Article

Land Use Identification through Social Network Interaction

Jesus S. Aguilar- Ruiz [1,2,*] , **Diana C. Pauca-Quispe** [2] , **Cinthya Butron-Revilla** [2] , **Ernesto Suarez-Lopez** [2] **and Karla Aranibar-Tila** [2]

1   School of Engineering, Pablo de Olavide University, ES-41013 Seville, Spain
2   Universidad Nacional de San Agustin de Arequipa, Arequipa 04001, Peru
*   Correspondence: aguilar@upo.es

Abstract: The Internet generates large volumes of data at a high rate, in particular, posts on social networks. Although social network data have numerous semantic adulterations and are not intended to be a source of geo-spatial information, in the text of posts we find pieces of important information about how people relate to their environment, which can be used to identify interesting aspects of how human beings interact with portions of land based on their activities. This research proposes a methodology for the identification of land uses using Natural Language Processing (NLP) from the contents of the popular social network Twitter. It will be approached by identifying keywords with linguistic patterns from the text, and the geographical coordinates associated with the publication. Context-specific innovations are introduced to deal with data across South America and, in particular, in the city of Arequipa, Peru. The objective is to identify the five main land uses: residential, commercial, institutional-governmental, industrial-offices and unbuilt land. Within the framework of urban planning and sustainable urban management, the methodology contributes to the optimization of the identification techniques applied for the updating of land use cadastres, since the results achieved an accuracy of about 90%, which motivates its application in the real context. In addition, it would allow the identification of land use categories at a more detailed level, in situations such as a complex/mixed distribution building based on the amount of data collected. Finally, the methodology makes land use information available in a more up-to-date fashion and, above all, avoids the high economic cost of the non-automatic production of land use maps for cities, mostly in developing countries.

Keywords: land use; social networks data; natural language processing; classification

1. Introduction

The current dynamic nature of cities has generated substantial changes in urban environments, making the analysis of geo-spatial information more complex [1]. The urban management and planning of cities has been nourished for years by Geographic Information Systems (GIS) tools, since they allow storing, modeling, and analyzing geo-spatial data. On the other hand, urban planning faces challenges such as organizational changes, staff availability, updating resources, and data availability [2]. In this context, land use maps are one of the most demanded resources by authorities and researchers, for conducting urban planning and environmental sustainability studies [3]. However, the production of these maps is costly and time-consuming [4]. Consequently, the production and management of land use maps evidence a technical problem [5] caused by economic constraints, mostly in developing countries.

The Internet generates large volumes of data at a high rate, in particular, posts on social networks. Social networks have become a source of data for researchers, because these data contain information that allows studying the interaction of citizens with their environment [6]. The concept of "user-generated content" [7,8], which encompasses how citizens share excerpts from their daily lives, can be translated to different application domains: the representation of urban boundaries, relationships between weather conditions and

traffic, citizen's activity patterns, urban transportation behavior and land use [6,9], urban planning, and spatial planning. Therefore, social networks are a valuable source of data that could be used to identify interesting aspects of a city's land use.

Although social network data have numerous semantic adulterations, and are not intended to be a source of geo-spatial information, in the text of posts we find pieces of important information about how people interact with their space, and can complement other sources in the process of identifying land uses [10]. These data can be composed of various fields such as message text, location, and images. In this context, various approaches have been developed and applied to analyze this data type; the semantic analysis of the text of the messages—unstructured information—being the basis of much of the research related to social networks. However, it is an underexploited resource in obtaining and analyzing geographic information [11]. Current land use identification and analysis studies rely primarily on publication metadata, such as time and geographical coordinates—structured information.

The various studies on land use classification have rarely considered the content of posts as a source of data to infer the type of use. In the content of social network posts, people communicate attitudes, activities, relationships, and other emotional states that are often complemented with a geo-spatial location. In written expressions it is possible to find words that allow inferring information about where a user is situated. For example, if a person posts the following message "having lunch with friends" on a social network, we could assume that they are in a restaurant, a house, a shopping mall, or another place where they can have lunch. In addition, if we associate two or more words such as "having lunch with friends in a restaurant", the connector "in" and the word "restaurant" provide us with a high probability that the place where he/she is located is, in fact, a restaurant. If the coordinates associated with the message are included, we can know the space where the restaurant is placed, and consequently, the type of land use to which that space belongs. Therefore, the potential of the content of social network publications to address the study of land use is remarkable.

