applied sciences

MDPI

Article

An Explainable Artificial Intelligence Model for Detecting Xenophobic Tweets

Gabriel Ichcanziho Pérez-Landa [1], Octavio Loyola-González [2,*] and Miguel Angel Medina-Pérez [1,2]

[1] School of Science and Engineering, Tecnologico de Monterrey, Carretera al Lago de Guadalupe Km. 3.5, Atizapán 52926, Mexico; gabriel.ich@exatec.tec.mx (G.I.P.-L.); migue@tec.mx (M.A.M.-P.)

[2] Altair Management Consultants, Calle de José Ortega y Gasset 22-24, 5th Floor, 28006 Madrid, Spain

* Correspondence: olg@altair.consulting

Abstract: Xenophobia is a social and political behavior that has been present in our societies since the beginning of humanity. The feeling of hatred, fear, or resentment is present before people from different communities from ours. With the rise of social networks like Twitter, hate speeches were swift because of the pseudo feeling of anonymity that these platforms provide. Sometimes this violent behavior on social networks that begins as threats or insults to third parties breaks the Internet barriers to become an act of real physical violence. Hence, this proposal aims to correctly classify xenophobic posts on social networks, specifically on Twitter. In addition, we collected a xenophobic tweets database from which we also extracted new features by using a Natural Language Processing (NLP) approach. Then, we provide an Explainable Artificial Intelligence (XAI) model, allowing us to understand better why a post is considered xenophobic. Consequently, we provide a set of contrast patterns describing xenophobic tweets, which could help decision-makers prevent acts of violence caused by xenophobic posts on Twitter. Finally, our interpretable results based on our new feature representation approach jointly with a contrast pattern-based classifier obtain similar classification results than other feature representations jointly with prominent machine learning classifiers, which are not easy to understand by an expert in the application area.

Keywords: Xenophobia; Twitter; Explainable Artificial Intelligence

check for updates

Citation: Pérez-Landa, G.I.; Loyola-González, O.; Medina-Pérez, M.A. An Explainable Artificial Intelligence Model for Detecting Xenophobic Tweets. *Appl. Sci.* **2021**, *11*, 10801. https://doi.org/10.3390/app112210801

Academic Editors: María Paz Sesmero Lorente, Plamen Angelov and Jose Antonio Iglesias Martinez

Received: 23 September 2021
Accepted: 27 October 2021
Published: 16 November 2021

1. Introduction

Xenophobia has been a social behavior present in people since the beginning of humanity. Fear and rejection of the different have led many people to distrust, belittle, and even hate other people who belong to a different social environment [1]. For example, recently, the uncertainty and lack of information about COVID-19 have generated acts of Xenophobia toward people of Chinese origin [2]. These xenophobic acts are more intense using social networks because people can show this discriminatory act non-physical, even with fake profiles or using bots [3,4].

As of March 2021, 5.1 billion people have access to the Internet, representing 65.6% of the world's population [5]. With the growth of the Internet, many social networks were created, which provide a virtual space for people to communicate through them by sharing posts, pictures, and videos [6]. Some of the most popular social networks today are Facebook, YouTube, WhatsApp, Twitter, LinkedIn, among others [7].

With the spread of social networks for interacting on the Internet, and the use of fake profiles, hate posts towards other peoples have been increasing as social media expands. Consequently, these posts cause not only the use of offensive language but even lead to acts of physical violence [8]. Increasingly, social media plays an essential role in amplifying and accelerating violence in the real world. The violent behavior present in social networks such as Twitter, Facebook, among others, has considerably increased the probability of committing acts of violence in the real world, which can be fatal [9].

The next examples of violence preceded by comments on social networks were extracted from *Citizens Crime Commission of New York City* [9]:

- SUMMARY {Date: 23 July 2017, Place: Nashville, TN, USA, Platform: Facebook}: *"A 20-year-old man and their 37-year-old mother were shot and killed in their home hours after the 20-year-old posted on Facebook multiple photos of himself with large wads of cash, jewelry, and shopping bags."*
- SUMMARY {Date: 6 October 2016, Place: St. Louis, MO, USA, Platform: Twitter}: *"An 18-year-old man fatally shot a 33-year-old police officer who was responding to a disturbance call. The shooter had repeatedly threatened violence on their Twitter page for months before the shooting."*

With the aim of stopping the hatred, racism, and Xenophobia present on the Internet, many web pages have rules for their users that prohibit these types of behavior. However, a post with violent content is visible until it is detected as brutal by some administrator user or a system that does not work in real-time, creating a wave of violence. In contrast, it is still posted [4]. Facebook has announced that there is no place for hate speech on their social network, and they would battle against racism and Xenophobia. However, the solution proposed by Facebook and Twitter indicates that the problem depends on human effort, leaving the users the responsibility of reporting offensive comments [10].

According to Pitsilis et al. [11], detecting offensive posts requires a great deal of work for human annotators, but this is a subjective task providing personal interpretation and bias. As Nobata et al. [12] mentioned, the need to automate the detection of abusive posts becomes very important due to the growth of communication among people on the Internet.

Each social network has its privacy policy, which could or could not allow developers to analyze the publications that users make on their platforms. For example, Facebook does not recognize the extraction of comments from publications, except that these comments are from a page that you manage [13]. Although there are pages such as *export comments* [14] that allow this information to be obtained. However, Facebook only allows downloading publications with less than 485 comments for a price of USD 11. On the one hand, Twitter natively has an API that enables developers to download their users' publications through Twitter Streaming API, and Twitter REST API [15].

Twitter is a social network characterized by the briefness of the posts, with a maximum of 280 characters. In the first quarter of 2019, Twitter reported 330 million users and 500 million tweets per day [16]. In the United States, Twitter is a powerful communication tool for politicians since it allows them to express their position and share their thoughts with many of the country's population. This opinion can dramatically change citizens' behavior, even if it was only written on Twitter [17]. Based on what was said previously, an open problem is detecting xenophobic tweets by using an automated Machine Learning model that allows experts to understand why the tweet has been classified as xenophobic.

Hence, this research focuses on developing an Explainable Artificial Intelligence model (XAI) for detecting xenophobic tweets. The main contribution of this research is to provide an XAI model in a language close to experts in the application area, such as psychologists, sociologists, and linguists. Consequently, this model can be used to analyze and predict the xenophobic behavior of users in social networks.

As a part of this research, we have created a Twitter database in collaboration with experts in international relations, sociology, and psychology. The experts have helped us to classify xenophobic posts in our Twitter database proposal. Then, based on this database, we have extracted new features using Natural Language Processing (NLP), jointly with the XAI approach, creating a robust and understanding model for experts in the field of Xenophobia classification, particularly experts in international relations.

This document is structured as follows: Section 2 provides preliminaries about Xenophobia and contrast pattern-based classification. Section 3 shows a summary of works related to Xenophobia and hate-speech classification. Section 4 introduces our approach for Xenophobia detection in Twitter. Section 5 describes our experimental setup. Section 6 contains our experimental results as well as a brief discussion of the results. Finally, Section 7 presents the conclusions and future work.

2. Preliminaries

2.1. Xenophobia

According to Wicker [18], Xenophobia is a hostile response that society has towards foreign people, providing stereotypes and prejudices against these unfamiliar people. Usually, this aggressive response has political or individual purposes, which seek to improve society's cohesion through discrimination of foreign people. This xenophobic response can become so strong as to be a distinctive feature of a population. Additionally, Wicker mentions that the fear and hatred with which Xenophobia is fostered are qualities based on subjective experiences, which have a background in the education and values that we receive by society. In this way, Xenophobia, being a hostile response, can also be considered as social behavior, which can be used to control communities, in which hatred of the other forms a way not only to generate identity but to promote acts of violence before third parties [19].

For Crush [20], Xenophobia proceeds through dynamic public rhetoric, which shows contempt through verbal offenses or acts of violence. Xenophobia actively stigmatizes immigrants, calling them "threats" to society, and thus making them *"the cause of social problems."*

It is essential to understand what are the differences between racism and Xenophobia. However, these phenomena are often intertwined, they can also present themselves in other ways, and each one of them entails different societal problems. Therefore, the means to solve and correct these social behaviors are different. On the one hand, racism implies discrimination against human beings based on their physical characteristics, such as skin tone, weight, height, and facial features, among others. On the other hand, Xenophobia denotes *"behavior specifically based on the perception that the other is alien or originates from outside the community or nation"* [21].

Some of the definitions of Xenophobia were presented at the international conference in Migration, Racism, Discrimination, and Xenophobia [21]:

- **By the standard dictionary definition:** Xenophobia is the intense dislike or fear of strangers or people from other countries.
- **As a sociologist's point of view:** Xenophobia is an attitudinal orientation of hostility against non-natives in a given population.
- **From world conference against racism:** Xenophobia describes attitudes, prejudices, and behavior that reject, exclude, and often vilify persons, based on the perception that they are outsiders or foreigners to the community, society, or national identity.

Xenophobia has managed to stay on social networks, which it has spread quickly. The number of discourses of violence that motivate discrimination or violence against immigrants as well as minority groups has grown enormously [22]. Due to this growth, experts have questioned states and social media companies on stopping this spread. Consequently, these online and offline hate speeches have increased social tension, leading to violent attacks that can end in death [23]. For these reasons and more than ever, social networks must take action to mitigate this behavior on their platforms and reduce the probability that people will be injured in the real world [24].

One problem that exists is how to know when a comment is xenophobic? According to Bucio [25], in Mexico, posts with hate hashtags are daily published, among which are: *#indio, #puto, #naco*, among others. The posts with hate hashtags cause various social problems such as classism, homophobia, racism, sexism, Xenophobia, etc. [25]. The most alarming thing about this xenophobic behavior on social networks is that public figures write some of these xenophobic comments. Additionally, Bucio mentioned that public figures are not penalized because their xenophobic posts are treated as "black humor or harmless comments"; allowing several people to spread hate speeches hidden in "humor" publications. They are also trying to lessen the fact that they are normalizing xenophobic behaviors. As they are "humor", the people who write these posts do not contemplate the consequences that their comments may have on people's lives, such as sadness, pain, distress, humiliation, isolation, and dignitary insult [26].

The problem of writing xenophobic posts is that we are unaware of how dangerous our behavior can be on social networks. At the time when we started to spread publications that incite discrimination, that promote hatred and violence towards others, we are complicit in the consequences that these may have [27]. Threats, insults, blows, even attacks that end in the death of third parties are caused day by day as a result of the normalization of xenophobic behavior on social networks [28].

Social networks are aware of xenophobic behavior; however, there are still no quick and precise measures to address this issue with the importance it needs. The lack of an automatic xenophobic publication detection tool makes them last longer online and can harm third parties while they are not deleted. There are even cases where "after deleting offensive posts", they tend to reappear after a while [4].

Finally, the classification of xenophobic comments on social networks is very recent [29–31]. According to Plaza-Del-Arco et al. [32], the rating of xenophobic posts is a poorly addressed topic. Besides, Loyola-González [33] mentions that there is currently a trend to transform unexplainable models (black-box) to explainable models (white-box), particularly in sectors such as health care. Hence, our proposal aims to classify xenophobic posts through an Explainable Artificial Intelligence model. With the use of XAI models, in such wat that experts can have a set of explainable patterns describing xenophobic posts.

2.2. Contrast Pattern-Based Classification

Nowadays, using Explainable Artificial Intelligence (XAI) models instead of the highly used black-box artificial intelligence models is a trend, especially for most important areas, for example, criminal justice, healthcare, finance, among others [34]. The main advantages of using XAI models are that they can achieve reasonable classifications results and provide an explication of the model in a language close to the experts in the related work [35].

The contrast pattern-based classifiers are a group of algorithms that follows the XAI approach [36]. These classifiers can provide similar classification results as other well-known classification models, such as Nearest Neighbor, Naïve–Bayes bagging, Decision Trees, boosting, and even support vector machines [37]. Additionally, according to Loyola-González [36], contrast pattern-based classifiers can be used to deal with class imbalance problems.

A pattern is an expression characterized in a particular language that portrays a collection of objects. In contrast, a contrast pattern is a pattern that frequently appears in a class and at the same time infrequently in the remaining courses [37]. Ordinarily, a pattern is represented by a conjunction of relational statements, each with the form:
[f_i # v_j], where v_j is a value within the space of feature f_i, and # is a relational operator taken from the set $\{ =, \neq, \leq, >, \in, \notin, \}$ [33,36,38]. For example, [*violent foreigners* = *"present"*] $\wedge$ [*hate-speech* ≥ 0.11], is a pattern describing post xenophobes.

Thereby, a contrast pattern-based classifier uses a group of patterns to construct a model capable of classifying query objects into a class [37]. Nowadays, there exist pattern-based classifiers such as: OCSVM [39], HeDEx [40], CAEP [37], iCAEP [41], PBC4cip [36], PCB4occ [42], and PBCEQ [43] from which PBC4cip and PBCEQ have shown to obtain the best results for class imbalance problems [43].

To build a contrast pattern-based classifier, there exist three main steps: Mining, filtering, and finally, the classification step [36]. Mining is committed to discovering a set of conceivable patterns by an exploratory examination employing a search-space, characterized by many inductive limitations provided by the user. Filtering is committed to choosing a set of high-quality patterns coming from the mining step. Classification is responsible for looking at the leading procedure for combining the data given by a subset of patterns and builds a precise model based on patterns [3].

For mining the contrast patterns, it is required to construct a Decision Tree (DT). A DT contains components of the tree structure based on the graph theory. Then, a DT can be reported as a directed graph where two vertices are associated by only one path [44]. The top-down approach is the most used method to make an inducing decision tree [33]. This approach is based on the divide and conquer method [45]. It begins by making a node at the root with all the objects of the training database D, and after, it splits the root node into two disjoint subsets, the left child D_l and the right child D_r; this process is performed again recursively until a halt criterion is reached [46].

