applied sciences

MDPI

Article

Analysis of Harassment Complaints to Detect Witness Intervention by Machine Learning and Soft Computing Techniques

Marina Alonso-Parra [1], Cristina Puente [1,*], Ana Laguna [1] and Rafael Palacios [1,2]

[1] Computer Science Department, ICAI School of Engineering, Comillas Pontifical University, 28015 Madrid, Spain; marinaalon-so98@gmail.com (M.A.-P.); alaguna@comillas.edu (A.L.); rafael.palacios@iit.comillas.edu (R.P.)

[2] Institute for Research in Technology (IIT), ICAI School of Engineering, Comillas Pontifical University, 28015 Madrid, Spain

* Correspondence: cristina.puente@comillas.edu

Abstract: This research is aimed to analyze textual descriptions of harassment situations collected anonymously by the Hollaback! project. Hollaback! is an international movement created to end harassment in all of its forms. Its goal is to collect stories of harassment through the web and a free app all around the world to elevate victims' individual voices to find a societal solution. Hollaback! pretends to analyze the impact of a bystander during a harassment in order to launch a public awareness-raising campaign to equip everyday people with tools to undo harassment. Thus, the analysis presented in this paper is a first step in Hollaback!'s purpose: the automatic detection of a witness intervention inferred from the victim's own report. In a first step, natural language processing techniques were used to analyze the victim's free-text descriptions. For this part, we used the whole dataset with all its countries and locations. In addition, classification models, based on machine learning and soft computing techniques, were developed in the second part of this study to classify the descriptions into those that have bystander presence and those that do not. For this machine learning part, we selected the city of Madrid as an example, in order to establish a criterion of the witness behavior procedure.

Keywords: social violence; natural language processing; text classification; machine learning; harassment complaints; bystander presence

Citation: Alonso-Parra, M.; Puente, C.; Laguna, A.; Palacios, R. Analysis of Harassment Complaints to Detect Witness Intervention by Machine Learning and Soft Computing Techniques. *Appl. Sci.* **2021**, *11*, 8007. https://doi.org/10.3390/app11178007

Received: 29 June 2021
Accepted: 20 August 2021
Published: 29 August 2021

1. Introduction

Today, most women and many men experience some form of sexual harassment in their lives. For many years, these situations were not denounced, but in 2018, there was a big media movement in which sexual harassment scandals by film directors, big executives, photographers and some others came to light. This triggered an important movement on social networks called #MeToo, where a multitude of people who had suffered harassment in their lives raised their voices.

The magnitude of this movement prompted a US study of the same name in 2019, which revealed that 76% of all women claimed to have experienced verbal sexual harassment, a figure that drops to 43% for all men [1]. The study also found that women are more likely than men to have experienced sexual assault, as shown in Figure 1.

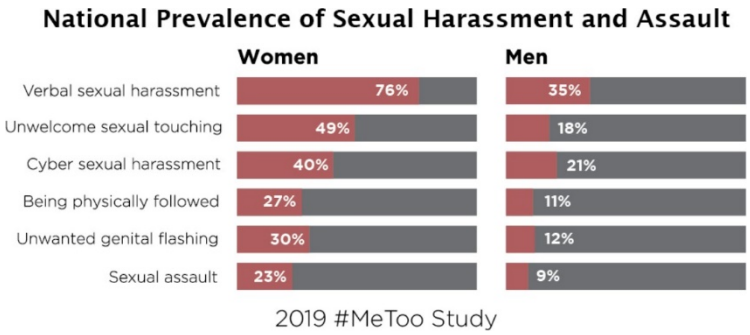

Figure 1. National prevalence of sexual harassment and assault [2].

These experiences usually have a negative impact on the victims, generating feelings of fear, shame and anger. They can even cause trauma and insecurities for the rest of the victims' lives. Therefore, this type of behavior can be considered as gender violence and therefore a violation of human rights. At the same time, they are in direct conflict with the Sustainable Development Goals where gender equality (goal 5) and just and safe societies (goal 16) are advocated [3]. Due to the great importance of these situations, gender-based violence and violence against women and girls are a major problem in many countries.

Due to the impact that these situations have on society, in recent years, many associations have gained prominence that seek to combat them. Among them is Hollaback!, which is an NGO founded in 2005 in New York with the aim of providing support, raising awareness and eliminating harassment in public places. It consists of a platform where victims or witnesses of harassment situations can register and publish their own experiences. The testimonies are accompanied by the geolocation of the incident, the date and time, the type of harassment and a description of what happened, and they can even upload a photo or description of the aggressor. Then, these experiences are published on the website so that other victims can read them and feel accompanied, reinforced by people who stood up to them or were attentive to an incident in the same area; in short, they feel part of a community [4].

Over the years, numerous studies have been conducted on the psychological effects that sexual harassment experiences have on the victims, from the point of view of psychology and sociology [5]. Currently, thanks to the advance of new technologies such as artificial intelligence, it has been possible to analyze and work on the issue of sexual harassment from a different perspective that has allowed a breakthrough in the studies. Machine learning techniques have been used to solve various problems, such as the use of text mining as a tool to locate and summarize more than 5000 articles on the topic to help researchers to find information [6]. Another tool related to the field of sexual harassment was the creation of a chatbot to psychologically assist victims, as presented in [7]. However, one of the most interesting uses of artificial intelligence in this field is for detecting situations of harassment and violence through Natural Language Processing (NLP) and deep learning, as in [8], or the use of sentiment analysis in social networks such as Twitter to understand the feeling of victims or harassers, as in [9]. In this sense, many studies have sought ways to refine models to increase detection accuracy such as [10], where, through ensemble methods, up to 97% accuracy was achieved by identifying sentiments.

NLP techniques are very powerful in extracting information from free-text descriptions that do not directly state such information [11,12]. One example is the analysis of social network messages to detect special behavior such as Jihadist ideology [13]. The analysis of a huge number of events is a problem that can be decomposed into smaller sets of problems because each harassment description is independent from the others. Therefore, to reduce the total computational time, it is beneficial to apply distributed computing or grid computing techniques [14].

Over recent years, the psychological effect of these incidents on victims has been analyzed, and the many factors that can modify it have been identified. In the case of street

harassment, it has been proven that the intervention of a witness is a very important factor in decreasing or increasing the psychological effect on the victim. This has led some NGOs to launch various initiatives to raise awareness of the behaviors that witnesses should have in the event of witnessing a scene of harassment. A good example would be the witness action guide, developed by CUP and Hollaback! [15].