Nevertheless, the language used in social network publications is very informal, includes many idioms, expressions specific to each region, misspellings, or other language deformations. For this reason, the processing of this type of information with specific objectives becomes an important challenge [12], in which numerous algorithmic techniques will concur, to form a methodology capable of extracting the valuable knowledge.

Most studies on land use are mainly oriented to classification tasks. When we use a data source such as the expressions of a language, it is necessary to propose a methodology that starts with data collection, followed by pre-processing and, finally, classification. The fundamental reason is that the expressions in each language and country—and even region—are different. For example, in order to carry out the present study, it was necessary to create a corpus and a dictionary for the Spanish language and for expressions specific to Peru.

This paper proposes a methodology for the identification of land uses using Natural Language Processing (NLP) from the contents of the popular social network Twitter. It will be approached by identifying keywords with linguistic patterns from the text, and the geographical coordinates associated with the publication. The methodology is composed of the following stages: data collection, corpus and dictionary creation, pre-processing, feature extraction, learning, prediction, and representation. At each stage, context-specific innovations are introduced to deal with data from South America and, in particular, the city of Arequipa, Peru. The objective is to identify the five main land uses: residential, commercial, institutional-governmental, industrial-offices, and unbuilt land.

Within the framework of urban planning and sustainable urban management, the methodology proposed in this study contributes to the optimization of the identification techniques applied for the updating of land use cadastres, since the results achieved an accuracy of about 90%, which motivates its application in the real context. In addition, this model would allow the identification of land use categories at a more detailed level,

in situations such as a complex/mixed distribution building based on the amount of data collected. Finally, the methodology makes land use information available in a more up-to-date and, above all, much less costly way.

The document is organized as follows: first, a review of the recent scientific literature related to the use of machine learning techniques applied to the categorization of land use, and specifically involving data from social networks, followed by an overview of the NLP-based methodology; next, a description of the application context in the city of Arequipa, containing examples of the results provided by the main phases of the methodology, together with the results achieved in comparison to several models; then, visual examples are presented to illustrate the contrast between results from our approach and what is registered in the cadastre; finally, conclusions and future work.

2. State-of-the-Art

2.1. Land Use and Urban Planning Applications

Recent proposals for urban growth cover aspects such as the potential exploitation of land uses: those that allow to accommodate a high urban diversity. Land uses are thus defined as a measure of the diversity of uses contained in a given space [13], which allows in a specific way to identify nearby uses or nearby activities developed in a limited spatial range. The term *land use* refers to the employment that human beings give to a portion of land based on the activities carried out in it. However, in order to use land uses in urban management and planning, an updated map with the most recent and accurate information is necessary [4,14] to make the right decisions [15]. Under this situation, several terms emerge, such as *urban computing* [16] and urban planning applications [17], which have become an emerging research area where urban problems in cities are studied using different data sources such as electronic devices—mainly smartphones, location-based social media, web pages—e.g., analyzed with Robotic Process Automation technologies, among other digital information.

Many approaches include the use of images (e.g., from satellites), sometimes together with some structured data (e.g., points of interest). For example, Liu et al. [1] use natural physical features from high spatial resolution images (HSR) and socio-economic semantic features (frequent characteristic words related to an urban land type) from social data to create a dictionary of land use words, including data from multiple sources such as OpenStreetMap road networks, Gaode's Points of Interest, and Tencent's real-time user density. On the other hand, Zhan et al. [18] use large-scale Twitter log data, and they propose a pre-processing method in which the raw data only contains coordinates and activity category information. One of the few papers that mentions data pre-processing is [8], which used Twitter posts and metadata. From each tweet, the authors use text to analyze whether a space is a restaurant, airport, or stadium. They perform pre-processing of the text of each tweet before using a term frequency-based technique, previously removing emoticons and other non-ASCII characters, as well as hashtags and the "@" character.

