modeling. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4901–4908. [CrossRef]
28. Molina-Villegas, A. La incidencia de las voces misóginas sobre el espacio digital en México. In *Jóvenes, Plataformas Digitales Y Lenguajes: Diversidad Lingüística, Discursos E Identidades*; Pérez Barajas, A.E., Arellano Ceballos (coord.), A.C., Eds.; Página Seis: Zapopan, Mexico, 2022; pp. 39–61.
29. Graff, M.; Moctezuma, D.; Miranda-Jiménez, S.; Tellez, E. A Python library for exploratory data analysis on twitter data based on tokens and aggregated destination information. *Comput. Geosci.* **2022**, *159*, 105012. Available online: https://www.sciencedirect.com/science/article/pii/S0098300421002946 (accessed on 19 June 2023). [CrossRef]
30. Wiegand, M.; Ruppenhofer, J.; Kleinbauer, T. Detection of Abusive Language: The Problem of Biased Datasets. In Proceedings of the 2019 Conference Of The North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2–9 June 2019; Volume 1, pp. 602–608. Available online: https://aclanthology.org/N19-1060 (accessed on 18 May 2023).
31. Cohen, J. Les nativistes face aux immigrés aux États-Unis. *Après-demain* **2020**, 25–27. Available online:. (accessed on 12 April 2023). [CrossRef]
32. Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; Wojatzki, M. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *arXiv* **2016**, arXiv:1701.08118.
33. García-Díaz, J.; Cánovas-García, M.; Colomo-Palacios, R.; Valencia-García, R. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Gener. Comput. Syst.* **2021**, *114*, 506–518. (accessed on 16 March 2023) [CrossRef]
34. Canales, A. El malestar con las migraciones: Perspectivas desde el Sur. *Anthropos* **2021**, *9*, 52.