The influence of bystanders in sexual harassment situations has always been analyzed from a psychological and sociological point of view, highlighting the importance it has in affecting the victims' feelings. Articles such as [16,17] state the importance of bystanders in both physical and online situations, explain the reasons for the lack of action of bystanders and propose measures to motivate action.

There is a lot of active research in the field of bystander intervention, and thanks to the new technologies, a different perspective can be obtained. This paper aims to help to understand the impact of witnesses through the analysis of the textual descriptions in the Hollaback! database; in order to achieve this, this study is divided as follows: In the next section, we perform an exploratory analysis of the information contained in the database to obtain statistics on the temporal and spatial distribution of the descriptions, the languages of the descriptions and the type of harassment. Once we explore the information, in Section 3, we take care of the processing of the descriptions with NLP tools (cleaning and standardization of the text, reduction in dimensionality, word embedding). Section 4 explains the development of an ML classification model with supervised learning tools. Finally, conclusions and suggestions for future work finish the paper in Section 5.

2. Exploratory Data Analysis

In this section, we present an overview of the data to obtain relevant statistics about the database, such as the locations of the complaints, the type of descriptions, the time slots in which they occurred, the type of complaints and the time periods in which the complaints were received. This information is very valuable in the process of establishing first-sight conclusions.

2.1. Database Description

The database used for this research contains approximately 12,000 different complaints collected since 2015 by the NGO Hollaback! It started in 2005, working to end gender-based harassment in public spaces, also known as street harassment. In 2015, it expanded its mission to work on harassment across all spaces including online, the workplace, transportation, protests and polling booths and all identities including women, LGBTQ+ folks, Black folks, Indigenous folks, people of color, religious minorities, people with disabilities, immigrants and all others who are treated as "less than" just for being who they are. They seek to uproot hate and harassment whether is perpetuated by individuals or institutions, and in the messy areas in between in issues such as voter suppression, police brutality and ICE raids.

The reports are made by victims or witnesses of harassment situations in public places and are composed of different fields:

- ID: This is a character string, unique in the database, assigned automatically by the page when the report is made, which allows it to be identified.
- ReportedAt: This is a value in date/time format (YYYYY-MM-DD T hh:mm:ss) and contains the date and time of the incident.
- Lng: This is a value in floating point (float) format, and it collects the longitude of the place where the incident occurred.
- Lat: This is also a value in floating point format (float), and it collects the latitude of the place where the incident occurred.

- Categories: This is a character string that collects the category or categories of the harassment suffered. The different categories are: verbal abuse, sexual gestures, inappropriate touching, being followed, homophobia, transphobia, racism, colorism, discrimination against people with disabilities or sizeism.
- ReportedByBystander: This is a Boolean value that is true if the story is reported by a witness and false if it is reported by the victim.
- Title: This is a string with the title of the description.
- Description: This is a character string with the description of the incident. This field will be used to make the classification.

With this information, our goal is to classify those events in which a witness or bystander is present at the scene. In regard to the current work, a person at the scene, different from the victim and the attacker, will be considered a witness. However, due to the great variety of descriptions and the different ways of recounting experiences, the difference between presence and non-presence can sometimes be complex. In the following, the considerations made for the classification will be detailed. Witness presence is considered whenever there is someone who may have seen the incident, whether they intervene or not. This is because if there is a person present during the event, the victim's feelings are already changed, regardless on whether the witness intervenes or not. If the victims or harassers are more than one person, they are not considered as witnesses, but as multiple victims or multiple harassers, which may complicate the analysis even further.

When speaking of a public place, it will be considered that a witness was present when it is explicitly stated that there were people or reference is made to some external person; otherwise, despite being public, it will be considered that there were no witnesses.

2.2. Information Analisis

To interpret the results correctly, it is important to bear in mind that actual harassment events and the number of reports are not strongly correlated. There are numerous social and cultural factors, such as awareness of the Hollaback! platform, which may favor more or less reporting depending on the conditions. Some of the factors thought to be affected are geographic location, time of occurrence, language of reporting or type of harassment.

First, we located the origin of the complaints. Thanks to the free tool called Carto [18], different data layers can be represented on a dynamic map. In this case, a representation of the reports was made on a world map using the latitude and longitude provided. In this way, it was possible to observe the general distribution of the reports, which reflects, to a large extent, the global scope of the Hollaback! movement.

In this picture, it can be seen that the complaints are mainly concentrated in Europe and the United States. In these places, the Hollaback! movement is strongest, and this is due to several factors. One of the most important is because they are the first places where the platform was launched and therefore the origin of the movement, namely, in New York. This image (Figure 2), also reflects the areas of the world where the movement has not yet reached, such as Africa, or is just beginning to gain momentum, such as India, Southeast Asia and South America.

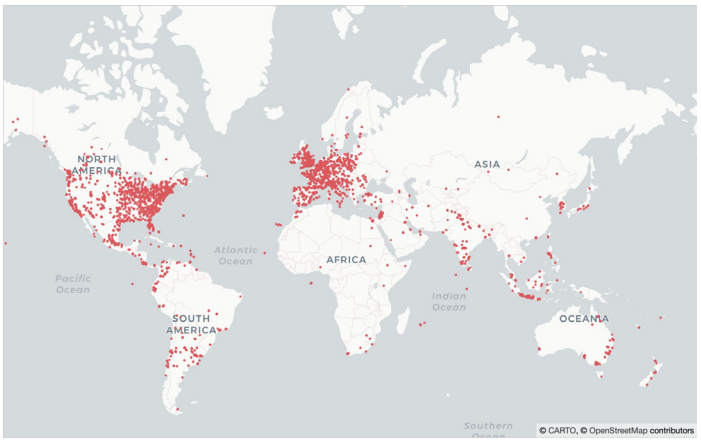

Figure 2. Worldwide distribution of complaints.

Another interesting datum is the distribution of the number of complaints over the years, showing how the success and scope of Hollaback! has varied since its creation. It can be observed that since it started in 2005, the number of users grew until 2014 where it had its highest point; since then, the number of complaints has been decreasing over the years. One might think that the drop is due to a decrease in the number of harassment situations; however, it is unlikely that the decrease is so large, especially when compared to general statistics. Therefore, it is more likely that the decrease is caused by the emergence of new similar platforms [19], such as StoPit Solutions [20], the #notme app [21], Spot [22] and some other similar apps that have caused this decrement in the distribution of users, as Figure 3 shows.