2.2. Social Networks and Text Mining

The incremental use of social networks has been producing large amounts of data that are being extracted, analyzed, and structured [11]. This allows gathering useful information that can reflect different areas such as the human dynamics in a city, identifying how people live and interact with the environment, health applications, natural hazards management, tourism, environmental monitoring, crimes and disturbances [19].

The ways to extract the information offered by a publication on social networks range from the analysis of its metadata (geo-tags, time and date, username, place name, etc.), the profile of the user who publishes it, to the application of text mining to the message of the publication [12]. The message is considered the backbone of many of the investigations related to the inference of location because of its enormous potential [20], apart from presenting several scientific challenges. Since the texts of social media are mostly generated by mobile devices and have no writing restrictions, it gives the user a great margin of

typographical error and brevity. These texts also include links, emoticons, use of informal language and idioms, spelling and grammatical errors, presence of user mentions and hashtags, HTML tags, use of acronyms and abbreviations. Therefore, the main challenge is to clean up the large amount of noise in the messages, in addition to deal with the unstructured format, in contrast with articles, web pages, or blogs, that have more content and make use of conventional grammar and semantic rules.

2.3. Text Mining

Text mining is a research field that aims to automatically discover or extract new knowledge from texts written in natural language [21,22]. According to the objective to be achieved, the text mining process has a set of stages that are classified into data collection, pre-processing, transformation, and analysis.

2.3.1. Data Collection

The collection is the first step of the text mining process, where unstructured data is captured from different sources such as blogs, reviews, news, and publications on social media, among others. The data is stored for future pre-processing and analysis.

2.3.2. Pre-Processing

The objective is to obtain clean and actionable data from the data collected through the application of different debugging and cleaning techniques. There is no specific order to perform the pre-processing task, so this must be determined empirically [23], experimenting with each technique individually, comparing the results, and combining those that perform best. The pre-processing of social media data presents certain challenges due to the use of informal language and the length of the messages (for instance, tweets are very short in length) [24].

The techniques commonly used in text pre-processing are the following [10,11,25–28]:

1. Normalization and noise reduction: this process unifies the text and removes irrelevant and meaningless elements. The literature recommends the deletion of URLs, HTML tags, special characters, emojis, or emoticons.
2. Remove punctuation marks: as punctuation marks do not add additional information, the elimination of them helps reduce the size of the data and increase the model efficiency.
3. Remove stopwords: these are terms that help build sentences but do not commonly provide meaning, like prepositions or articles.
4. Processing of abbreviations, acronyms, and entities.
5. Spell correction.
6. Tokenization: it splits text into significant words delimited by blank spaces, commas, periods, or any other special character.
7. Lemmatization: it is the search for the words' lemma to unify the terms that give the same information (for example, derivational forms of verbs).
8. Part of Speech Tagging (PoS): grammatical tagging of each word.

2.3.3. Data Transformation

Data transformation involves the extraction and selection of characteristics (SC) from text data (string) to generate a suitable representation for computational learning. In classification, the transformation is crucial as the SC directly impacts the result of the classifier. The selection depends on the type of document and the classification chosen. The objective is to find characteristics with relevant information that improve the precision of the classifier. Then, the text is represented as a value using a binary representation or as a SC technique, such as Term Frequency (TF), Inverse Document Frequency (IDF), or Term Frequency-Inverse Document Frequency (TF-IDF). From these representations, the global feature space is generated from all the texts used for training. The next paragraph presents some techniques to obtain the characteristics of a text:

1. Term Frequency-Inverse Document Frequency (TF-IDF): it seeks to find the balance between the Frequency Term (importance of a word w_k locally in a text T) and IDF (a global measure of the importance of w_k in the corpus) [27].

2. N-grams: they are sequences of words grouped according to the value given to N. For example, the 1-grams (unigrams) of T = *"Tomorrow is Thursday"* are $W_1^T = \{$*Tomorrow, is, Thursday*$\}$, the 2-grams (bigrams) are $W_2^T = \{$*Tomorrow is, is Thursday*$\}$; then, given a text T with m words, a set of n-grams of size $m - n + 1$ is obtained.