According to Loyola-González [33], the extraction of contrast patterns obtained from only one decision tree generates very few contrast patterns. Nevertheless, the pattern's extraction from several equal decision trees causes several duplicate patterns. To deal with this problem, collecting diverse decision trees for extracting the contrast patterns can mitigate this problem [47]. Each pattern is the conjunction of the properties $\mathrm{f_i}$ # $\mathrm{v_j}$ in a path from the root node to a leaf node; that is, any path from the root to a leaf decides the conjunction of properties, making a pattern. Finally, only those patterns satisfying the contrast pattern condition are kept [3]. For example, Figure 1 shows an example of a hypothetic decision tree for Xenophobia classification.

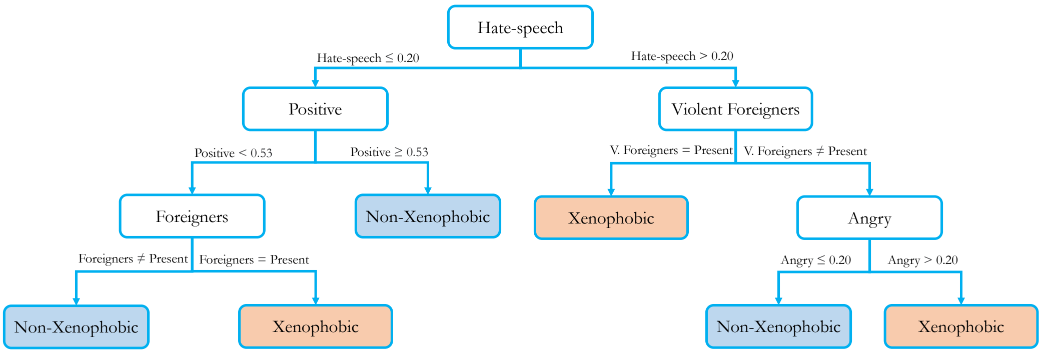


Figure 1. Decision Tree example.

From this decision tree, we can extract the following six contrast patterns:

- $\mathrm{P_1}$ = [$\textit{Hate-speech} \leq 0.20$] ∧ [$\textit{Positive} < 0.53$] ∧ [$\textit{Foreigners} \neq$ *"Present"*]
- $\mathrm{P_2}$ = [$\textit{Hate-speech} \leq 0.20$] ∧ [$\textit{Positive} < 0.53$] ∧ [$\textit{Foreigners} =$ *"Present"*]
- $\mathrm{P_3}$ = [$\textit{Hate-speech} \leq 0.20$] ∧ [$\textit{Positive} \geq 0.53$]
- $\mathrm{P_4}$ = [$\textit{Hate-speech} > 0.20$] ∧ [$\textit{Violent Foreigners} =$ *"Present"*]
- $\mathrm{P_5}$ = [$\textit{Hate-speech} > 0.20$] ∧ [$\textit{Violent Foreigners} \neq$ *"Present"*] ∧ [$\textit{Angry} \leq 0.20$]
- $\mathrm{P_6}$ = [$\textit{Hate-speech} > 0.20$] ∧ [$\textit{Violent Foreigners} \neq$ *"Present"*] ∧ [$\textit{Angry} > 0.20$]

from which $\mathrm{P_1}$, $\mathrm{P_3}$, $\mathrm{P_5}$ corresponds to the *Non-Xenophobia* class and the remaining patterns ($\mathrm{P_2}$, $\mathrm{P_4}$, $\mathrm{P_6}$) corresponds to the *Xenophobia* class.

For filtering the contrast patterns, there exist two main approaches, based on set theory (i) and quality measures (ii) [33]. The first approach removes duplicate and specific patterns and also removes redundant items. The second approach allows generating a pattern ranking in which the raking position is based on the discriminative power of the patterns. It can be explained the three main steps of the contrast pattern filtering process based on the set theory as follows:

- **Removing duplicated contrast patterns:** when two or more contrast patterns have the same items and cover the same objects, they are colled duplicate patterns. This problem usually happens because the contrast patterns come from several decision trees built from the same training database. Then, all the duplicated contrast patterns are removed to reduce the contrast patterns, and only one is kept.
- **Removing specific contrast patterns:** commonly, some specific patterns are extracted in the mining process. Let us suppose that there are two contrast patterns $\mathrm{P_1}$ and $\mathrm{P_2}$, that belong to the same class. The pattern $\mathrm{P_2}$ is considered more specific than

P_1 if all the items contained in P_1 are also in P_2 but not vice versa. For example, let P_1 = [*Hate-speech* $\leq$ 0.20] $\wedge$ [*Positive* $\geq$ 0.53] and P_2 = [*Hate-speech* $\leq$ 0.20] $\wedge$ [*Positive* $\geq$ 0.53]$\wedge$ [*Negative* $<$ 0.24]. Since all the items presented in P_1 are also in P_2, but P_2 has one more item, then P_2 is considered more specific than P_1. Therefore, according to [48], P_2 should be removed.

- **Removing redundant items:** An item I_1 is more general than another item I_2 if all of the objects presented in I_1 are also in $_2$, but not the other way around. We also say that I_2 is redundant with I_1. If two items in a pattern are redundant, the most general item is eliminated [3]. We can provide the next example of a pattern with redundant items: [*Hate-speech* $>$ 0.20] $\wedge$ [*Hate-speech* $>$ 0.41], which is simplified to [*Hate-speech* $>$ 0.41] because an entry with a hate-speech greater than 0.41 is also greater than 0.20.

To explain the second approach of filtering the contrast pattern based on quality measures, let us take the support as our quality measure. Let p be a pattern, and C = $\{C_1, C_2, C_3, \ldots, C_n\}$ a set of classes such that $C_1 \cup C_2 \cup C_3 \cup \ldots \cup C_n = U$; then, support for p is obtained after dividing the number of objects that belong to C_i that were described by p by the total number of objects that are in C_i [36]. After obtaining all the supports for each contrast pattern, they can be ranked from higher to lower support. Then, it can be selected only the first n patterns and remove the rest of the patterns [38]. Additionally, a minimum threshold can be proposed for considering a pattern with enough quality and then eliminate all those patterns that do not reach the minimum threshold.

There are two main strategies for classifying the contrast patterns: unweighted (i) and weighted (ii) scores. The unweighted strategy is easy to compute and understand but is not suitable for all kinds of problems, such as class imbalance problems because it tends to be biased towards the majority class [36]. The weighted approach is more computationally expensive but is ideal for handling both balance and imbalance problems [33]. Specifically, for our Xenophobia detection problem on social networks, we have an imbalanced database. In this way, we decided to use PBC4cip [36] as our contrast pattern-based classifier, not only because it has been proved to obtain the best classification results jointly with PBCQE [43] for class imbalance problems, but also because PBC4Cip provides significantly fewer patterns than others contrast pattern-based classifiers [36]. PBC4cip weights the sum of supports in each class at the training stage by considering all contrast patterns covering a query object and class imbalance level. This preparing plan is distinctive from conventional classifiers, which only sum the supports [3].

The contrast pattern-based classifiers have been used to solve real-world problems, where they have managed to obtain similar or best results to other classifiers. It can be mentioned some of the most relevant applications where contrast pattern-based classifiers have been applied, such as improvement of road safety [49], rule construction from crime pattern [50], the discovery of an unusual rule within a cerebrovascular examination [51], describing political figures [52], the observation of sales trends in dynamic markets [53], bot detection on Twitter [3], bot detection on Web Log Files [54], detection of alarm patterns in industrial alarm floods [55], complex activity recognition in smart homes [56], discriminating deviant behaviors in MOBA games [57], summarizing significant changes in network traffic [58], among others.

3. Related Work

In this section, we present previous works related to our research. All these works have similar semantics since their objective is to identify undesirable behaviors in social networks using Machine Learning.

Pitropakis [59] addressed the issue of Xenophobia classification on Twitter. For that, they created a Xenophobia database on Twitter using keywords associated with Xenophobia. Additionally, they used a geolocation filter to focus on the UK, USA, and Canada countries. Their database consisted of labeling 6085 tweets, of which 3971 belong to the Non-Xenophobia class and 2114 to the Xenophobia class. Finally, to classify the tweets, they used Term Frequency–Inverse Document Frequency (TFIDF) [60] as their feature extraction method, and they also used word n-grams of length one to three and character n-grams of size one to four to create their tokens. They used Support Vector Machines (SVM) [61], Naïve–Bayes (NB) [62], and Logistic Regression (LR) [63] as their classifier models. They obtained 0.84 in the F1 score test, 0.87 in the recall, and 0.85 in precision.

Plaza-Del-Arco et al. [32] compared three different approaches to deal with Spanish hate speech on social networks. The first approach used supervised machine learning classifiers, while the second used deep learning techniques, and the last was performed using lexicon-based techniques. The problems addressed in their investigation were misogyny and Xenophobia classification in Twitter. To accomplish that, Plaza-Del-Arco et al. use a supervised machine learning approach using the Term Frequency–Inverse Document Frequency [60] jointly with the Naïve–Bayes [62], SupportVector Machines [61], Logistic Regression [63], Decision Tree, and Ensemble Voting (EV) machine learning classifiers. Furthermore, the FastText word embedding jointly with Recurrent Neural Networks (RNN) [64] and Long-Short-Term Memory (LSTM) [65] were used. Finally, the last approach used was to build an emotion lexicon dictionary made of words related to misogyny and Xenophobia. Finally, using the supervised machine learning approach, they obtained their best results 0.754 in the accuracy, 0.747 in precision, 0.739 in the recall, and 0.742 in the F1 score test. These results were obtained by using the Ensemble Voting classifier with unigrams and bigrams.

Charitidis et al. [66] proposed an ensemble of classifiers for the classification of tweets that threaten the integrity of journalists. They brought together a group of specialists to define which posts had a violent intention against journalists. Something worth noting is that they used five different Machine Learning models among which are: Convolutional Neural Network (CNN) [67], Skipped CNN (sCNN) [68], CNN+Gated Recurrent Unit (CNN+GRU) [69], Long-Short-Term Memory [65], and LSTM+Attention (aLSTM) [70]. Charitidis et al. used those models to create an ensemble and tested their architecture in different languages obtaining an F1 Score result of 0.71 for the German language and 0.87 for the Greek language. Finally, with the use of Recurrent Neural Networks [64] and Convolutional Neural Networks [67], they extracted essential features such as the word or character combinations and the word or character dependencies in sequences of words.

Pitsilis et al. [11] used Long-Short-Term Memory [65] classifiers to detect racist and sexist posts issued short posts, such as those found on the social network Twitter. Their innovation was to use a deep learning architecture using Word Frequency Vectorization (WFV) [11]. Finally, they obtained a precision of 0.71 for classifying racist posts and 0.76 for sexist posts. To train the proposed model, they collected a database of 16,000 tweets labeled as neutral, sexist, or racist.

Sahay et al. [71] proposed a model using NLP and Machine Learning techniques to identify comments of cyberbullying and abusive posts in social media and online communities. They proposed to use four classifiers: Logistic Regression [63], Support Vector Machines [61], Random Forest (RF) (RF, and Gradient Boosting Machine (GB) [72]. They concluded that SVM and gradient boosting machines trained on the feature stack performed better than logistic regression and random forest classifiers. Additionally, Sahay et al. used Count Vector Features (CVF) [71] and Term Frequency-Inverse Document Frequency [60] features.

Nobata et al. [12] focused on the classification of abusive posts as neutral or harmful, for which they collected two databases, both of which were obtained from Yahoo!. They used the Vowpal Wabbit regression model [73] that uses the following Natural Language Processing features: N-grams, Linguistic, Syntactic and Distributional Semantics (LS, SS, DS). By combining all of them, they obtained a performance of 0.783 in the F1-score test and 0.9055 AUC.

It is essential to highlight that all the investigations above collected their database; therefore, they are not comparable. A summary of the publications mentioned above can be seen in Table 1. The previously related works seek the classification of hate posts on social networks through Machine Learning models. These investigations have relatively similar results that range between 0.71 and 0.88 in the F1-Score test.

Beyond the performance that these classifiers can have, the problem of using black-box models is that we cannot be sure what factors determine whether a message is abusive. Today we need to understand the background of the behavior of ML models to make better decisions [33,74]. This is why this work takes on the characteristics of previous works but proposes a radical change in its intelligibility, offering experts in the field the possibility of having a transparent tool that helps them classify xenophobic posts and understand why these posts are considered in this way.

Table 1. Summary of previous work in terms of the problem they address, the data source used, features extracted, classifiers used, evaluation metrics, and the result obtained in the evaluation.

Author	Problem	Database Origin	Extracted Features	Methods	Evaluation Metrics	Performance
Pitropakis et al.	• Xenophobia	• Twitter	• Word n-grams • Char n-grams • TF-IDF	• LR • SVM • NB	• F1 • Rec • Prec	• 0.84 F1 • 0.87 Rec • 0.85 Prec
Plaza-Del-Arco et al.	• Misogyny and Xenophobia	• Twitter	• TF-IDF • FastText • Emotion lexicon	• LR • SVM • NB • Vote • DT • LSTM	• F1 • Rec • Prec • Acc	• 0.742 F1 • 0.739 Rec • 0.747 Prec • 0.754 Acc
Charitidis et al.	• Hate speech to journalists	• Wikipedia • Twitter • Facebook • Other	• Word or character combinations • Word or character dependencies in sequences of words	• CNN • sCNN • CNN + GRU • LSTM • aLSTM	• F1	• English: 0.82 • German: 0.71 • Spanish: 0.72 • Fr:ench 0.84 • Greek: 0.87
Pitsilis et al.	• Sexism • Racism	• Twitter	• Word Frequency Vectorization	• LSTM • RNN	• F1	• Sexism: 0.76 • Racism 0.71
Sahay et al.	• Cyberbullying	• Train: Twitter and YouTube • Test: Kaggle	• Count Vector Features • TF-IDF	• LR • SVM • RF	• AUC • Acc	• 0.779 AUC • 0.974 Acc
Nobata et al.	• Abusive language	• Yahoo! Finance and News	• N-grams • Linguistic semantics • Syntactic semantics • Distributional semantics	• Vowpal Wabbit's regression	• F1 • AUC	• 0.783 F1 • 0.906 AUC

4. Our Approach for Detecting Xenophobic Tweets

Our approach for Xenophobia detection in social networks consists of three steps: the Xenophobia database creation labeled by experts (Section 4.1); creating a new feature representation based on a combination of sentiments, emotions, intentions, relevant words, and syntactic features stemming from tweets (Section 4.2); and providing both contrast patterns describing Xenophobia texts and an explainable model for classifying Xenophobia posts (Section 4.3).