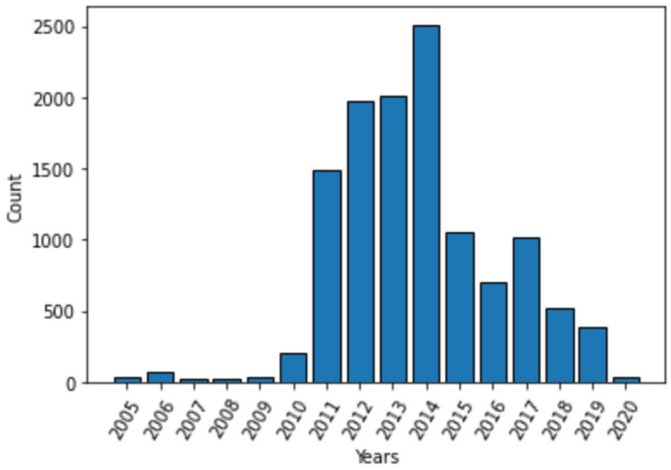

Figure 3. Number of users of the iHollaback! app since its creation.

Subsequently, the distribution of complaints was studied according to four time intervals: early morning (1 a.m. to 6 a.m.), morning (7 a.m. to 12 a.m.), afternoon (1 p.m. to 6 p.m.) and evening (7 p.m. to 12 p.m.). The objective was to see in which intervals it is more frequent to suffer a harassment experience and also to see how these experiences are distributed in a city.

In this case, the city of Madrid was chosen in order to better understand the results. We chose Madrid because there were many complaints collected in this city during a limited amount of time. Additionally, the city is very well divided into mostly residential neighborhoods, and other regions with nightlife activity where more harassment events are expected. Despite there being other cities, we chose Madrid as an example used to show the information that could be obtained from the database and that would allow for a better understanding of the dynamics of the city.

In Figure 4, it is clear that the lowest number of incidents occurred in the early morning, but that they are more concentrated in the center.

In the rest of the time intervals, the concentration of complaints continues to be in the most central sector; however, in the mornings, it tends to be displaced to the north, and in the afternoons, to the southeast. The majority concentration in the center is due to Hollaback! being better known among foreigners who move mainly in that area. This can be verified by looking at the language of the descriptions, with English almost always in the center and Spanish on the outskirts. Therefore, it cannot be deduced that there is more harassment in one area or another, only that it is more reported.

Figure 4. Distribution of the complaints by time intervals in Madrid.

Figure 5 also shows a large difference between the early morning hours and the rest of the day. To check whether this occurred in the rest of the world, a histogram was made by time intervals of all the reports.

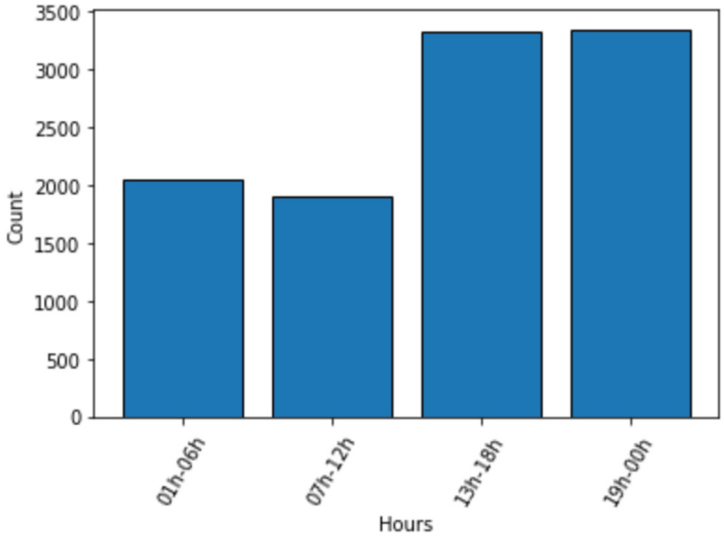

Figure 5. Histogram of complaints by time intervals.

As it seems logical, in Figure 5, we observe that the early morning and morning are the time intervals with the fewest complaints. Seems interestingly, despite the fact that in the early morning, there is a third less traffic on the streets, there are more complaints than in the morning. Another interesting fact in the graph is that in the afternoon and at night, there is the same number of complaints. Therefore, as in the previous case, since there are fewer people at night, there is a greater probability of being harassed.

Another interesting datum collected in the database is the type of harassment suffered by the victims. These data can be compared with those presented in the introduction and elaborated by #MeToo [1], which took into account all types of harassment and not only those in public places. In the study, as seen in Figure 6, the most frequent group was victims of verbal harassment; according to the study, these victims represented 76% of the cases in women. In the case of the database, the figure is around 60%, but all those descriptions listed as "other" should be considered. In the study, the second category was unwanted touching, and this also occurs with the descriptions in the database, with "groping" being the second most frequent. Therefore, it can be seen that the database, despite not covering as many areas as the #MeToo study, has, in general, the same qualitative results.

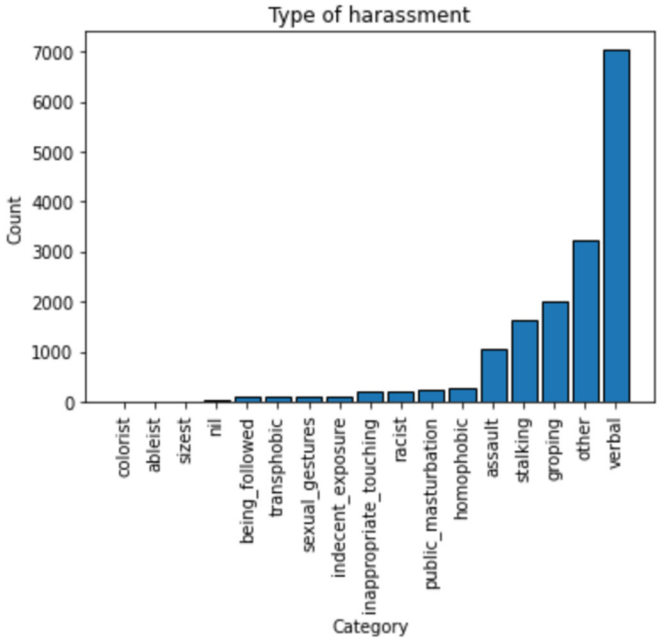

Figure 6. Frequency of type of harassment.