3. Bag-of-PoS: application of n-grams in the PoS labels of the text to be analyzed.

2.3.4. Data Analysis

There exist different methods to analyze unstructured data, which are classified into information extraction, summary, grouping, and categorization [21,29]. The present study uses categorization in order to classify a set of documents into topics or categories, which requieres a correct combination of NLP and machine learning techniques.

In the field of machine learning, classification tasks are mainly supervised, which consist of training the system and then testing with information about the classes before the real classification process [30]. Among the most popular techniques are Logistic Regression, Naive Bayes, Support Vector Machines, Decision Trees, Neural Networks, and k-Nearest Neighbor. The Naive Bayes approach is one of the most employed classifiers for analyzing text documents. It works under the principle of using the probabilities of words and categories that allow determining the classes for given documents, and assumes a property that makes the technique very fast: conditional independence among variables with respect to the target variable. In summary, Bayesian classifiers are simple and powerful in terms of the degree of certainty, and therefore a good choice for approaching NLP problems.

3. Materials and Methods

The area of study is the city of Arequipa, which is the second most populated urban area in Peru, with more than a million of inhabitants and a yearly growth rate of 2.3%, according to the last census conducted in 2017. This study covers the historical center of Arequipa (see Figure 1), with an approximate area of 3.46 square kilometers and about 2251 properties located within 56 blocks. The categorization made is based on the real estate properties declared in the area.
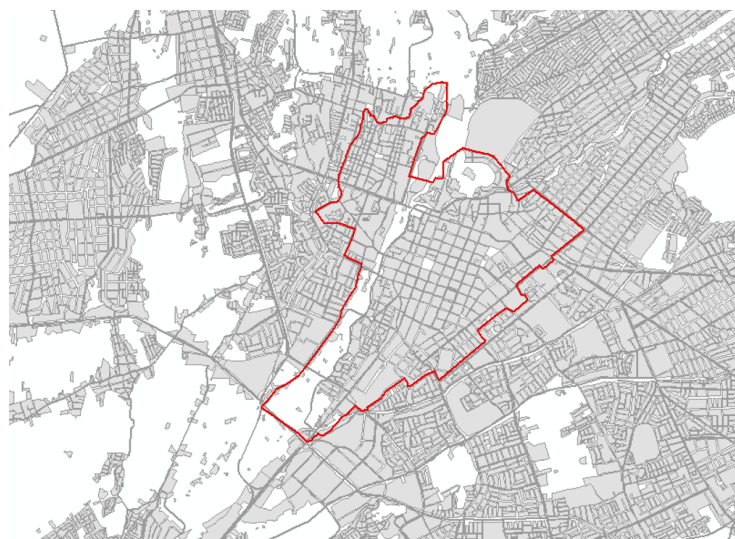

Figure 1. Map of the historical center of the city of Arequipa (own elaboration from data based on PlaMCha, 2017–2027).

The Master Plan for Arequipa's Historic Center and Surrounding Area (PlaMCha 2017–2027) describes the geographical characteristics of the historical center, which has

three defined zones; old zone, monumental zone, and buffer zone located in the coordinates S: 16° 23′53.33″ W: 71° 32′12.67″ [31]. The oldest area, where the main monuments and the Main Square of the city are located, is called Damero (transl. checkboard) and is constituted by blocks of 111.4 m per side and separated by streets of 10.3 m, which gives the characteristic unitary image and covers an area of 1.41 km^2 (141 ha.). The Monumental Zone has 2.12 km^2 (212 ha.), and the Buffer Zone has 3.46 km^2 (346 ha.) and covers the first 2 zones.

The historical center was selected because it brings together different types of land use such as historical, artistic, cultural, workplaces, residential, commercial, and public spaces. The large influx of tourists and residents in this area makes it possible to collect a large amount of data from social networks, unlike in other parts of the city.