4.1. Creating the Xenophobia Database

For collecting our xenophobic database, we used the Twitter API [15] using the Tweepy Python library [75] implementation to filter the tweets by language, location, and keywords. The Twitter API offers free access to all Twitter data that the users generate, not only the text of the tweets that each user posts on Twitter, but also the user's information such as the number of followers, the date where the Twitter account was created, among others. Figure 2 shows the pipeline to develop our Xenophobia database.

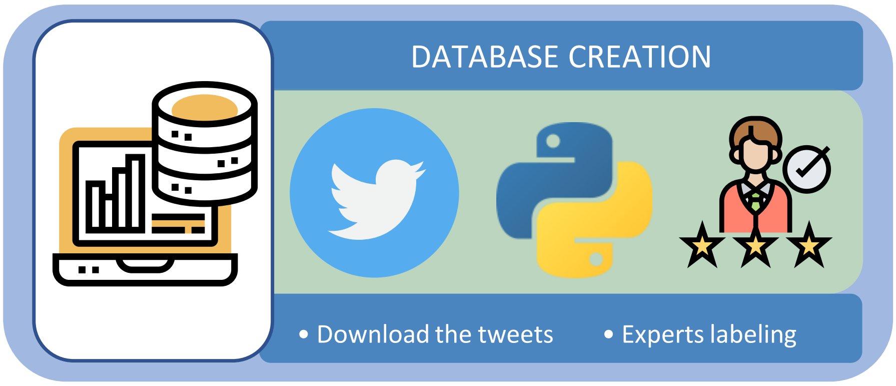


Figure 2. The creation of the Xenophobia database consisted of downloading tweets through the Twitter API jointly with the Python Tweepy library. Then, Xenophobia experts took it upon themselves to manually label the tweets.

We decided to keep only the raw text of each tweet to make a Xenophobia classifier based only on text. We made this decision to extrapolate this approach to other platforms because each social network has additional information that could not exist or is difficult to access on other platforms [76]. For example, detailed profile information as geopositioning, account creation date, preferred language; among others, are characteristics challenging to obtain (even not provided) in other social networks. In this way, the exclusion of additional information from the text allows focusing on its classification based solely on natural language processing techniques such as sentiment, semantic and syntactic analysis [77], which is more versatile for applying to any platform containing posts. As an additional configuration for obtaining the analyzed tweets, we used the (geo_search) Tweepy method with the parameters (query= "USA", granularity="country"); consequently, it allowed us to collect tweets issued from the USA and using the English language.

These data were collected in five weeks, from 27 June to 31 August 2021. The tweets publication date corresponds with the collection's date of the same. Each week 2000 tweets were downloaded. For the labeling process, we were supported by five experts. Two were psychologists, two were experts in international relations, and the last expert was a sociologist. These experts were in charge of labeling the tweets manually.

Since a single Twitter API return can return, at most, 100 tweets per looked term, we followed the same scheme used by Pitropakis et al. [59]. We used a set of keywords regarding Xenophobia instead of a single immigration term. Some of our xenophobic keywords were the same as the ones used by Pitropakis et al., such as *immigration*, *migrant*, and *deport them all*. While our experts proposed a new set of keywords, among which are: *illegal aliens*, *backcountry*, and *violent*. Nevertheless, we also used a set of neutral terms to make our database more diversified, such as sports, food, travel, love, money, among others. As a result, a total of 10,073 tweets were annotated.

The collected tweets were labeled in two categories where 8056 tweets were labeled as non-xenophobic, 2017 as xenophobic, where 79.97% of the labels correspond with the non-Xenophobia class and the remainder, 20.03%, belong to the Xenophobia class. Table 2 shows two random examples of tweets belonging to each class. Finally, our collected database was divided into 20 batches of 504 tweets, each one. Each expert was in charge of labeling four batches for a total of 2016 tweets. After the first labeling process, a second process was done by one of our experts in international relations. This second process was to inspect again all the tweets labeled as xenophobic and look for any discrepancy.

Table 2. Examples of tweets classified as xenophobic, non-xenophobic.

Class	Example
Non-xenophobic	No wonder why the 4Chan CHUDs have misunderstood the meaning of this movie and then made it their foundational text. https://t.co/96M7rHy3fc i just received the best text in the world i truly love my friends so fucking much
Xenophobic	@Jones17Charlene @FugginAlex The majority of the illegal alien migrant minors are not being brought by their real relatives anyways. Furthermore, if they are that sickly when they get here, I do not want them to make it into our cities. @learnin_as_i_go @orr_it You want a country of criminal foreigners all different colours smoking drugs being gay and living in a council block baby machine! No middle class just poor people!!!!

4.2. A New Feature Representation for Xenophobia Detection

When we classify data in a text format, and we want to obtain good results, we must have a good machine learning classifier and an appropriate feature representation [78]. The problem with the feature representations based on numeric terms of the frequency of the words is that they are not accessible for human understanding [79]. In this way, our feature representation proposal was created to obtain good classification results and an interpretable model. Figure 3 shows the pipeline to develop our new feature representation proposal.

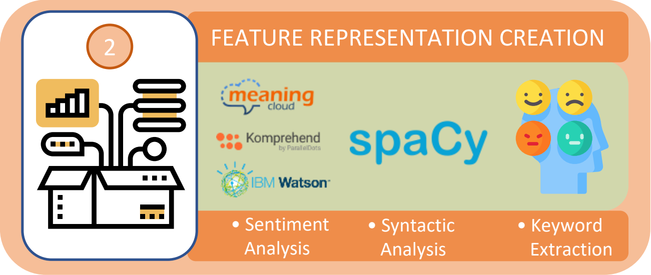


Figure 3. The creation of our feature representation proposal consisted of using three NLP APIs (Parallel Dots [80], Meaning Cloud [81], IBM NLU [82]) to perform sentiment analysis and the Python spaCy library [83] for extracting the syntactic features and the keywords of the tweets.

To create our new feature representation proposal and make it interpretable, we decided to base our representation on three main bases: sentiment analysis, syntactic analysis, and the extraction of keywords (semantic analysis) related to the Xenophobia class. The sentimental analysis has been proved to be a good tool for making text classification using interpretable models [84]. The use of sentiment, syntactic and semantic analysis improves the classification results compared to the use of only sentiment analysis [77,85]. The first step was to perform the sentiment analysis; Figure 3 shows the different cloud services that we use to perform the sentiment analysis. Table 3 shows a summary of the features extracted from the cloud services and the field to which they belong.

Table 3. Set of features extracted from different cloud services.

Source	Features	Field
Parallel Dots	Negative, Positive	Sentiment
Meaning cloud	Agreement, Score Tag	Sentiment
Parallel Dots	Bored, Sad, Happy, Excited	Emotion
IBM NLU	Joy, Disgust, Anger	Emotion
Parallel Dots	News, Spam, Marketing, Feedback, Complaint, Appreciation	Intent
Parallel Dots	Abusive, Hate-speech, Neither	Abusive content

The second step was to perform the syntactic analysis; we used the spaCy python library [83]. We employ the *en_core_web_lg* pipeline that is specifically designed for blogs, news, and comments. This pipeline is pre-trained on the OntoNotes Release 5.0 data [86]. With this library, we were enabled to extract different linguistic features using the spaCy Part-of-speech tagging implemented on its pipeline. For more information related to Universal POS tags, it can be visited the following document: https://universaldependencies.org/docs/u/pos/ (accessed on 15 April 2021). The linguistic features extracted from spaCy are:

- **ADJ:** (INT) The number of adjectives presented in the tweet.
- **AUX:** (INT) The number of auxiliary verbs presented in the tweet.
- **NUM:** (INT) The number of numbers presented in the tweet.
- **PROPN:** (INT) The number of proper nouns presented in the tweet.
- **ALPHAS:** (INT) The number of words presented in the tweet that are not stopwords.
- **HASHTAGS:** (INT) The number of hashtags presented in the tweet.
- **URLs:** (BOOL) If the tweet has a URL or not.

Finally, the third step was to extract the most representative xenophobic words of the database according to their frequency. Beyond the fact that our database was created using xenophobic keywords, not all the tweets with a xenophobic word were labeled as xenophobic. Furthermore, there are new terms related to Xenophobia that were not proposed at creating the database. The process to extract the new xenophobic keywords was as follows:

- **Clean the tweet:** To clean the tweet, we normalize the tweets by removing stopwords, unknown characters, numbers, URLs, user mentions, and then apply lemmatization. Lemmatization is a normalization technique [87], generally defined as "the transformation of all inflected word forms contained in a text to their dictionary look-up form" [88].
- **Get the frequency of the words:** For each class (Xenophobic and Not-xenophobic) it was generated a list of all the words that belong to the class, then it was counted the frequency of each term, and it was gotten a dictionary where the word was the key, and the frequency was the value.
- **Extract the xenophobic keywords:** After getting the frequency of the words, they were sorted by the highest to the lowest frequency, and it was selected only the 20 most used words. It was considered two conditions to determine if a comment might be regarded as a xenophobic keyword. The first condition: if the word only belongs to the xenophobic class, this means that the term is present in the 20 most used words list of the Xenophobia class and did not belong to the other list. The second condition: if the word is presented in both lists, but the absolute frequency of the word is more significant in the Xenophobia list than the non-Xenophobia list.

When we consider the proportion of the tweets that belong to the Xenophobia and no-Xenophobia class, we can realize that for each tweet that was labeled as xenophobic, there are four tweets labeled as non-xenophobic. If a word has the same use frequency in both classes, we can say that the word is four times more used in the xenophobic class. The above process was used again to obtain bigrams, sequences of two words that appear together or near each other. As a result, the following list of words was obtained. Five are unigrams, and five are bigrams: country, illegal, foreigners, alien, criminal, back country, illegal alien, violent foreigners, criminal foreigners, criminal migrant.

Table 4 shows the number of features grouped by different key labels for our INTER feature representation. In total, 37 features were used to construct our new feature representation proposal. Of which 20 were from the sentiment analysis, seven were extracted from the syntactic analysis, and the last ten were from the xenophobic keyword extraction process described above. Finally, Table 5 shows an example of two tweets extracted from EXD, one belonging to the non-Xenophobia class and the other to the Xenophobia class. These tweets were transformed using our interpretable feture representation and Table 6 shows each feature grouped by different key labels.

Table 4. Distribution of the features presented in our INTER feature representation. The overall column shows the total number of features.

Number of Features Grouped by Different Key Labels.						
Sentiment	**Emotion**	**Intent**	**Abusive Content**	**Xenophobia Keywords**	**Syntactic Features**	**Overall**
4	7	6	3	10	7	37

Table 5. Example of tweets belonging to the non-Xenophobia and Xenophobia class.

Class	Tweet
Non-Xenophobia	Immigrant families deserve to live without fear in Massachusetts, especially amid the #COVID19 pandemic. It's a moral imperative. Let us align our laws with our values! Pass the #SafeCommunitiesAct ASAP! @MassGovernor @KarenSpilka @SpeakerDeLeo #MALeg
Xenophobia	@EUTimesNET I do not know what liberal idiot runs your site but the USA is not a hellhole. We may have racist terrorists running around burning things but Europe has violent migrants raping women, vandalizing churches and attacking Christians. You're far from a model region.

Table 6. Example of our interpretable feature representation for tweets belonging to the Xenophobia and non-Xenophobia class grouped by different key labels.

(a) Sentiment features.

Class	Negative	Positive	Agreement	Score tag
Non-Xenophobia	0.202	0.458	AGREEMENT	Neutral
Xenophobia	0.707	0.094	DISAGREEMENT	Negative

(b) Emotion features.

Class	Bored	Sad	Happy	Excited	Joy	Disgust	Anger
Non-Xenophobia	0.0240	0.1341	0.2205	0.2371	0.1766	0.2272	0.1721
Xenophobia	0.0251	0.0566	0.0718	0.2963	0.0047	0.5043	0.4510

(c) Intent features.

Class	News	Spam	Marketing	Feedback	Complaint	Appreciation
Non-Xenophobia	0.6940	0.0410	0.0730	0.1770	Null	Null
Xenophobia	0.8260	0.0400	0.0480	0.0800	Null	Null

(d) Abusive content features.

Class	Abusive	Hate-speech	Neither
Non-Xenophobia	0.0092	0.8499	0.1408
Xenophobia	0.0005	0.9990	0.0005

(e) Unigram Xenophobia Keywords.

Class	Country	Illegal	Foreigners	Alien	Criminal
Non-Xenophobia	not present	not present	not present	not present	not present
Xenophobia	not present	not present	present	not present	not present

(f) Bigram Xenophobia Keywords.

Class	Back country	Illegal alien	Violent foreigners	Criminal foreigners	Criminal migrant
Non-Xenophobia	not present	not present	not present	not present	not present
Xenophobia	not present	not present	present	not present	not present

(g) Syntactic features.

Class	ADJ	AUX	NUM	PROPN	ALPHAS	HASHTAGS	URLs
Non-Xenophobia	2	1	0	5	19	not present	3
Xenophobia	3	0	0	2	20	not present	0

4.3. Mining Contrast Patterns

Nowadays, various understandable classifications models can bring us an explanation about the classification [89]. Nevertheless, from all the understandable classification models, the contrast pattern-based models are one of the most prominent [90], not only for the explanatory power but also because different studies have proved that these models provide good results for class imbalance problems [36,90]. Figure 4 shows the pipeline to extract the contrast patterns.