Finally, in the database, there is a category to indicate whether the description has been uploaded by a witness who witnessed the incident. This is a good indicator of whether the witness took an active role during the harassment, since if they report the incident, it is likely that witnesses were involved. We see that only 5% of the descriptions were uploaded by witnesses, which could imply that most witnesses do not take direct action. This fact does not necessarily imply that there was no witness who took action. However, if descriptions are uploaded, it indicates that witnesses probably did take action, and the fact that only 5% of descriptions were uploaded by witnesses, as seen in Figure 7, shows that the number of witnesses who give importance to these situations is still very small (important enough to put them in a blog).

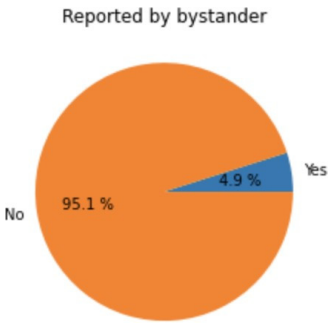

Figure 7. Volume of harassments reported by bystanders.

However, it is also important to see the relationship between the number of reports made in a year and the number of reports made by witnesses in the same year, in order to know if they have increased in proportion. Figure 8 shows that over the years, if we discard 2020 because it is incomplete, the number of reports made by witnesses has increased slightly in relation to the number of reports made. This may suggest that people are more aware of these situations.

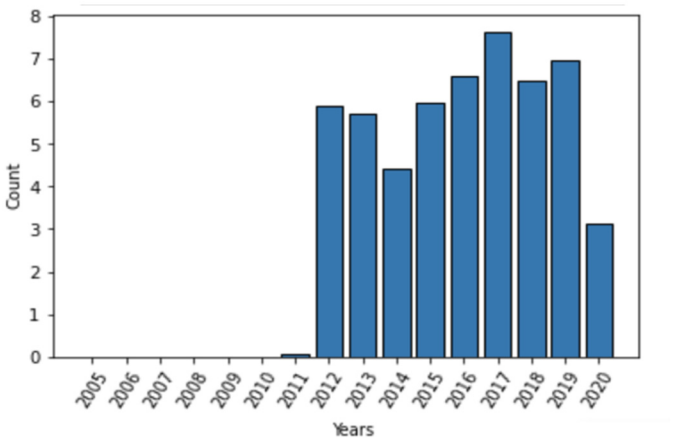

Figure 8. Complaints reported by bystanders over the years.

In this paper, the descriptions made by victims were processed to detect the presence of witnesses. This allows for a better labeling of the data for future work to make witnesses aware of the importance of their intervention.

3. Text Preprocessing and Exploratory NLP Techniques

To create the classification models with the database descriptions, it is important that the text be as homogeneous as possible. Text processing was performed in a series of steps and algorithms described in Figure 9 that will be explained one by one below, which aim to reduce ambiguity without losing important information.

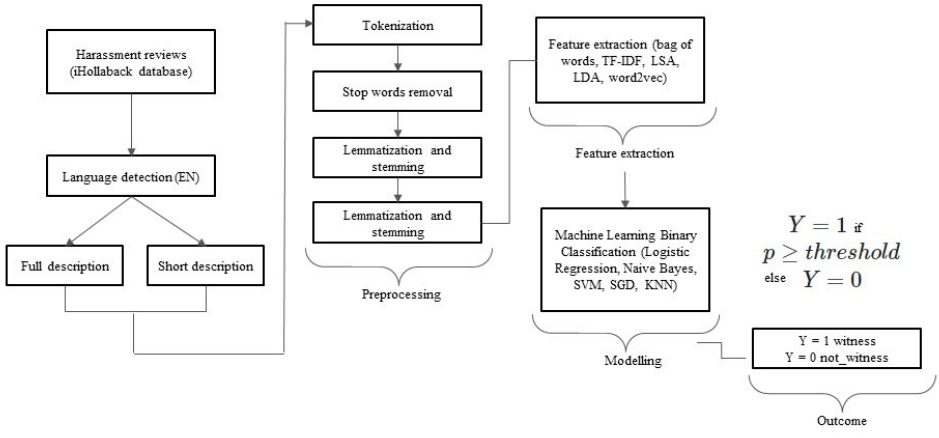

Figure 9. Workflow diagram of the harassment classification pipeline.

At the beginning of the algorithm, all the data were imported to be used in the preliminary analysis through the CSV module [23] and by means of a dictionary that uses the description number as a key and contains attributes, as seen in Figure 10.

```
hour_and_date:2010-10-04T15:43:58+00:00
longitude:-122.67611111111
latitude:45.523333333333
categories:["verbal"]
reported_by_bystander:false
description:Portland, OR is a place where I actually encountered little sexual harassment on the street! Is it ironic that I

"Not to be disrespectful, but that is a very pretty backside." UGH. As I biked by him on the way up the block, I yelled back
```

Figure 10. Data structure of one dictionary element coming from the CSV file.

Next, we needed to clean the data. Two functions with two different modules were used to perform this task. The first one was bs4 from BeautifulSoup [24] which removed html tags that may be in the descriptions.

The second module was Contractions [25] to separate possible contractions from the text and treat them as two words. This process had the expected effect and managed to remove even the most complicated contractions such as "gonna" and "wanna"; however, it did not deal with the genitive ('s).

To tokenize the text, the wordtokenize module of the nltk [26] library was used. Having previously eliminated the contractions avoids the situation where the tokenization separates text according to the apostrophes, instead taking into account the complete words, except for the 's used in English, passing them to the normalization process, which aims to make all the tokens equal and easy to process. The first step was to convert the words to lowercase so that, for example, "HOLA" and "hola" are not considered two different words. The second step consisted of eliminating the genitive ('s) since it was found that it was separated during tokenization and generated errors during processing if it was not eliminated. The third one eliminated punctuation and other foreign characters that may have remained after tokenization; however, it kept accents and letters from other languages in case the processing was conducted in another language. Figure 11 shows an example of the text after processing.

```
Frase de ejemplo:
I wasn't <p> gonna </p> <main> go</main>. However, my father's friend Marco convinced me to go & I went.


Frase tras limpieza:
I was not  going to   go. However, my father's friend Marco convinced me to go & I went.


Frase tras tokinización:
['I', 'was', 'not', 'going', 'to', 'go', '.', 'However', ',', 'my', 'father', "'s", 'friend', 'Marco', 'convinced', 'me', 'to', 'go', '&', 'I', 'went', '.']


Frase tras normalización:
['i', 'was', 'not', 'going', 'to', 'go', 'however', 'my', 'father', 'friend', 'marco', 'convinced', 'me', 'to', 'go', 'i', 'went']
```

Figure 11. Text preprocessing results.