3.1. Data Sources

3.1.1. Types of Land Use

The land use map of the historical center of Arequipa, made by the PlaMCha, defines 14 categories of land use (see Table 1—left). However in this work, the re-categorization carried out by [13] is considered, which initially takes into account 48 categories of land use and then groups them into 5: residential, commercial, industrial-offices, institutional-government, and unbuilt land (see Table 1—right). The residential category groups houses, apartments, and condominiums. The commercial category comprises properties that carry out typical activities of stores, bookstores, shopping centers, restaurants, bars, hotels, lodgings, parking lots, and entertainment venues. The industrial-offices category groups industrial, companies and offices land uses. The institutional-government category includes buildings related to education (institutes, academies, schools and universities), health (hospitals and clinics), cultural (cultural centers and museums), management (administrative, financial or governmental grounds), and religious centers. Finally, the category of unbuilt land is mostly composed of agricultural land (crops), vacant land (buildings that do not exceed 1% of the surface), the Chili river course—which runs north–southwest through the historical area.

Table 1. Re-categorization of land use categories [13].

Categories	Re-Categorization
Residential	Residential
Commerce Lodging Parking	Commercial
Industry	Industrial-Offices
Education Health Cultural Administrative Religious	Institutional-Governmental
Vacant land Crop Chili River Others	Unbuilt land

3.1.2. Cartographic Map

This study focuses on determining the categories of land use in the historical center of Arequipa; hence, it is necessary to define the boundaries and characteristics of the area as a function of polygons.The information is obtained from the documents developed in the PlaMCha, and the cadastral maps of the historical center developed by the Technical Team of the Municipal Planning Institute in the Pilot Project named "Altura para la Cultura" (transl. Height for Culture) [31]. Cadastral maps were provided as GIS data, representing set of polygons for blocks and lots in the historical center (see Figure 1).

3.1.3. Geo-Tagged Twitter Publications

Tweets allow the categorization of land uses where tweets were generated. A tweet is a short message of 280 characters maximum composed of text, emojis, and attachments, published on the Twitter platform, which is considered one of the largest sources of information fed by millions of users [32]. Twitter Application Programming Interfaces (APIs) allow capturing tweets encoded in JavaScript Object Notation (JSON) format with their associated attributes and values. A tweet can have around 150 associated attributes, although this research only requires the ID, user ID, text, timestamp, geodata, and language of the message. All collected tweets contained the precise location (latitude and longitude).

3.2. Experimentation

The methodology is divided into a sequence of tasks, which begin with the data collection, followed by splitting the data for building the corpus and for validating. The validation data go through a previous classification using PoS patterns before entering the classification algorithm. For clearer interpretation, the results of the tagged data are presented on a map of the study area according to the related coordinates of the tweet. Finally, quality metrics are shown to validate the approach. The methodology is graphically illustrated in Figure 2.
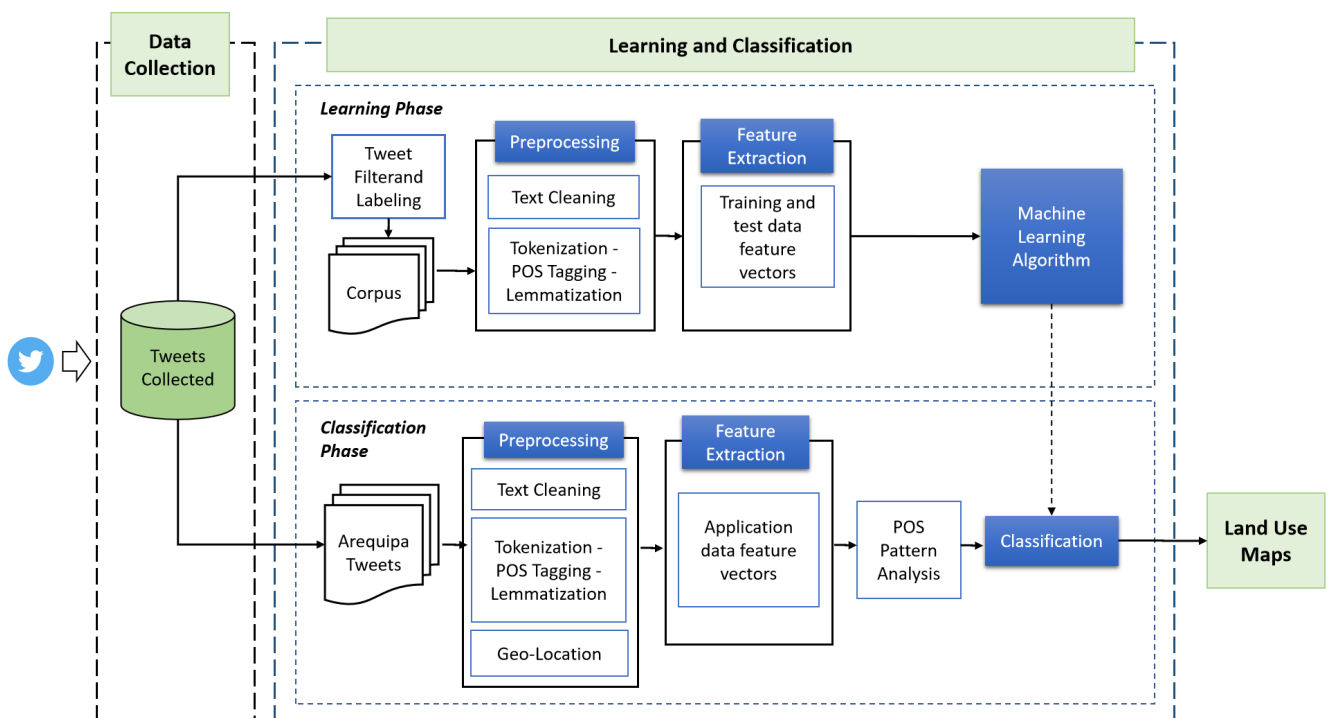