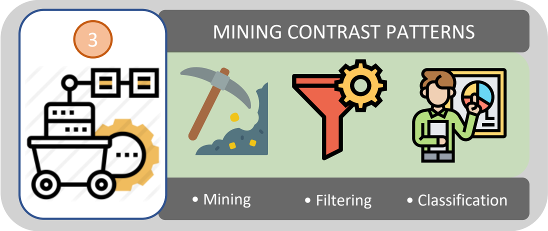


Figure 4. The extraction of the contrast patterns consist on three phases mining, filtering and classification.

According to Dong and Bailey [38], a pattern is a condition on data tuples that evaluates to either true or false. To be considered a pattern, the succinct state must be much simpler and smaller than the original length of the data. Ordinarily, a pattern is represented by a conjunction of relational statements, each with the form: [f_i # v_j], where v_j is a value within the space of feature f_i, and # is a relational operator taken from the set $\{ =, \neq, \leq, >, \in, \notin, \}$ [33,36,38]. For example, [*violent foreigners* = *"present"*] $\wedge$ [*hate−speech* ≥ 0.11], is a pattern describing post xenophobes.

In comparison, contrast patterns are a type of pattern whose supports differ significantly among the analyzed databases [38]. There are three steps to build a contrast pattern-based classifier: mining, filtering, and classification [3,33]:

- **Mining:** it is committed to looking for a set of candidate patterns by an exploratory examination using a search-space, characterized by a group of inductive limitations given by the user.
- **filtering:** it is committed to choosing a set of high-quality patterns from the mining stage; this step permits equal or superior results than using all the patterns extracted at the mining step.
- **Classification:** it is responsible for looking for the finest methodology to combine the data provided by a subset of patterns and construct an accurate model that is based on patterns.

We decided to use the Random Forest Miner (RFMiner) [91] as our algorithm for mining contrast patterns during the first step. García-Borroto et al. [92] conducted a large number of experiments comparing several well-known contrast pattern mining algorithms that are based on decision trees. According to the results obtained in their experiments, García-Borroto et al. have shown that RFMiner is capable of creating diversity of trees. This feature allows RFMiner to obtain more high-quality patterns compared to other known pattern miners. The filtering algorithms can be divided into two groups: based on set theory and based on quality measure [33]. For our filtering process, we start using the set theory approach. We remove redundant items from patterns and duplicated patterns. Furthermore, we choose only general patterns. After this filtering process, we kept the patterns with higher support.

Finally, we decided to use PBC4cip [36] as our contrast pattern-based classifier for the classification phase due to the good results that PBC4cip has reached in class imbalance problems. This classifier uses 150 trees by default; nevertheless, after many experiments classifying the patterns, we use only 15 trees, looking for the simplest model with good classification results in the AUC score metric. We repeated this process, reducing the number of trees and minimizing the AUC loss and the number of trees. A stop criterion was executed when the AUC score obtained in our experiments was more than 1% compared with the results that PBC4Cip reaches with the default number of trees.

5. Experimental Setup

This section shows the methodology designed to evaluate the performance of the tested classifiers. For our experiments, we use two databases: our Experts Xenophobia Database (EXD), which consists of 10,057 tweets labeled by experts in the fields of international relations, sociologists, and psychologists. Additionally, we use the Xenophobia database created by Pitropakis et al. [59]; for this article, we will refer to this database as Pitropakis Xenophobia Database (PXD). Table 7 shows the number of tweets per class for the PXD and EXD databases before and after applying the cleaning method. Figure 5 shows the flow diagram to obtain our experimental results. The flow diagram starts from getting each database and then transforming it using different feature representations and finishing bringing the performance of each classifier. Below, we will briefly explain what each of the steps in the said figure consists of:

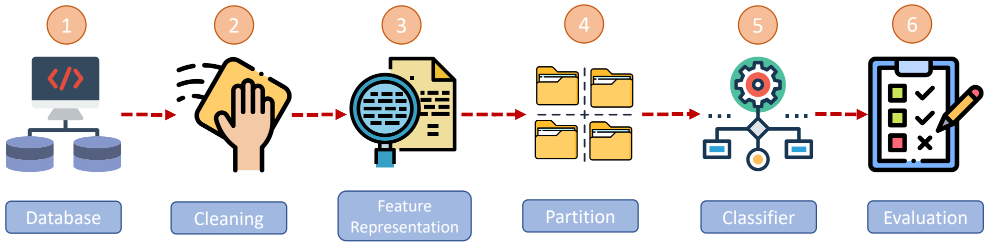


Figure 5. Flow diagram for the procedure of getting the classification results of the Xenophobia databases.

1. **Database:** The first step consisted of obtaining the Xenophobia databases used to train and validate all the tested machine learning classifiers detailed in step number five.
2. **Cleaning:** For each database, our proposed cleaning method was used to obtain a clean version of the database. Our cleaning method was specially designed to work with databases made on Twitter. It removes unknown characters, hyperlinks, retweet text, and user mentions. Additionally, our cleaning method converts the most used slang to their original word, removes stop words, and applies lemmatization. Finally, it removes tweets that are identical before and after applying the normalization process described above.
3. **Feature representation:** For each database, different feature representations were used to convert tweets into representations used by the tested machine learning classifiers. We use the following well-known feature representations: Bag Of Words (BOW) [93], Term Frequency-Inverse Document Frequency (TFIDF) [60], Word To Vector (W2V) [94], and our interpretable proposal (INTER).
4. **Partition:** For each feature representation, five partitions were generated. Each partition was performed using the Distribution Optimally Balanced Stratified Cross-Validation (DOB-SCV) method [95]. According to Zeng and Martinez [95], the principal advantage of DOB-SCV is that it keeps a better distribution balance in the feature space when splitting a sample into groups called folds. This property empowers the cross-validation training set better to capture the distribution features within the actual data set.
5. **Classifier:** For each partition, the following machine learning classifiers were used: C4.5 (C45) [96], k-Nearest Neighbor (KNN) [97], Rusboost (RUS) [98], UnderBagging (UND) [99], and PBC4cip [36]. Except for KNN, the other classifiers are based on decision trees. The classifiers mentioned above have been implemented in the KEEL software [100], except for PBC4cip, which is a package available for the Weka Data-Mining software tool [101]; it can be taken from https://sites.google.com/view/leocanetesifuentes/software/multivariate-pbc4cip (accessed on 20 October 2020).
6. **Evaluation:** For each classifier, we used the following performance evaluations metrics: F1 score and Area Under the ROC Curve (AUC) [102]. These metrics are widely used in the literature for class imbalance problems [103,104].

Table 7. Comparison between the number of tweets belonging to the non-xenophobic and xenophobic classes before and after using the cleaning method. The class imbalance ratio (IR) is calculated as the proportion between the number of objects belonging to the majority class and the number of objects belonging to the minority class [36]. The higher the IR value, the more imbalanced the database is.

	Before Cleaning Method				**After Cleaning Method**			
Database	**No Xenophobia**	**Xenophobia**	**Total**	**IR**	**No Xenophobia**	**Xenophobia**	**Total**	**IR**
PXD	3971	2114	6085	1.88	3826	1988	5814	1.92
EXD	8056	2017	10,073	3.99	8054	2003	10,057	4.02

On the one hand, our INTER feature representation method proposal is designed to be interpretable and provide a set of feelings, emotions, and keywords from a given text. On the other hand, the feature representation BOW, TFIDF, and W2V transform an input text into a numeric vector [105]. According to Luo et al. [79] these numeric transformations are considered black-box and prevent them from being human-readable.

We can also mention that there are methods based on neural networks built from the numeric feature representation methods achieving interpretable results [106]. On the one hand, the interpretability of the neural networks is based on highlighting the keywords that a text has to belong to a class [106]; on the other hand, our approach seeks to obtain more interpretability features such as feelings, emotions, and intentions; this can allow an expert to understand why a text is considered to be xenophobic with more detail.

Table 8 shows a summary of the information presented above, synthesizing which classifiers and feature representation are interpretable and which classifiers use contrast patterns or not. From Table 8a, and the C4.5's definition stated by Ting et al. [96], García et al. [107], and Dong and Bailey [38] (see Section 4 for more detail), we can comment that the tree-based classifiers are interpretable; however, only PBC4cip uses contrast patterns. Finally, Table 8b shows that the only feature representation being interpretable is our INTER feature representation proposal.

Table 8. Summary of the characteristics of the classifiers and the interpretability of the feature representations.

(a) Characteristics of the classifiers.

Classifier	**Is interpretable?**	**Is it contrast pattern-based?**
C45	✓	✗
KNN	✗	✗
RUS	✓	✗
UND	✓	✗
PBC4cip	✓	✓

(b) Interpretability of the feature representations.

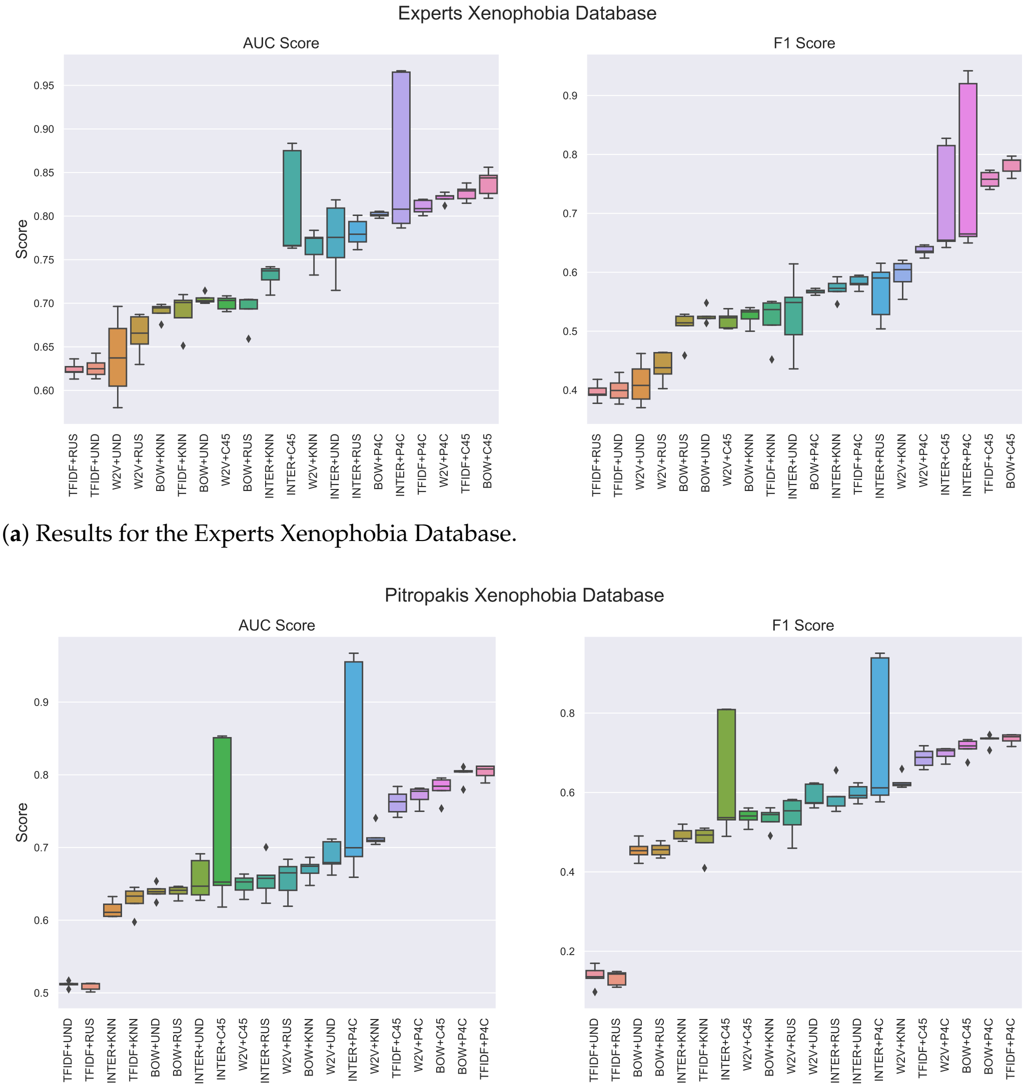
Feature representation	**Is interpretable?**
BOW	✗
TFIDF	✗
W2V	✗
INTER	✗

6. Experimental Results and Discussion

For a better understanding of our experiment results, we have split this section into two subsections: in Section 6.1, we show all the classification results for both metrics, Area Under the Curve (AUC) and F1 score, and in Section 6.2, we present an analysis of the obtained patterns describing the Xenophobia class.

6.1. Classification Results

Using the methodology proposed in Section 5, we can analyze the classification results obtained on both EXD and PXD databases. Figure 6 show box-and-whisker plots for both databases regarding AUC and F1 score metrics. The box-and-whiskers plot, also known as a boxplot, is a chart used in descriptive data analysis [108].



(**a**) Results for the Experts Xenophobia Database.

(**b**) Results for the Pitropakis Xenophobia Database.

Figure 6. Box-and-whisker plots for the AUC and F1 score metrics. The boxes are sorted in ascending order according to their median.

The boxplots are very useful to compare the distribution between many groups where each box of the boxplot represents the distribution of a group. In our case, each box represents the combination's results for both AUC and F1 score metrics presented in Table 9. Boxplots show the next five-number summary of a group:

- **The minimum score:** is the lowest score present in the set, excluding outliers. In the chart, it is represented as the line below the box.
- **Lower quartile:** also known as the first quartile or Q1, the lower quartile is a line where 25% of the scores fall below this value. In the chart, it is represented as the bottom line of the box.
- **Median**: also known as second quartile or Q2, the median is a line where half of the scores are less than this value, and half are greater. In the cart, it is represented as the middle line of the box.
- **Upper quartile**: also known as third quartile or Q3, the upper quartile is a line where 75% of the scores fall below this value. In the chart, it is represented as the upper line of the box.
- **Maximum score**: is the highest score present in the set, excluding outliers. In the chart, it is represented as the line above the box.