Another issue was language detection. During the reading process of the descriptions, it was found that they existed in several different languages; therefore, it was imperative to know the language of each one in order to be able to treat them. The most common languages and therefore the ones detected were: English, Spanish, French, Italian and German; the rest of the descriptions were saved as "other languages".

To perform this detection, we made use of the stop words of each dictionary which, as previously mentioned, are the most frequent words in a language. For each language, the number of stop words in the text was counted, and the language in which this number was the highest was chosen.

To deal with "other languages", a minimum count of stop words was established, if they did not reach the "other" category. This minimum started as 5, but later, it was realized that it did not work for long texts since words such as "a" occur in all languages. Therefore, we ended up filtering according to size and establishing that at least 10% of the words had to be stop words to belong to a language.

This process still presented some problems with similar languages such as Spanish and Portuguese. If we wanted more precision, we could solve it by comparing the whole

dictionary instead of the stop words; however, this would be a very slow process and would consume a lot of resources.

For this paper, only the English descriptions were used, which are the most numerous in the database; the rest of the descriptions were stored for future analysis. As mentioned before, sometimes, there are some languages that are mislabeled due to stop words; this happens in the case of English. Since these words would greatly bias the classification algorithm if they were introduced, it is essential to eliminate them. For this purpose, a function was introduced that checks, word by word, all the descriptions labeled as English and eliminates those words that do not belong to the English dictionary. This was conducted by obtaining the complete dictionary thanks to the words module of the nltk.corpus library [26]. Lemmatization and stemming, as Figure 12 shows, were used to reduce the number of words in the memory, leave them with the same root and be able to compare them as equal words. The Lan-casterStemmer and the WordNetLemmatizer from the nltk.stem library were used for this process. During the process, it was found that the stemmer worked well; however, the lemmatizer needed information about the type of word to be lemmatized in order to be more accurate. This led to the implementation of a function to detect the type of word thanks to the postag module of nltk [26].

```
Ejemplo: I was walking with my best friend to go pick up my little brother from elementary school. I saw a car pull up r
Ejemplo stemming: ['walk', 'best', 'friend', 'go', 'pick', 'littl', 'broth', 'el', 'school', 'saw', 'car', 'pul', 'real'
Ejemplo lematización: ['walk', 'best', 'friend', 'go', 'pick', 'little', 'brother', 'elementary', 'school', 'saw', 'car'
```

Figure 12. Stemming and lemmatization results.

Next, we applied some document representation models. One of the models of document representation was the bag-of-words technique. In this case, it was applied to gain an idea of which were the most frequent words in the texts. For this purpose, two tools were used: wordcloud [27], which allows us to gain a more pleasant view of the most important words, and FreqDist from nltk [26], which allows us to know the count exactly.

In Figure 13, we see that verbs such as "say" and "walk" are the most frequent words, with around 10,000 occurrences in the 8000 English documents, suggesting that most people had something happen to them while they were walking or that someone spoke to them.

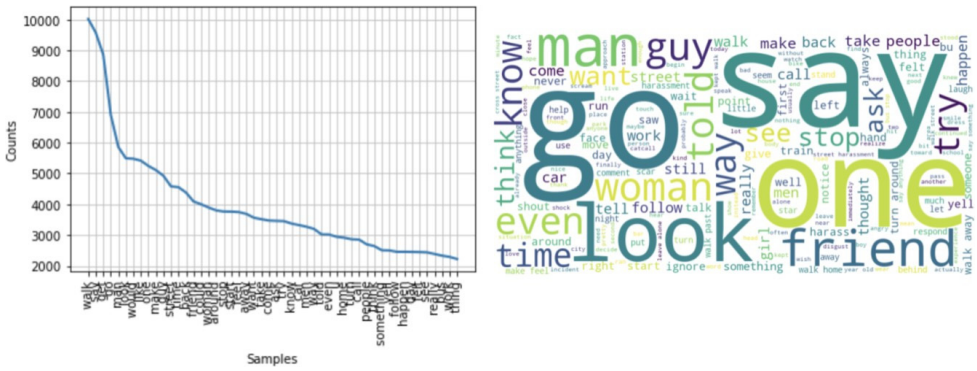

Figure 13. Bag-of-words representation model.

Another way to represent the document was through the traditional term frequency –inverse document frequency (TF-IDF) [28] technique which, as mentioned before, allows us to see which words are more salient in a document, in relation to others, as the example of Figure 14. As with the bag-of-words technique, this representation was used later as an input for the algorithms.

```
['place', 'actually', 'encounter', 'little', 'sexual', 'harassment', 'street', 'grateful', 'able', 'exist',
'public', 'without', 'visual', 'appraisal', 'week', 'stop', 'expect', 'one', 'day', 'walk', 'bike', 'fairly',
'unpopulated', 'sidewalk', 'middle', 'afternoon', 'man', 'walk', 'behind', 'caught', 'say', 'disrespectful',
'pretty', 'backside', 'ugh', 'bike', 'way', 'block', 'yell', 'back', 'disrespectful', 'could', 'hear', 'murmur',
'well', 'still', 'pretty', 'submit']

disrespectful
appraisal
unpopulated
visual
bike
```

Figure 14. TF-IDF results.

Another technique to reduce the dimensionality of the data is LSA [29], and it is based on the use of singular value decomposition; more specifically, the reduction is conducted through the truncated SVD. In this project, LSA was applied to see how the text would be divided into two generative topics according to the words contained. This reduction was conducted with the truncated SVD model of the decomposition part of sklearn [30] and with TF-IDF as the text representation.

After its application, the following results were obtained: the most important and characteristic words of topic 0 are: walk, say, man, guy, guy, look, like, street; and those of topic 1 are: walk, car, yell, street, drove, hey, shout. Additionally, it can be seen that the texts are divided into two categories, with 0 being more predominant than 1. Figure 15 shows an example where the first sentence would belong to category 0 in 17% and to category 1 in 0.5%.

	0	1
place actually encounter little sexual harassment street grateful able exist public without visual appraisal week stop expect one day walk bike fairly unpopulated sidewalk middle afternoon man walk behind caught say disrespectful pretty backside ugh bike way block yell back disrespectful could hear murmur well still pretty submit	0.173634	0.067673
walk get coffee man corner make sound walk back home start monologue look good kept go shout behind shut hand full coffee muffin key would take picture ridiculous submit	0.170807	0.055620

Figure 15. LSA results.

Another dimensionality reduction technique based on machine learning is Latent Dirichlet Allocation (LDA) [31], which was implemented through the gensim library [32] and using the BoW model. In this case, we again attempted to separate the texts into two topics using weighted words. The result is as follows: topic 0: 0.020 * walk + 0.015 * say + 0.014 * get + 0.011 * man + 0.011 * go, and topic 1: 0.014 * say + 0.013 * walk + 0.013 * get + 0.012 * go + 0.009 * make.

In Figure 15, we can see that the themes of category 0 largely coincide with those of LSA. Category 1 differs and tends to repeat words from category 0. This may be caused by the use of BoW instead of TF-IDF.

Finally, it is important to note that the natural division into two topics was not conducted according to the desired criterion of whether there is a token or not. Therefore, classification algorithms with supervised learning had to be implemented in order to perform it.

The best-known technique for vector representation of documents is Word2Vec [33], and during the project, it was applied to the database to see the representation of these words. For this purpose, the gensim model library was used [32]. One of the results is that each word has a unique representation.

Another result is that each word has a vectorial representation of 100 dimensions, passed as a parameter to the algorithm.

Word2vec is very useful if you want to find words similar to a given word, and therefore words that will be found in the same contexts. For example, if one were to search for the words most similar to “beautiful”, one would obtain the result in Figure 16.

```
[('gorgeous', 0.9526709318161011), ('baby', 0.9347470998764038),
('hello', 0.916921854019165), ('nice', 0.9122025966644287),
('sexy', 0.90898597240448), ('damn', 0.8988128900527954), ('hey',
0.8908101320266724), ('hi', 0.8842440247535706), ('babe',
0.879196047782898), ('cute', 0.8746659755706787)]
```

Figure 16. Word2vect results for similar words to "beautiful".

As it can be seen, the algorithm is largely correct with synonyms such as "gorgeous", "nice", "sexy" and "cute". However, it does include some words that are often directly related to the word but not similar to it. For example, "hi" is very often used in "hi beautiful" or "babe" for phrases such as "beautiful babe". These expressions can often be found in descriptions, and therefore, despite not being synonyms, the algorithm considers them similar.

4. Classification Models

We designed classification models for three different scenarios: full descriptions, simplified descriptions and both descriptions. For each scenario, five classification models were developed: logistic regression, naive Bayes, support vector machines, SGD (stochastic graph descent) classifier and k-nearest neighbors.

4.1. Full Description Models

Since the classification models developed require supervised learning, it was essential to have a number of perfectly labeled descriptions to build the models. Therefore, avoiding unsupervised techniques, 99 textual descriptions were read and labeled manually. Within the set of 99 events, 33 had presence of witnesses, and 66 did not. These descriptions were evenly divided into two sets, the training and the test set. Two types of feature extraction techniques were applied to the descriptions: TF-IDF and BoW, from the sklearn feature-extraction package [30], and a binary model was applied to the labels with the LabelBinarizer module of the sklearn preprocessing library. After fitting and computing the models several times, it was found that the most efficient parameters for feature extraction were the minimum and maximum df at 0 and 1 and the number of n-grams limited to 1.

Subsequently, different classification models were built to check which one is the most efficient for the given database. These models are: logistic regression, naive Bayes multinomial, support vector machines (SVMs), k-nearest neighbors and an SGD classifier. The evaluation of the models was based on the metrics related to the "with witness" category. Considering the purpose of Hollaback!'s awareness campaign, the goal of this first automatic analysis was to conduct a "pre-selection" of descriptions with bystanders in order to avoid the marketing staff reading thousands of stories. Having said this, the most important metric is the sensitivity of the true class (in our case, the true positive, real descriptions with a witness predicted as true).

In order to conduct a more precise comparison, all the metrics of the exposed models are collected in Figure 17. These metrics are calculated based on the confusion matrices of each model. The acronyms ST and CT refer to the categories "with witness" or "without witness". Although the table collects information from both categories, the one that will be taken into account is "with witness". The models evaluated are logistic regression (LR), multinomial naive bayes (NB), a support vector machine (SVM), stochastic gradient descendant (SGD) and k-nearest neighbors (KNNs) as some of them had been tested before in different fields, returning good results [13].

Model	Accuracy	Precision w/o bystander	Precision w/ bystander	Sensitivity w/o bystander	Sensitivity w/ bystander	F1-score w/o bystander	F1-score w/ bystander
LR Bow	0,72	0.70	1.00	1.00	0.17	0.82	0.29
LR tfidf	0,66	0.66	-	1.00	0.00	0.80	-
NB Bow	0,66	0.70	0.50	0.85	0.30	0.77	0.38
NB Tf-Idf	0,66	0.66	-	1.00	0.00	0.80	-
SVM bow	0,72	0.75	0.61	0.85	0.47	0.80	0.53
SVM tfidf	0,68	0.68	0.66	0.97	0.11	0.80	0.19
SGD bow	0,68	0.68	0.66	0.97	0.11	0.80	0.19
SGD tfidf	0,68	0.68	0.66	0.97	0.11	0.80	0.19
KNN bow	0,70	0.71	0.63	0.90	0.29	0.79	0.40
NKK tfidf	0,70	0.71	0.63	0.90	0.29	0.79	0.40

Figure 17. Results of the different models using full descriptions.

All models have an accuracy of around 70%. Moreover, except for the logistic regression and naive Bayes, they also have an accuracy of around 63% for the model with a control. This means that 63% of the time, an item drawn from the class with a control will be well classified. It can also be seen that most models have a low sensitivity for detecting tokens, the highest being the SVM for BoW with 47%. This would mean that by having a witness description, the algorithm has a 47% chance of categorizing it as such.

4.2. Short Description Models

Considering the unsatisfactory results of the first models with the full descriptions, other approaches should be considered.

The feature analysis demonstrated that full descriptions were very generic and complex. Just a couple of features out of 100 had a strong weight referring to the witness presence. Apart from those relevant automatic features, we decided to create an alternative, mostly manual selective database, considering the most representative expressions used by victims to indicate the presence of witnesses and other synonym expressions, as well as expressions that might suggest that there were no witnesses. The manually created database has only two fields, the descriptions and the labels—an example is shown in the figure below.