In Figure 6, the best combination of embedding method and classifier is the one that has more score in the median. On the one hand, in Figure 6a, when EXD is used, the combination with a higher median is BOW+C45 for both AUC and F1 metric scores. On the other hand, Figure 6b shows that for PXD, the best combination is TFIDF+P4C. It is worth mentioning that the best combinations of embedding methods and classifiers that maximize the AUC and F1 score results for EXD and PXD are not interpretable.

Taking into account all the interpretable combinations that are presented in the Table 8 and comparing their results presented in Figure 6, we can see that for both databases, the combination INTER+P4C is the one providing the best score metric results.

Table 9 shows the results of each partition for both databases regarding AUC and F1 score metrics. We highlight the INTER+P4C results in Table 9a,b. In this way, we can compare each database's differences easily when using the same combination. It is essential to compare the average (AVG) results and their standard deviation (STD); this allows us to understand the differences between each database.

From Table 9, we can obtain the following observations:

- For EXD, the combinations with the lowest STD in both metrics are W2V+P4C and BOW+P4C, while for PXD, they are TFIDF+P4C and TDIDF+RUS.
- For EXD, the combinations with the highest AVG in both metrics are INTER+P4C and BOW+C45, while for PXD, they are TFIDF+C45, BOW+P4C.
- For EXD, the best combination in terms of AVG and STD is BOW+C45, while for PXD, it is TFIDF+P4C. The difference between these models is, on the one hand, that BOW+C45 in EDX achieves 0.040 more than TFIDF+P4C in both AUC and F1 score metrics. On the other hand, TFIDF+P4C in PDX has approximately 0.003 fewer STD than BOW+C45 in EXD.

Table 10 shows a comparison of results between EXD and PDX for the combinations of the embedding method and classifier that are interpretable. When we compare both AVG and STD for the interpretable combinations present in Table 10 we can observe the following characteristics.

- The best interpretable combinations are INTER+P4C and INTER+C45 for both databases. INTER+P4C shows a difference of 0.050 or greater in both AUC and F1 scores than INTER+C45 in EXD. This difference is greater in PXD.
- On the one hand, on AVG, the results obtained in AUC are higher in EXD than in PXD. On the other hand, the AVG results for the F1 score are higher in EXD only for the P4C and C45 classifiers.
- The STD in EXD for the INTER+P4C and INTER+C45 combinations is less than in PXD for both AUC and F1 score metrics.
- The combination with the lowest STD is INTER+UND using PXD, while the combination with the highest STD is INTER+P4C using PXD.

With the above observations, we can conclude that the results obtained by INTER+P4C are better than the other interpretable combinations. Furthermore, INTER+P4C has a higher AVG, and it is more robust according to their STD in EXD than in PXD.

Notice that Table 11 shows very relevant information for the analysis of results. In this table, we can observe exceptional characteristics. Let us first analyze the results obtained by EXD (see Table 11a).

- In two of the five partitions, the INTER+P4C configuration had the best results for both AUC and F1 scores (partition one and partition two).
- In terms of F1 score, INTER+P4C always placed in the top five positions, while for AUC, there were two occasions in which it was in seven places.
- When we compare the best partition of INTER+P4C, partition number two, against the best partition of BOW+C45, partition one. We can see that the best results of INTER+P4C are 0.11 AUC points and 0.15 F1 score points superior.

- For all the embedding methods, the P4C classifier is always within the top ten of 20 regardless of partition or metric. Not only that, within the top five of each partition of each metric, there are always at least two combinations of embedding methods jointly with a classifier that uses P4C as a classifier.

Table 9. Results of the AUC and F1 score metrics. Each row represents the combination of an embedding method jointly with a classifier. In AVG the best interpretable combination is **INTER+P4C**.

(a) Classification results for the Experts Xenophobia Database.

COMBINATION	Partition 1		Partition 2		Partition 3		Partition 4		Partition 5		AVG		STD	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
BOW+C45	0.856	0.791	0.847	0.790	0.820	0.759	0.844	0.797	0.826	0.772	0.839	0.782	0.013	0.014
BOW+KNN	0.696	0.536	0.695	0.533	0.699	0.540	0.689	0.521	0.675	0.500	0.691	0.526	0.008	0.014
BOW+P4C	0.802	0.568	0.800	0.565	0.798	0.561	0.804	0.573	0.805	0.569	0.802	0.567	0.003	0.004
BOW+RUS	0.704	0.514	0.704	0.525	0.659	0.459	0.704	0.529	0.694	0.509	0.693	0.507	0.017	0.025
BOW+UND	0.706	0.514	0.703	0.525	0.702	0.522	0.714	0.548	0.700	0.522	0.705	0.526	0.005	0.012
INTER+C45	0.883	0.827	0.875	0.815	0.766	0.653	0.763	0.642	0.766	0.654	0.811	0.718	0.056	0.084
INTER+KNN	0.737	0.573	0.742	0.581	0.727	0.567	0.740	0.592	0.709	0.546	0.731	0.572	0.012	0.015
INTER+P4C	0.965	0.920	0.967	0.942	0.786	0.650	0.792	0.665	0.808	0.661	0.864	0.768	0.084	0.134
INTER+RUS	0.770	0.528	0.761	0.504	0.779	0.600	0.794	0.615	0.801	0.590	0.781	0.568	0.015	0.043
INTER+UND	0.752	0.494	0.715	0.436	0.819	0.614	0.776	0.557	0.809	0.549	0.774	0.530	0.038	0.061
TFIDF+C45	0.838	0.773	0.829	0.758	0.831	0.769	0.820	0.746	0.815	0.741	0.827	0.757	0.008	0.013
TFIDF+KNN	0.651	0.452	0.703	0.548	0.710	0.550	0.683	0.510	0.701	0.537	0.690	0.519	0.021	0.037
TFIDF+P4C	0.805	0.579	0.818	0.592	0.819	0.595	0.809	0.581	0.800	0.568	0.810	0.583	0.007	0.010
TFIDF+RUS	0.636	0.418	0.621	0.393	0.613	0.378	0.627	0.403	0.621	0.391	0.624	0.397	0.008	0.014
TFIDF+UND	0.643	0.430	0.625	0.399	0.613	0.376	0.632	0.412	0.618	0.387	0.626	0.401	0.010	0.019
W2V+C45	0.690	0.506	0.706	0.526	0.708	0.538	0.703	0.523	0.694	0.504	0.700	0.519	0.007	0.013
W2V+KNN	0.756	0.584	0.784	0.620	0.776	0.615	0.732	0.554	0.775	0.605	0.764	0.596	0.018	0.024
W2V+P4C	0.820	0.636	0.823	0.644	0.827	0.646	0.823	0.633	0.812	0.624	0.821	0.637	0.005	0.008
W2V+RUS	0.687	0.464	0.653	0.427	0.630	0.403	0.666	0.438	0.684	0.463	0.664	0.439	0.021	0.023
W2V+UND	0.671	0.436	0.580	0.370	0.605	0.385	0.637	0.408	0.696	0.462	0.638	0.412	0.042	0.033

(b) Classification results for the Pitropakis Xenphobia Database.

COMBINATION	Partition 1		Partition 2		Partition 3		Partition 4		Partition 5		AVG		STD	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
BOW+C45	0.754	0.675	0.796	0.733	0.793	0.729	0.778	0.710	0.784	0.717	0.781	0.713	0.015	0.021
BOW+KNN	0.674	0.544	0.686	0.561	0.677	0.549	0.665	0.526	0.648	0.491	0.670	0.534	0.013	0.025
BOW+P4C	0.780	0.706	0.811	0.745	0.805	0.737	0.804	0.735	0.806	0.736	0.801	0.732	0.011	0.013
BOW+RUS	0.627	0.435	0.647	0.478	0.645	0.467	0.636	0.443	0.641	0.456	0.639	0.456	0.007	0.016
BOW+UND	0.624	0.421	0.654	0.490	0.643	0.464	0.639	0.453	0.636	0.443	0.639	0.454	0.010	0.023
INTER+C45	0.851	0.809	0.853	0.810	0.648	0.531	0.652	0.537	0.618	0.490	0.725	0.635	0.105	0.143
INTER+KNN	0.633	0.520	0.622	0.504	0.611	0.484	0.605	0.477	0.605	0.483	0.615	0.494	0.011	0.016
INTER+P4C	0.967	0.951	0.955	0.939	0.700	0.612	0.659	0.576	0.688	0.593	0.794	0.734	0.137	0.172
INTER+RUS	0.700	0.656	0.623	0.552	0.644	0.566	0.658	0.590	0.662	0.590	0.657	0.591	0.025	0.036
INTER+UND	0.627	0.593	0.635	0.586	0.647	0.571	0.691	0.624	0.682	0.615	0.656	0.598	0.026	0.019
TFIDF+C45	0.741	0.658	0.763	0.689	0.773	0.704	0.784	0.718	0.749	0.669	0.762	0.687	0.015	0.022
TFIDF+KNN	0.598	0.410	0.633	0.493	0.640	0.505	0.623	0.474	0.645	0.510	0.628	0.478	0.017	0.037
TFIDF+P4C	0.789	0.716	0.812	0.745	0.808	0.740	0.799	0.730	0.812	0.745	0.804	0.735	0.009	0.011
TFIDF+RUS	0.513	0.146	0.505	0.115	0.513	0.149	0.513	0.143	0.501	0.109	0.509	0.132	0.005	0.017
TFIDF+UND	0.517	0.170	0.512	0.132	0.511	0.151	0.513	0.136	0.505	0.097	0.512	0.137	0.004	0.024
W2V+C45	0.642	0.531	0.629	0.507	0.663	0.561	0.658	0.553	0.653	0.541	0.649	0.539	0.012	0.019
W2V+KNN	0.713	0.625	0.741	0.659	0.710	0.620	0.704	0.613	0.708	0.617	0.715	0.627	0.013	0.017
W2V+P4C	0.750	0.671	0.777	0.705	0.780	0.710	0.766	0.691	0.782	0.710	0.771	0.698	0.012	0.015
W2V+RUS	0.619	0.460	0.641	0.519	0.684	0.582	0.674	0.579	0.665	0.554	0.657	0.539	0.023	0.046
W2V+UND	0.679	0.573	0.677	0.574	0.708	0.624	0.662	0.561	0.712	0.621	0.688	0.591	0.019	0.026

Table 10. Comparison of the interpretable combinations in EXD and PXD.

Experts Xenophobia Database					Pitropakis Xenophobia Database				
COMBINATION	AVG		STD		COMBINATION	AVG		STD	
	AUC	F1	AUC	F1		AUC	F1	AUC	F1
INTER+P4C	0.864	0.768	0.084	0.134	INTER+P4C	0.794	0.734	0.137	0.172
INTER+C45	0.811	0.718	0.056	0.084	INTER+C45	0.725	0.635	0.105	0.143
INTER+RUS	0.781	0.568	0.015	0.043	INTER+RUS	0.657	0.591	0.025	0.036
INTER+UND	0.774	0.530	0.038	0.061	INTER+UND	0.656	0.598	0.026	0.019

Table 11. Ranking of the AUC and F1 score metrics. Each cell represents the combination of an embedding method jointly with a classifier. The best interpretable combination is **INTER+P4C**.

(a) Ranks for the Experts Xenophobia Database.

RANK	Partition 1		Partition 2		Partition 3		Partition 4		Partition 5	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
1	**INTER+P4C**	**INTER+P4C**	**INTER+P4C**	**INTER+P4C**	TFIDF+C45	TFIDF+C45	BOW+C45	BOW+C45	BOW+C45	BOW+C45
2	INTER+C45	INTER+C45	INTER+C45	INTER+C45	W2V+P4C	BOW+C45	W2V+P4C	TFIDF+C45	TFIDF+C45	TFIDF+C45
3	BOW+C45	BOW+C45	BOW+C45	BOW+C45	BOW+C45	INTER+C45	TFIDF+C45	**INTER+P4C**	W2V+P4C	**INTER+P4C**
4	TFIDF+C45	TFIDF+C45	TFIDF+C45	TFIDF+C45	TFIDF+P4C	**INTER+P4C**	TFIDF+P4C	INTER+C45	INTER+UND	INTER+C45
5	W2V+P4C	W2V+P4C	W2V+P4C	W2V+P4C	INTER+UND	W2V+P4C	BOW+P4C	W2V+P4C	**INTER+P4C**	W2V+P4C
6	TFIDF+P4C	W2V+KNN	TFIDF+P4C	W2V+KNN	BOW+P4C	W2V+KNN	INTER+RUS	INTER+RUS	BOW+P4C	W2V+KNN
7	BOW+P4C	TFIDF+P4C	BOW+P4C	TFIDF+P4C	**INTER+P4C**	INTER+UND	**INTER+P4C**	INTER+KNN	INTER+RUS	INTER+RUS
8	INTER+RUS	INTER+KNN	W2V+KNN	INTER+KNN	INTER+RUS	INTER+RUS	INTER+UND	TFIDF+P4C	TFIDF+P4C	BOW+P4C
9	W2V+KNN	BOW+P4C	INTER+RUS	BOW+P4C	W2V+KNN	TFIDF+P4C	INTER+C45	BOW+P4C	W2V+KNN	TFIDF+P4C
10	INTER+UND	BOW+KNN	INTER+KNN	TFIDF+KNN	INTER+C45	INTER+KNN	INTER+KNN	INTER+UND	INTER+C45	INTER+UND
11	INTER+KNN	INTER+RUS	INTER+UND	BOW+KNN	INTER+KNN	BOW+P4C	W2V+KNN	W2V+KNN	INTER+KNN	INTER+KNN
12	BOW+UND	BOW+RUS	W2V+C45	W2V+C45	TFIDF+KNN	TFIDF+KNN	BOW+UND	BOW+UND	TFIDF+KNN	TFIDF+KNN
13	BOW+RUS	BOW+UND	BOW+RUS	BOW+RUS	W2V+C45	BOW+KNN	BOW+RUS	BOW+RUS	BOW+UND	BOW+UND
14	BOW+KNN	W2V+C45	TFIDF+KNN	BOW+UND	BOW+UND	W2V+C45	W2V+C45	W2V+C45	W2V+UND	BOW+RUS
15	W2V+C45	INTER+UND	BOW+UND	INTER+RUS	BOW+KNN	BOW+UND	BOW+KNN	BOW+KNN	W2V+C45	W2V+C45
16	W2V+RUS	W2V+RUS	BOW+KNN	INTER+UND	BOW+RUS	BOW+RUS	TFIDF+KNN	TFIDF+KNN	BOW+RUS	BOW+KNN
17	W2V+UND	TFIDF+KNN	W2V+RUS	W2V+RUS	W2V+RUS	W2V+RUS	W2V+RUS	W2V+RUS	W2V+RUS	W2V+RUS
18	TFIDF+KNN	W2V+UND	TFIDF+UND	TFIDF+UND	TFIDF+UND	W2V+UND	W2V+UND	TFIDF+UND	BOW+KNN	W2V+UND
19	TFIDF+UND	TFIDF+UND	TFIDF+RUS	TFIDF+RUS	TFIDF+RUS	TFIDF+RUS	TFIDF+UND	W2V+UND	TFIDF+RUS	TFIDF+RUS
20	TFIDF+RUS	TFIDF+RUS	W2V+UND	W2V+UND	W2V+UND	TFIDF+UND	TFIDF+RUS	TFIDF+RUS	TFIDF+UND	TFIDF+UND

(b) Ranks for the Pitropakis Xenophobia Database.