This database is composed of 97 labeled descriptions, of which 55 indicate the presence of witnesses and 42 indicate non-presence. As with the full descriptions, these are represented as a bag-of-words and TF-IDF and then input to all previously used models.

To avoid detailing case by case as before, and due to the similarity of the process, Figure 18 shows the results obtained through the confusion matrices.

Description	Intervention
Because there were other people in the elevator	1
Because there weren't any other people	0
There was no one	0
There was someone	1
Two ladies were next to him	1
It involved bystanders	1

Figure 18. Short textual description model.

Except for the SGDClassifier, which returns worse results, all models have the same metrics. They have an accuracy of 69%, a slightly higher precision than the models with full descriptions (67%) and a much higher sensitivity (89%). This means that for short descriptions, 89% of the time, if the description has tokens, it will be well classified, and 67% of the time, if a description is taken from those classified as having tokens, it will be well

classified. Combining these two metrics, we can see that they also have a higher F1-score (76%) than the 53% that was returned with the best model with full descriptions.

As seen in Figure 19, the simplified selective descriptions return better results than the full descriptions when training and testing the models. However, the aim of the project is to classify the existing descriptions in the database. Therefore, the next step was to test the application of the simplified and apparently more efficient models to the full real descriptions.

Model	Accuracy	Precision w/o bystander	Precision w/ bystander	Sensitivity w/o bystander	Sensitivity w/ bystander	F1-score w/o bystander	F1-score w/ bystander
LR Bow	0.69	0.75	0.67	0.43	0.89	0.55	0.76
LR tfidf	0.69	0.75	0.67	0.43	0.89	0.55	0.76
NB Bow	0.69	0.75	0.67	0.43	0.89	0.55	0.76
NB Tf-idf	0.69	0.75	0.67	0.43	0.89	0.55	0.76
SVM bow	0.69	0.75	0.67	0.43	0.89	0.55	0.76
SVM tfidf	0.69	0.75	0.67	0.43	0.89	0.55	0.76
SGD bow	0.57	0.50	0.77	0.85	0.35	0.63	0.48
SGD tfidf	0.57	0.50	0.77	0.85	0.35	0.63	0.48
KNN bow	0.67	0.69	0.67	0.43	0.85	0.53	0.75
KNN tfidf	0.69	0.75	0.66	0.43	0.89	0.55	0.76

Figure 19. Results of the different models using short descriptions.

4.3. Case of Use: Simple Description Model to Categorize Complete Descriptions

In this section, we analyze the effectiveness of the models elaborated with simplified descriptions to predict the category of complete descriptions. Again, to avoid going into detail about each model, the table of metrics obtained by means of the confusion matrices [34] is presented and analyzed in Figure 20.

Model	Accuracy	Precision w/o bystander	Precision w/ bystander	Sensitivity w/o bystander	Sensitivity w/ bystander	F1-score w/o bystander	F1-score w/ bystander
LR Bow	0.51	0.95	0.40	0.28	0.96	0.43	0.56
LR tfidf	0.40	1.00	0.36	0.12	1.00	0.21	0.53
NB Bow	0.62	0.79	0.46	0.59	0.69	0.68	0.55
NB Tf-idf	0.60	0.96	0.45	0.42	0.96	0.58	0.61
SVM bow	0.60	0.96	0.45	0.42	0.96	0.58	0.61
SVM tfidf	0.51	0.85	0.40	0.29	0.96	0.43	0.56
SGD bow	0.42	1.00	0.36	0.13	1.00	0.23	0.53
SGD tfidf	0.50	1.00	0.40	0.26	1.00	0.41	0.57
KNN bow	0.66	0.76	0.49	0.71	0.55	0.73	0.52
KNN tfidf	0.42	1.00	0.37	0.14	1.00	0.25	0.54

Figure 20. Simple model metrics applied to full descriptions.

In this case, the models greatly improved their F1-score with respect to the models with complete descriptions, due to a great increase in sensitivity. However, the accuracy and precision dropped to a great extent, having an accuracy for the category "with witness" of approximately 40%. This would imply that only 40% of the elements in the token category would be well classified. A good text processing and feature extraction applied to the complete descriptions before inference would allow us to reach the accuracy of 67% described in 4.2.

5. Conclusions

Several conclusions can be drawn from the exploratory data analysis presented in this paper. We launched this first analysis through all the information contained in our database, containing different cities and countries from all over the world. The first conclusion is that the predominant language of the reports is English, and therefore the text analysis must be carried out in English to obtain the most out of the database. The second conclusion is that there is a low rate of reports made by witnesses, which shows that there

is a need to make people more aware of the importance of intervening and reporting incidents. Intervention of a witness protects the feelings of the victim and contributes to the Hollaback! project for reducing harassment and violence. Another conclusion that can be drawn is that there is an uneven distribution of reports around the world, which may suggest the need to expand the reach of the platform to places where it is not known. Stronger emphasis on guarantees of anonymity should help to break down cultural barriers and lead to more reports being made around the world.

In the machine learning process focused on the city of Madrid and selecting English as the language, it can be concluded that, related to the text processing tools, more advanced feature engineering techniques could be applied [35]. As with many other research works, we appreciated the importance of the feature extraction process for attaining a proper performance of the machine learning models. The results in our target class ("with witness") considerably improved when we amended the feature selection process. Thus, unsupervised learning NLP techniques such as Word2Vec would be extremely useful for vocabulary augmentation based on semantic proximity and to find more selective clusters of features characterizing the descriptions with bystanders.

Multiple conclusions can be drawn from the binary classification algorithms. Regarding the different machine learning models, the behavior was very similar among all of them, except for SGD. As described in Section 4, the approach with an exhaustive feature selection had better results, reinforcing the importance of a good feature extraction in the text, not only during training but also before processing the test data. With this approach, we reached an accuracy of 68%, a precision of 67% and a sensitivity of 89% in the target class (descriptions with witnesses). Those metrics could also increase even further when augmenting the training dataset. Thus, this model could be used as a first automatic approach to filter out descriptions in which bystanders were present.

Author Contributions: Conceptualization, M.A.-P. and A.L.; methodology, C.P. and R.P.; software, M.A.-P. and R.P.; validation, A.L. and C.P.; investigation, A.L., C.P. and M.A.-P.; writing—original draft preparation, C.P., M.A.-P., A.L. and R.P.; writing—review and editing, A.L. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: All data were collected from: https://www.ihollaback.org/resources/ (accessed on 28 May 2021).

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

References

1. UCSD Center on Gender Equality and Health. *A National Study on Sexual Harassment and Assault*; UCSD Center on Gender Equality and Health: San Diego, CA, USA, 2019.
2. Stop Street Harassment. Available online: http://www.stopstreetharassment.org/resources/statistics (accessed on 28 May 2021).
3. ONU. Sustainable Development Goals. Available online: https://sustainabledevelopment.un.org/?menu=1300 (accessed on 28 May 2021).
4. Hollaback! Available online: https://www.ihollaback.org (accessed on 28 May 2021).
5. Katz, R.C.; Hannon, R.; Whitten, L. Effects of gender and situation on the perception of sexual harassment. *Sex Roles* **1996**, *34*, 35–42. https://doi.org/10.1007/BF01544794.
6. Karami, A.; Spinel, M.Y.; White, C.N.; Ford, K.; Swan, S. A Systematic Literature Review of Sexual Harassment Studies with Text Mining. *Sustainability* **2021**, *13*, 6589. https://doi.org/10.3390/su13126589.
7. Bauer, T.; Devrim, E.; Glazunov, M.; Jaramillo, W.L.; Mohan, B.; Spanakis, G. #MeTooMaastricht: Building a Chatbot to Assist Survivors of Sexual Harassment. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science*; Cellier, P., Driessens, K., Eds.; Springer: Cham, Switzerland, 2020; Volume 1167. https://doi.org/10.1007/978-3-030-43823-4_41.

8. Fan, H.; Du, W.; Dahou, A.; Ewees, A.A.; Yousri, D.; Elaziz, M.A.; Elsheikh, A.H.; Abualigah, L.; Al-qaness, M.A.A. Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. *Electronics* **2021**, *10*, 1332. https://doi.org/10.3390/electronics10111332.
9. Basu, P.; Tiwari, S.; Mohanty J.; Karmakar, S. Multimodal Sentiment Analysis of #MeToo Tweets using Focal Loss (Grand Challenge). In Proceedings of the 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), New Delhi, India, 24–26 September 2020; pp. 461–465. https://doi.org/10.1109/BigMM50055.2020.00076.
10. Haralabopoulos, G.; Anagnostopoulos, I.; McAuley, D. Ensemble Deep Learning for Multilabel Binary Classification of User-Generated Content. *Algorithms* **2020**, *13*, 83. https://doi.org/10.3390/a13040083.
11. SAS. Natural Language Processing. What It Is and Why It Matters. Available online: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html (accessed on 28 May 2021).
12. Alberich, M. Procesamiento del Lenguaje Natural. *Guía Introd*. **2007**, *2007*, 27.
13. Sánchez-Rebollo, C.; Puente, C.; Palacios, R.; Piriz, C.; Fuentes, J.P.; Jarauta, J. Detection of Jihadism in Social Networks Using Big Data Techniques Supported by Graphs and Fuzzy Clustering. *Complexity* **2019**, *2019*, 1238780. https://doi.org/10.1155/2019/1238780.
14. Latorre, J.M.; Cerisola, S.; Ramos, A.; Palacios, R. Analysis of stochastic problem decomposition algorithms in computational grids. *Ann. Oper. Res*. **2009**, *166*, 355–373. https://doi.org/10.1007/s10479-008-0476-1.
15. CUP y Hollaback! Show Up. Your Guide to Bystander Intervention. Available online: https://www.ihollaback.org (accessed on 28 May 2021).
16. Fairbairn, J. Before #MeToo: Violence against Women Social Media Work, Bystander Intervention, and Social Change. *Societies* **2020**, *10*, 51. https://doi.org/10.3390/soc10030051.
17. Puigvert, L.; Vidu, A.; Melgar, P.; Salceda, M. BraveNet Upstander Social Network against Second Order of Sexual Harassment. *Sustainability* **2021**, *13*, 4135. https://doi.org/10.3390/su13084135.
18. Carto Maps. Available online: https://carto.com (accessed on 28 May 2021).
19. Fast Company. Available online: https://www.fastcompany.com/90303329/these-apps-try-to-make-reporting-sexual-harassment-less-of-a-nightmare-do-they-work (accessed on 19 July 2021).
20. STOPit Solutions. Available online: https://stopitsolutions.com/stopit-products/ (accessed on 19 July 2021).
21. #NotMe App. Available online: https://www.not-me.com/en/ (accessed on 19 July 2021).
22. Spot. Available online: https://talktospot.com/index (accessed on 19 July 2021).
23. CSV File Reading and Writing. Available online: https://docs.python.org/3/library/csv.html (accessed on 28 May 2021).
24. Beautiful Soup Documentation. Available online: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed on 28 May 2021).
25. Contractions v.0.0.25. Available online: https://pypi.org/project/contractions/ (accessed on 28 May 2021).
26. NLTK 3.5 Documentation. Available online: https://www.nltk.org (accessed on 28 May 2021).
27. Wordcloud 1.7.0. Available online: https://pypi.org/project/wordcloud/ (accessed on 28 May 2021).
28. Salton, G.; McGill, M. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, NY, USA, 1983.
29. Thomas, L.; Peter, F.; Darrell, L. An Introduction to Latent Semantic Analysis. *Discourse Process.* **1998**, *25*, 259–284. https://doi.org/10.1080/01638539809545028.
30. Scikit Learn, API Reference. Available online: https://scikit-learn.org/stable/modules/classes.html (accessed on 28 May 2021).
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res*. **2003**, *3*, 993–1022.
32. Gensim. About Us. Available online: https://radimrehurek.com/gensim/ (accessed on 28 May 2021).
33. Goldberg, Y.; Levy, O. Word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv* **2014**, arXiv:1402.3722.
34. Juan Ignacio Barrios Arce. La Mtriz de Confusión y sus Métricas. Available online: https://www.juanbarrios.com/matriz-de-confusion-y-sus-metricas/ (accessed on 28 May 2021).
35. Puente, C.; Palacios, R.; González-Arechavala, Y.; Sánchez-Úbeda, E.F. Non-Intrusive Load Monitoring (NILM) for Energy Disaggregation Using Soft Computing Techniques. *Energies* **2020**, *13*, 3117. https://doi.org/10.3390/en13123117.