RANK	Partition 1		Partition 2		Partition 3		Partition 4		Partition 5	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
1	**INTER+P4C**	**INTER+P4C**	**INTER+P4C**	**INTER+P4C**	TFIDF+P4C	TFIDF+P4C	BOW+P4C	BOW+P4C	TFIDF+P4C	TFIDF+P4C
2	INTER+C45	INTER+C45	INTER+C45	INTER+C45	BOW+P4C	BOW+P4C	TFIDF+P4C	TFIDF+P4C	BOW+P4C	BOW+P4C
3	TFIDF+P4C	TFIDF+P4C	TFIDF+P4C	BOW+P4C	BOW+C45	BOW+C45	TFIDF+C45	TFIDF+C45	BOW+C45	BOW+C45
4	BOW+P4C	BOW+P4C	BOW+P4C	TFIDF+P4C	W2V+P4C	W2V+P4C	BOW+C45	BOW+C45	W2V+P4C	W2V+P4C
5	BOW+C45	BOW+C45	BOW+C45	BOW+C45	TFIDF+C45	TFIDF+C45	W2V+P4C	W2V+P4C	TFIDF+C45	TFIDF+C45
6	W2V+P4C	W2V+P4C	W2V+P4C	W2V+P4C	W2V+KNN	W2V+UND	W2V+KNN	INTER+UND	W2V+UND	W2V+UND
7	TFIDF+C45	TFIDF+C45	TFIDF+C45	TFIDF+C45	W2V+UND	W2V+KNN	INTER+UND	W2V+KNN	W2V+KNN	W2V+KNN
8	W2V+KNN	INTER+RUS	W2V+KNN	W2V+KNN	**INTER+P4C**	**INTER+P4C**	W2V+RUS	INTER+RUS	**INTER+P4C**	INTER+UND
9	INTER+RUS	W2V+KNN	BOW+KNN	INTER+UND	W2V+RUS	W2V+RUS	BOW+KNN	W2V+RUS	INTER+UND	**INTER+P4C**
10	W2V+UND	INTER+UND	W2V+UND	W2V+UND	BOW+KNN	INTER+UND	W2V+UND	**INTER+P4C**	W2V+RUS	INTER+RUS
11	BOW+KNN	W2V+UND	BOW+UND	BOW+KNN	W2V+C45	INTER+RUS	**INTER+P4C**	W2V+UND	INTER+RUS	W2V+RUS
12	W2V+C45	BOW+KNN	BOW+RUS	INTER+RUS	INTER+C45	W2V+C45	W2V+C45	W2V+C45	W2V+C45	W2V+C45
13	INTER+KNN	W2V+C45	W2V+RUS	W2V+RUS	INTER+UND	BOW+KNN	INTER+RUS	INTER+C45	BOW+KNN	TFIDF+KNN
14	INTER+UND	INTER+KNN	INTER+UND	W2V+C45	BOW+RUS	INTER+C45	INTER+C45	BOW+KNN	TFIDF+KNN	BOW+KNN
15	BOW+RUS	W2V+RUS	TFIDF+KNN	INTER+KNN	INTER+RUS	TFIDF+KNN	BOW+UND	INTER+KNN	BOW+RUS	INTER+C45
16	BOW+UND	BOW+RUS	W2V+C45	TFIDF+KNN	BOW+UND	INTER+KNN	BOW+RUS	TFIDF+KNN	BOW+UND	INTER+KNN
17	W2V+RUS	BOW+UND	INTER+RUS	BOW+UND	TFIDF+KNN	BOW+RUS	TFIDF+KNN	BOW+UND	INTER+C45	BOW+RUS
18	TFIDF+KNN	TFIDF+KNN	INTER+KNN	BOW+RUS	INTER+KNN	BOW+UND	INTER+KNN	BOW+RUS	INTER+KNN	BOW+UND
19	TFIDF+UND	TFIDF+UND	TFIDF+UND	TFIDF+UND	TFIDF+RUS	TFIDF+UND	TFIDF+UND	TFIDF+RUS	TFIDF+UND	TFIDF+RUS
20	TFIDF+RUS	TFIDF+RUS	TFIDF+RUS	TFIDF+RUS	TFIDF+UND	TFIDF+RUS	TFIDF+RUS	TFIDF+UND	TFIDF+RUS	TFIDF+UND

When we analyze the results obtained on PXD (see Table 11b), we can observe a behavior very similar to that described above for the INTER+P4C combination. The main feature that PXD share with EXD is that the first two partitions have the best AUC and F1 scores. However, in the following partitions, a lower position is observed in the position obtained by INTER+P4C. This agrees with the STD shown in Table 9. Considering the rank distribution obtained by INTER+P4C in both AUC and F1 score metrics, it can be concluded that EXD Table 11a, has more consistent results than those obtained in PXD (see Table 11b).

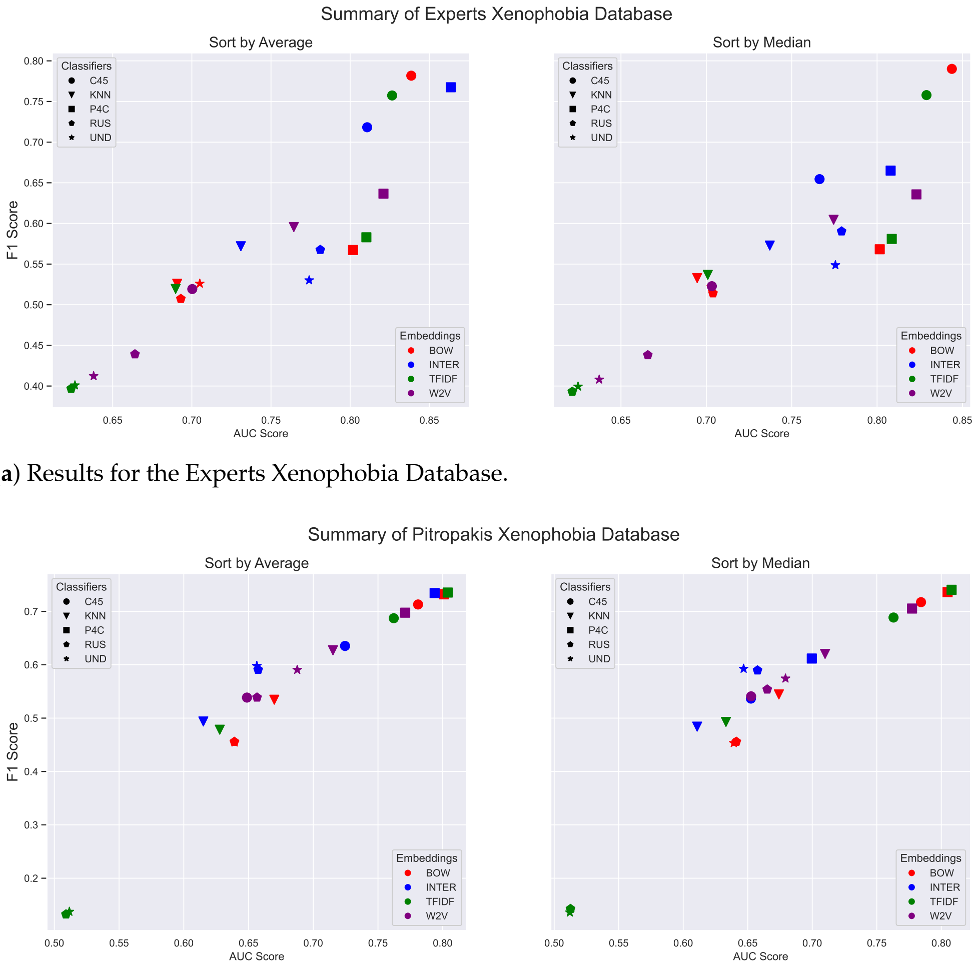
Finally, Figure 7 seeks to synthesize the results presented in this section. We can easily see that each point represents a combination embedding method (color) jointly with a classifier (shape) with these graphs. It obtains the best results for the AUC and F1 scores. It is important to note that each subgraph in Figure 7 shows the embedding method jointly with classifier results regarding AUC and F1 score metrics for EDX and PDX databases. We also see that each subgraph, in turn, contains two graphs.

On the one hand, in Figure 7a the left section, each embedding method jointly with a classifier combination is ordered according to the AVG result obtained for the AUC and F1 score. On the other hand, in Figure 7a the right section, each combination is ordered by the median. Additionally, we can see that each subgraph has the same scale for both the F1 score and AUC score; with this, we can see that the closer the points are to each other, the embedding method jointly with a classifier combination is more robust.

For example, in Figure 7a, the green circle represents TFIDF+C45. Both in the graphs, orders by AVG and by median are in a very similar position. This means that TFIDF+C45 is a very reliable combination, offering promising results for both metrics.

In both Figure 7a,b, regardless of the embedding method, the P4C classifier generally obtains good results this classifier shows to get better results in the AUC metric than for the

F1 score. Nevertheless, the classifier C45 also has good results for both AVG and median but works best for the embeddings BOW and TFIDF than for INTER and W2V.



(**a**) Results for the Experts Xenophobia Database.

(**b**) Results for the Pitropakis Xenophobia Database.

Figure 7. The color represents the embedding method, while the shape represents the classifier. The *X*-axis is the result of the AUC score. The *Y*-axis is the result of the F1 score. The graphs are ordered by mean and median according to the results of Table 9.

6.2. Extracted Patterns

This section discusses the interpretable contrast patterns obtained from the Expert Xenophobic database. The combination INTER+P4C extract better contrast patterns in terms of support in EDX than PXD. For this reason, we decided to use the contrast patterns from EDX. In Table 12, we can see ten representative contrast patterns. Five belong to the Xenophobia class, and five belong to the non-Xenophobia class. These patterns are arranged in descending order by their support.

According to Loyola-González et al. [3], the contrast pattern-based classifiers provide a model that is easy for a human to understand. The readability of the contrast patterns is very wide as they have few items. The first observations we can make about Table 12 shows the Xenophobia class's contrast patterns having slightly more support than for the non-Xenophobia class. The patterns describing the Xenophobia class are more straightforward in terms of many items than the patterns for the non-Xenophobia class. It is important to note that the patterns describing the Xenophobia class are formed by the presence of a negative feeling or emotion and a keyword.

Table 12. Example of contrast patterns extracted from the Experts Xenophobic Database.

Class	ID	Items	Supp
Xenophobic	CP_1	[*foreigners* = *"present"*] $\wedge$ [*disgust* > 0.15]	0.12
	CP_2	[*illegal* = *"present"*] $\wedge$ [*angry* > 0.19] $\wedge$ *hashtags* = *"not present"* $\wedge$ [*foreigners* = *"present"*]	0.11
	CP_3	[*foreigners* = *"present"*] $\wedge$ [*sad* $\leq$ 0.15]	0.10
	CP_4	[*angry* > 0.17] $\wedge$ [*violentForeigners* = *"present"*]	0.07
	CP_5	[*criminalForeigners* = *"present"*]	0.06
Non-Xenophobic	CP_6	[*positive* > 0.53] $\wedge$ [*joy* > 0.44] $\wedge$ [*negative* < 0.11] $\wedge$ [*hate-speech* $\leq$ 0.04]	0.09
	CP_7	[*angry* $\leq$ 0.17] $\wedge$ [*hate-speech* $\leq$ 0.06] $\wedge$ *negative* < 0.10 $\wedge$ [*country* = *"not present"*]	0.08
	CP_8	[*illegal* = *"not present"*] $\wedge$ [*foreigners* = *"not present"*] $\wedge$ [*backCountry* = *"not present"*] $\wedge$ [*joy* > 0.42]	0.08
	CP_9	[*positive* > 0.53] $\wedge$ [*angry* $\leq$ 0.13] $\wedge$ [*spam* $\leq$ 0.56] $\wedge$ [*ALPHAS* > 9.50]	0.06
	CP_{10}	[*hate-speech* $\leq$ 0.11] $\wedge$ [*foreigners* = *"not present"*]	0.05

Combining a keyword plus a sentiment or intention is crucial since we can contextualize the keyword and extract the word's true meaning. On the one hand, the CP_4 pattern shows us how the bigram "violent foreigners" has 0.07 support for the Xenophobia classification when the emotion that accompanies the text has at least a little anger. On the other hand, the CP_5 pattern is significant since it shows that even without the need for an associated feeling or emotion, the bigram "criminal foreigners" has the support of 0.06 of the Xenophobia class, this means that when this set of words is present is an excellent indicator for detecting Xenophobia.

The contrast patterns obtained for the non-Xenophobia class have more items than for the non-Xenophobia class. Only CP_{10} has two items, while the others have four items. Coherently concerning xenophobic patterns, the patterns extracted for the non-Xenophobia class are more associated with positive feelings and emotions or with very little presence of negative feelings, emotions, or intentions.

The CP_6 and CP_9 patterns are essential since they show that it is possible to identify non-xenophobic tweets only through their feelings, emotions, and intentions. A simple interpretation of the CP_6 pattern is that if a tweet has a positive sentiment, its emotion is joy, and it has very little hate-speech intent. Then, it is most likely a non-xenophobic tweet.

Additionally, we can observe that the absence of the words "foreigners", "country", "illegal", together with the little or no presence of negative feelings, emotions, and intentions, greatly help classify non-Xenophobia. To conclude, it is important to note how the word "foreigners" is significant for detecting Xenophobia. By itself, it does not provide enough information to determine whether or not it is Xenophobia. Still, by contextualizing it with feelings and emotions, we can discern whether or not the tweet is xenophobic.

On the one hand, the simple absence of the words "foreigners" and "illegal" are indicators of the non-Xenophobia of the message. While on the other hand, the presence of the words "foreigners", "criminal", "violent", indicates that the message is more likely to be classified as xenophobic. The critical thing about contrast patterns is that they are an excellent opportunity for understanding the process of classifying a message. These patterns extract helpful characteristics of Xenophobia. With this information, experts can have a better understanding of xenophobic behavior in social networks, but also by identifying this type of behavior, possible acts of violence in real life can be prevented [8].

7. Conclusions and Future Work

With the growth of the Internet around the world, people are increasingly connected. The use of social networks has spread more and more. In turn, this has allowed the increase in hate speeches on the Internet. Detecting hate speech in social networks has become an area of great interest for investigation, especially misogyny, racism, sexism. However, detecting Xenophobia in social networks is a problem that the scientific community has not sufficiently studied. Due to the insufficient study of Xenophobia in social networks, few databases currently focus on this topic. Additionally, there has been no proposal to deal with this hate speech using Explainable Artificial Intelligence models. Therefore, in this paper, our proposal uses contrast pattern-based classifiers to detect Xenophobia in social networks, specifically on Twitter.

This paper obtains results by using two databases related to Xenophobia, the Pitropakis Xenophobia Database (PXD) and our proposal, the Experts Xenophobia Database (EXD). EXD has the main characteristic of being labeled by experts in psychology, sociology, and international relations. Additionally, this paper compares three of the most popular state-of-the-art feature representations against our interpretable feature representation proposal based on keywords, feelings, emotions, and intentions. Furthermore, five of the most prominent classifiers were used jointly with each of the feature representations mentioned above for classifying both Xenophobia databases.

From our experimental results, on the one hand, the best Xenophobia classification results in PXD were obtained using Term Frequency–Inverse Document Frequency (TFIDF) as feature representation and PBC4cip as a classifier. On average, TFIDF+PBC4Cip obtained 0.804 in AUC and 0.735 for F1 score with a standard deviation of 0.009 and 0.011, respectively. However, using our INTER+PBC4cip interpretable proposal, the following results were obtained on average: 0.794 in AUC and 0.734 in F1 score with a standard deviation of 0.137 and 0.172, respectively. On the other hand, when EXD was used, the combination of Bag of Words (BOW) jointly with C45 maximized the results of the F1 score, while on the other hand, the combination INTER jointly with PBC4cip maximized the AUC results. On average, BOW+C45 obtained 0.839 in AUC and 0.782 for F1 score with a standard deviation of 0.013 and 0.014, respectively. In contrast, our interpretable proposal obtained 0.864 in AUC and 0.768 in the F1 score on average, with a standard deviation of 0.084 and 0.134.

Our experimental results show that the best combinations of feature representation jointly with an interpretable classifier obtain results on average similar to the non-interpretable varieties. However, it is essential to mention that combinations such as TFIDF+PBC4cip or BOW+C45 receive good results for both AUC and F1 scores and are also quite robust, presenting a small value in their standard deviation. Nevertheless, it is essential to mention that our interpretable feature representation proposal, jointly with a contrast pattern-based classifier, is the only combination that produces interpretable results that experts in the application domain can understand. The use of keywords in conjunction with feelings, emotions, and intentions helps to contextualize the reasons why a post is considered xenophobic or not. As Luo et al. mentioned, feature representations based on numerical transformations are considered black-box; consequently, the results obtained by using black-box approaches are complicated to be understandable by an expert in the application area.

After using the same methodology in both databases, our experimental results show that classifiers trained in EXD obtain better outcomes for both AUC and F1 score metrics than those trained in PXD. We are confident that our expertly labeled Xenophobia database is a valuable contribution to dealing with Xenophobia classification on social media. It is necessary to have more databases focused on Xenophobia to increase the research lines on this problem. Additionally, having more Xenophobia databases can improve the quality of future Xenophobia classification models.

In future work, we want to extend this proposal to other social networks such as Facebook, Instagram, or YouTube, among others. For this, a proposal is to increase our database with entries from other social networks. Each social network has different privacy policies that make extracting posts from its users complicated; consequently, making it different research for each social network. Nevertheless, this proposal aims to create a model that is more adaptable to the classification of Xenophobia in social networks and can take advantage of the differences in the way of writing of each social network.

Author Contributions: Conceptualization, O.L.-G.; methodology, G.I.P.-L. and O.L.-G.; software, G.I.P.-L., O.L.-G., and M.A.M.-P.; validation, O.L.-G. and M.A.M.-P.; formal analysis, G.I.P.-L.; investigation, G.I.P.-L.; resources, O.L.-G. and M.A.M.-P.; data curation, O.L.-G. and G.I.P.-L.; writing—original draft preparation, G.I.P.-L.; writing—review and editing, O.L.-G. and M.A.M.-P.; visualization, G.I.P.-L.; supervision, O.L.-G. and M.A.M.-P.; project administration, G.I.P.-L. and O.L.-G.; funding acquisition, G.I.P.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the National Council of Science and Technology of Mexico under the scholarship grant 1048425.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We want to especially thank Galo Ben Yair Delgado and Laura Moreno as experts in International Relations, María José Sanabria and Néstor José Méndez as experts in Psychology, and Norma Soto as an expert in Sociology, who helped us to label our collected Xenophobia database.

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

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
XAI	Explainable Artificial Intelligence
ML	Machine Learning
AUC	Area Under the Receiver Operating Characteristic Curve
SVM	Support Vector Machine
NB	Naïve–Bayes
GB	Gradient Boosting Machine
LR	Logistic Regression
EV	Ensemble Voting
RNN	Recurrent Neural Networks
LSTM	Long-Short-Term-Memory
CNN	Convolutional Neural Network
sCNN	Skipped Convolutional Neural Network
GRU	Gated Recurrent Unit
RNN	Recurrent Neural Networks
DT	Decision Tree
RF	Random Forest
KNN	k-Nearest Neighbor
RUS	Rusboost
UND	Under Bagging
PBC4cip	Pattern-Based Classifier for Class imbalance problems
WFV	Word Frequency Vectorization
CVF	Count Vector Features
TFIDF	Term Frequency-Inverse Document Frequency
BOW	Bag Of Words
W2V	Word To Vec
INTER	Interpretable Feature Representation
DOB SCV	Distribution Optimally Balanced Stratified Cross-Validation
PXD	Pitropakis Xenophobia Database
EXD	Experts Xenophobia Database
STD	Standard Deviation
AVG	Average

References

1. Yakushko, O. Hatred of strangers: Defining Xenophobia and related concepts. In *Modern-Day Xenophobia: Critical Historical and Theoretical Perspectives on the Roots of Anti-Immigrant Prejudice*; Springer International Publishing: Cham, Switzerland, 2018; pp. 11–31. [CrossRef]
2. Huang, J.; Liu, R. Xenophobia in America in the Age of Coronavirus and Beyond. *J. Vasc. Interv. Radiol. JVIR* **2020**, *31*, 1187–1188. [CrossRef]
3. Loyola-González, O.; Monroy, R.; Rodríguez, J.; López-Cuevas, A.; Mata-Sánchez, J.I. Contrast Pattern-Based Classification for Bot Detection on Twitter. *IEEE Access* **2019**, *7*, 45800–45817. [CrossRef]
4. Chetty, N.; Alathur, S. Hate speech review in the context of online social networks. *Aggress. Violent Behav.* **2018**, *40*, 108–118. [CrossRef]
5. Internet World Stats. World Internet Users Statistics and 2020 World Population Stats. Available online: https://www.internetworldstats.com/stats.htm (accessed on 17 July 2020).
6. Vinerean, S.; Cetina, I.; Dumitrescu, L.; Tichindelean, M. The effects of social media marketing on online consumer behavior. *Int. J. Bus. Manag.* **2013**, *8*, 66. [CrossRef]
7. Clement, J. Global Social Media Ranking 2019. Available online: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on 7 April 2020).
8. Waseem, Z.; Hovy, D. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 13–15 June 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 88–93. [CrossRef]
9. Citizens Crime Commission of New York City. Social Media Use Preceding Real-World Violence. Available online: http://www.nycrimecommission.org/social-media-use-preceding-real-world-violence.php (accessed on 7 April 2020).
10. BBC News. Facebook, Google and Twitter Agree German Hate Speech Deal. Available online: https://www.bbc.com/news/world-europe-35105003 (accessed on 7 April 2020).
11. Pitsilis, G.K.; Ramampiaro, H.; Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.* **2018**, *48*, 4730–4742. [CrossRef]
12. Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web, Montreal, QC, Canada, 11–15 April 2016; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2016; pp. 145–153. [CrossRef]
13. Socialfy. Socialfy—Social Media Marketing Platform. Available online: https://socialfy.pw/facebook-export-comments (accessed on 18 April 2020).
14. Export Comments. Export Facebook, Instagram, Twitter, YouTube, VK, TikTok, Vimeo Comments to CSV/Excel. Available online: https://exportcomments.com/ (accessed on 18 April 2020).
15. Twitter. Twitter API. Available online: https://developer.twitter.com/en/docs/twitter-api (accessed on 12 May 2020).
16. Clement, J. Twitter: Number of Active Users 2010–2019. Available online: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (accessed on 8 April 2020).
17. Buccoliero, L.; Bellio, E.; Crestini, G.; Arkoudas, A. Twitter and politics: Evidence from the US presidential elections 2016. *J. Mark. Commun.* **2020**, *26*, 88–114. [CrossRef]
18. Wicker, H.R. Xenophobia. In *International Encyclopedia of the Social & Behavioral Sciences*; Smelser, N.J., Baltes, P.B., Eds.; Pergamon: Oxford, UK, 2001; pp. 16649–16652. [CrossRef]
19. Farmbry, K. *Migration and Xenophobia: A Three Country Exploration*; Rowman & Littlefield: Lanham, MD, USA, 2019.
20. Crush, J. Global migration. In *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed.; Wright, J.D., Ed.; Elsevier: Oxford, UK, 2015; pp. 169–173. [CrossRef]
21. Inter-Agency. International Migration, Racism, Discrimination and Xenophobia. Available online: https://www.refworld.org/docid/49353b4d2.html (accessed on 12 May 2020).
22. Arrocha, W. Combating Xenophobia and hate through compassionate migration: The present struggle of irregular migrants escaping fear and extreme poverty. *Crime Law Soc. Chang.* **2019**, *71*, 245–260. [CrossRef]
23. Kerr, P.; Durrheim, K.; Dixon, J. Xenophobic Violence and Struggle Discourse in South Africa. *J. Asian Afr. Stud.* **2019**, *54*, 995–1011. [CrossRef]
24. Gagliardone, I.; Gal, D.; Alves, T.; Gabriela, M. *Countering Online Hate Speech*; UNESCO: Paris, France, 2015.
25. Bucio, R. Contra el discurso de odio en redes sociales: Palabras que hieren a México. Available online: https://www.conapred.org.mx/index.php?contenido=registro_encontrado&tipo=2&id=4594 (accessed on 12 May 2020).
26. Alkiviadou, N. Hate speech on social media networks: Towards a regulatory framework? *Inf. Commun. Technol. Law* **2019**, *28*, 19–35. [CrossRef]
27. Timmermann, W.K. The Relationship between Hate Propaganda and Incitement to Genocide: A New Trend in International Law Towards Criminalization of Hate Propaganda? *Leiden J. Int. Law* **2005**, *18*, 257–282. [CrossRef]
28. Marantz, A. Free Speech is Killing us, Noxious Language Online is Causing Real-World Violence. What can we do about it? Available online: https://www.nytimes.com/2019/10/04/opinion/sunday/free-speech-social-media-violence.html (accessed on 12 May 2020).

29. Frenda, S.; Ghanem, B.; Montes, M.; Rosso, P. Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4743–4752. [CrossRef]
30. 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.
31. Lingiardi, V.; Carone, N.; Semeraro, G.; Musto, C.; D'Amico, M.; Brena, S. Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behav. Inf. Technol.* **2019**, 1–11. [CrossRef]
32. Plaza-Del-Arco, F.M.; Molina-González, M.D.; Ureña López, L.A.; Martín-Valdivia, M.T. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *Acm Trans. Internet Technol.* **2020**, *20*. [CrossRef]
33. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access* **2019**, *7*, 154096–154113. [CrossRef]
34. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
35. Loyola-González, O.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; García-Borroto, M. Cost-Sensitive Pattern-Based classification for Class Imbalance problems. *IEEE Access* **2019**, *7*, 60411–60427. [CrossRef]
36. Loyola-González, O.; Medina-Pérez, M.A.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; Monroy, R.; García-Borroto, M. PBC4cip: A new contrast pattern-based classifier for class imbalance problems. *Knowl. Based Syst.* **2017**, *115*, 100–109. [CrossRef]
37. Zhang, X.; Dong, G. Overview and Analysis of Contrast Pattern Based Classifica-tion. In *Contrast Data Mining*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2016; pp. 175–194.
38. Dong, G.; Bailey, J. *Contrast Data Mining: Concepts, Algorithms, and Applications*, 1st ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2012.
39. Zhang, T.; Wang, J.; Xu, L.; Liu, P. Fall Detection by Wearable Sensor and One-Class SVM Algorithm. In *Intelligent Computing in Signal Processing and Pattern Recognition, Proceedings of the International Conference on Intelligent Computing, ICIC 2006, Kunming, China, 16–19 August 2006*; Huang, D.S., Li, K., Irwin, G.W., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 858–863. [CrossRef]
40. Kang, S.; Ramamohanarao, K. A Robust Classifier for Imbalanced Datasets. In *Advances in Knowledge Discovery and Data Mining*; Tseng, V.S., Ho, T.B., Zhou, Z.H., Chen, A.L.P., Kao, H.Y., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 212–223.
41. Zhang, X.; Dong, G.; Ramamohanarao, K. Information-Based Classification by Aggregating Emerging Patterns. In *Intelligent Data Engineering and Automated Learning—IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents*; Leung, K.S., Chan, L.W., Meng, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2000; pp. 48–53.
42. Aguilar, D.L.; Loyola-González, O.; Medina-Pérez, M.A.; Cañete-Sifuentes, L.; Choo, K.K.R. PBC4occ: A novel contrast pattern-based classifier for one-class classification. *Future Gener. Comput. Syst.* **2021**, *125*, 71–90. [CrossRef]
43. Chen, X.; Gao, Y.; Ren, S. *A New Contrast Pattern-Based Classification for Imbalanced Data*; ISCSIC '18; Association for Computing Machinery: New York, NY, USA, 2018. [CrossRef]
44. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
45. Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **1994**, *16*, 235–240. [CrossRef]
46. Rokach, L.; Maimon, O. *Data Mining with Decision Trees*, 2nd ed.; World Scientific: Singapore, 2014. [CrossRef]
47. García-Borroto, M.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; Medina-Pérez, M.A.; Ruiz-Shulcloper, J. LCMine: An efficient algorithm for mining discriminative regularities and its application in supervised classification. *Pattern Recognit.* **2010**, *43*, 3025–3034. [CrossRef]
48. García, S.; Grill, M.; Stiborek, J.; Zunino, A. An empirical comparison of botnet detection methods. *Comput. Secur.* **2014**, *45*, 100–123. [CrossRef]
49. El Mazouri, F.Z.; Abounaima, M.C.; Zenkouar, K. Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France. *J. Big Data* **2019**, *6*, 5. [CrossRef]
50. Usha, D.; Rameshkumar, K.; Baiju, B.V. Association Rule Construction from Crime Pattern Through Novelty Approach. In *Advances in Big Data and Cloud Computing*; Peter, J.D., Alavi, A.H., Javadi, B., Eds.; Springer: Singapore, 2019; pp. 241–250.
51. Wulandari, C.P.; Ou-Yang, C.; Wang, H.C. Applying mutual information for discretization to support the discovery of rare-unusual association rule in cerebrovascular examination dataset. *Expert Syst. Appl.* **2019**, *118*, 52–64. [CrossRef]
52. Loyola-González, O.; López-Cuevas, A.; Medina-Pérez, M.A.; Camiña, B.; Ramírez-Márquez, J.E.; Monroy, R. Fusing pattern discovery and visual analytics approaches in tweet propagation. *Inf. Fusion* **2019**, *46*, 91–101. [CrossRef]
53. Weng, C.H.; Huang, T.C.K. Observation of sales trends by mining emerging patterns in dynamic markets. *Appl. Intell.* **2018**, *48*, 4515–4529. [CrossRef]
54. Loyola-González, O.; Monroy, R.; Medina-Pérez, M.A.; Cervantes, B.; Grimaldo-Tijerina, J.E. An Approach Based on Contrast Patterns for Bot Detection on Web Log Files. In *Advances in Soft Computing*; Batyrshin, I., Martínez-Villaseñor, M.d.L., Ponce Espinosa, H.E., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 276–285.
55. Hu, W.; Chen, T.; Shah, S.L. Detection of Frequent Alarm Patterns in Industrial Alarm Floods Using Itemset Mining Methods. *IEEE Trans. Ind. Electron.* **2018**, *65*, 7290–7300. [CrossRef]

56. Tabatabaee Malazi, H.; Davari, M. Combining emerging patterns with random forest for complex activity recognition in smart homes. *Appl. Intell.* **2018**, *48*, 315–330. [CrossRef]
57. Cavadenti, O.; Codocedo, V.; Boulicaut, J.F.; Kaytoue, M. What Did I Do Wrong in My MOBA Game? Mining Patterns Discriminating Deviant Behaviours. In Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2016; pp. 662–671. [CrossRef]
58. Chavary, E.A.; Erfani, S.M.; Leckie, C. *Summarizing Significant Changes in Network Traffic Using Contrast Pattern Mining*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 2015–2018.
59. 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*, 11. [CrossRef]
60. Sparck Jones, K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]
61. Suthaharan, S. Support Vector Machine. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*; Springer: Boston, MA, USA, 2016; pp. 207–235. [CrossRef]
62. Kononenko, I. Semi-naive bayesian classifier. In *Machine Learning—EWSL-91*; Kodratoff, Y., Ed.; Springer: Berlin/Heidelberg, Germany, 1991; pp. 206–219.
63. Greiner, R.; Su, X.; Shen, B.; Zhou, W. Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers. *Mach. Learn.* **2005**, *59*, 297–322. [CrossRef]
64. Mikolov, T.; Kombrink, S.; Burget, L.; Černocký, J.; Khudanpur, S. Extensions of recurrent neural network language model. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Czechia, Prague, 22–27 May 2011; pp. 5528–5531. [CrossRef]
65. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
66. Charitidis, P.; Doropoulos, S.; Vologiannidis, S.; Papastergiou, I.; Karakeva, S. Towards countering hate speech against journalists on social media. *Online Soc. Netw. Media* **2020**, *17*, 100071. [CrossRef]
67. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6. [CrossRef]
68. Kim, Y.H.; An, G.J.; Sunwoo, M.H. CASA: A Convolution Accelerator using Skip Algorithm for Deep Neural Network. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Hokkaido, Japan, 26–29 May 2019; pp. 1–5. [CrossRef]
69. Dey, R.; Salem, F.M. Gate-variants of Gated Recurrent Unit (GRU) neural networks. In Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 6–9 August 2017; pp. 1597–1600. [CrossRef]
70. Fernández-Díaz, M.; Gallardo-Antolín, A. An attention Long Short-Term Memory based system for automatic classification of speech intelligibility. *Eng. Appl. Artif. Intell.* **2020**, *96*, 103976. [CrossRef]
71. Sahay, K.; Khaira, H.S.; Kukreja, P.; Shukla, N. Detecting cyberbullying and aggression in social commentary using nlp and machine learning. *Int. J. Eng. Technol. Sci. Res.* **2018**, *5*.
72. Ayyadevara, V. *Gradient Boosting Machine*; Apress: Berkeley, CA, USA, 2018; pp. 117–134. [CrossRef]
73. Langford, J.; Li, L.; Strehl, A. Vowpal Wabbit. 2011. Available online: https://github.com/JohnLangford/vowpalwabbit/wiki (accessed on 10 October 2020).
74. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
75. Roesslein, J. Tweepy: Twitter for Python! Available online: https://github.com/tweepy/tweepy (accessed on 12 May 2020).
76. Shin, D. Analysis of online social networks: A cross-national study. *Online Inf. Rev.* **2010**, *34*, 473–495. [CrossRef]
77. Al-Smadi, M.; Al-Ayyoub, M.; Jararweh, Y.; Qawasmeh, O. Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features. *Inf. Process. Manag.* **2019**, *56*, 308–319.
78. Yuxin, D.; Xuebing, Y.; Di, Z.; Li, D.; Zhanchao, A. Feature representation and selection in malicious code detection methods based on static system calls. *Comput. Secur.* **2011**, *30*, 514–524. [CrossRef]
79. Luo, H.; Liu, Z.; Luan, H.; Sun, M. Online Learning of Interpretable Word Embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 1687–1692. [CrossRef]
80. Parallel Dots. Text Analysis APIs. Available online: https://komprehend.io/ (accessed on 15 April 2021).
81. Meaning Cloud. Sentiment Analysis API. Available online: https://www.meaningcloud.com/developer/sentiment-analysis (accessed on 15 April 2021).
82. IBM. Watson Natural Language Understanding. Available online: https://www.ibm.com/cloud/watson-natural-language-understanding (accessed on 15 April 2021).
83. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo*. **2020**. [CrossRef]
84. Vo, A.D.; Nguyen, Q.P.; Ock, C.Y. Semantic and syntactic analysis in learning representation based on a sentiment analysis model. *Appl. Intell.* **2020**, *50*, 663–680. [CrossRef]

85. Liu, H.; Cocea, M. Fuzzy rule based systems for interpretable sentiment analysis. In Proceedings of the 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), Doha, Qatar, 4–6 February 2017; pp. 129–136. [CrossRef]
86. Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; et al.. OntoNotes Release 5.0. LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium, 2013. Available online: https://catalog.ldc.upenn.edu/LDC2013T19 (accessed on 15 April 2021).
87. Larkey, L.S.; Ballesteros, L.; Connell, M.E. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-Occurrence Analysis. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 11–15 August 2002; Association for Computing Machinery: New York, NY, USA, 2002; pp. 275–282. [CrossRef]
88. Al-Shammari, E.; Lin, J. A Novel Arabic Lemmatization Algorithm. In Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, Singapore, 24 July 2008; Association for Computing Machinery: New York, NY, USA, 2008; pp. 113–118. [CrossRef]
89. García-Borroto, M.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A. A New Emerging Pattern Mining Algorithm and Its Application in Supervised Classification. In *Advances in Knowledge Discovery and Data Mining*; Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 150–157.
90. Rodríguez, J.; Medina-Pérez, M.A.; Gutierrez-Rodríguez, A.E.; Monroy, R.; Terashima-Marín, H. Cluster validation using an ensemble of supervised classifiers. *Knowl. Based Syst.* **2018**, *145*, 134–144. [CrossRef]
91. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
92. García-Borroto, M.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A. Finding the best diversity generation procedures for mining contrast patterns. *Expert Syst. Appl.* **2015**, *42*, 4859–4866. [CrossRef]
93. Harris, Z. Distributional structure. *Word* **1954**, *10*, 146–162. [CrossRef]
94. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; Curran Associates Inc.: Red Hook, NY, USA, 2013; Volume 2, pp. 3111–3119.
95. Zeng, X.; Martinez, T.R. Distribution-balanced stratified cross-validation for accuracy estimation. *J. Exp. Theor. Artif. Intell.* **2000**, *12*, 1–12. [CrossRef]
96. Ting, K. An instance-weighting method to induce cost-sensitive trees. *IEEE Trans. Knowl. Data Eng.* **2002**, *14*, 659–665. [CrossRef]
97. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
98. Seiffert, C.; Khoshgoftaar, T.; Hulse, J.V.; Napolitano, A. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. Part* **2010**, *40*, 185–197. [CrossRef]
99. Barandela, R.; Valdovinos, R.; Sánchez, J. New applications of ensembles of classifiers. *Pattern Anal. Appl.* **2003**, *6*, 245–256. [CrossRef]
100. Alcalá-Fdez, J.; Sánchez, L.; García, S.; del Jesus, M.J.; Ventura, S.; Garrell, J.M.; Otero, J.; Romero, C.; Bacardit, J.; Rivas, V.M.; et al. KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Comput.* **2009**, *13*, 307–318. [CrossRef]
101. Frank, E.; Hall, M.A.; Holmes, G.; Kirkby, R.; Pfahringer, B.; Witten, I.H. Weka: A machine learning workbench for data mining. In *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*; Springer: Berlin, Germany, 2005; pp. 1305–1314.
102. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *AI 2006: Advances in Artificial Intelligence*; Sattar, A., Kang, B.H., Eds.; Springe: Berlin/Heidelberg, Germany, 2006; pp. 1015–1021.
103. Halimu, C.; Kasem, A.; Newaz, S.H.S. Empirical Comparison of Area under ROC Curve (AUC) and Mathew Correlation Coefficient (MCC) for Evaluating Machine Learning Algorithms on Imbalanced Datasets for Binary Classification. In Proceedings of the 3rd International Conference on Machine Learning and Soft Computing, Da Lat, Vietnam, 25–28 January 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–6.
104. Jeni, L.A.; Cohn, J.F.; De La Torre, F. Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 245–251. [CrossRef]
105. Hasan, T.; Matin, A. Extract Sentiment from Customer Reviews: A Better Approach of TF-IDF and BOW-Based Text Classification Using N-Gram Technique. In Proceedings of the International Joint Conference on Advances in Computational Intelligence, Virtual, 25 August 2021; Uddin, M.S., Bansal, J.C., Eds.; Springer: Singapore, 2021; pp. 231–244.
106. Arras, L.; Horn, F.; Montavon, G.; Müller, K.R.; Samek, W. What is relevant in a text document?: An interpretable machine learning approach. *PLoS ONE* **2017**, *12*, e0181142.
107. García, S.; Fernández, A.; Herrera, F. Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. *Appl. Soft Comput.* **2009**, *9*, 1304–1314. [CrossRef]
108. Lem, S.; Onghena, P.; Verschaffel, L.; Van Dooren, W. The heuristic interpretation of box plots. *Learn. Instr.* **2013**, *26*, 22–35. [CrossRef]