Figure 2. Outline of the proposed methodology.

Data Collection

In this phase, the model connects to the Twitter Streaming API using the Tweepy library to capture geo-referenced tweets across South America. The Streaming API has certain limitations, but it allows the download of 100% of tweets that meet the defined filter, as long as they are less than 1% of the global volume of publications at a given time [33]. The application was run on a server for a period of 10 months (from April 2019 to February 2020—up to the beginning of the pandemic) and data was stored in a MySQL database.

A corpus related to land use categorization was not available in Spanish, so it was necessary to build one. The collected data was divided into two sets: the first set consists of 3870 tweets located in the historical center of Arequipa, that will then be used to identify the type of land use. The second set is used for the creation of the corpus and consists of 42,318 tweets located anywhere in South America.

These tweets are previously processed to remove the duplicates, blanks, single word, and only numbers. As a result, we obtained a total of 24,995 messages that were semi-automatically categorized into a land use type according to their content and geographical coordinates, leaving a total of 4538 useful tweets in different languages, which were randomly split into training and test. The result was distributed into the defined categories and divided into subcategories to avoid data imbalance (see Table 2).

Semi-automatic categorization consisted of two phases: automatic categorization using the Freeling library (successful in about 73%), and manual categorization of the remaining tweets, examining the presence of keywords (restaurant, school, park, house, etc.), along with the occurrence of specific sequence of tags (e.g., verb-preposition-noun).

Table 2. Distribution of classes in the corpus.

Commercial			Institutional-Government			Industrial Offices	Residential	Unbuilt Land
2539			1212			138	219	431
Commercial	Commercial Restaurant	Commercial Service	Institutional	Institutional Education	Institutional Cultural	Industrial Offices	Residential	Unbuilt land
1177	874	488	497	371	344	138	219	431

3.3. Pre-Processing

The pre-processing is divided into two phases: the first phase consists of cleaning the text to remove noise and correct data, and the second phase is the data filtering to identify only the tweets which are within a block of the historical center. The tweets to be used as classifier input go through the two pre-processing phases, while the corpus tweets only go through the first phase. The steps are detailed as follows:

1. Noise removal, language detection, and Spanish translation. URLs, symbols, and HTML tags are removed [23,34]. For translation and detection, the *Googletrans* library for Python is used, which allows processing large amounts of records without a query limit; the translation is of the entire text and not word by word so that the result is a meaningful sentence, as Google Translate Ajax API does. Hashtags and mentions are excluded from the translation.
2. Hashtag processing. Hashtags are words that tag a tweet to a topic that is generally related to its content [35]. In this study, these words are preserved and translated individually if they are in English.
3. Elimination of punctuation marks and replacement of mentions. All punctuation marks are removed, except for the symbol @ which is replaced by the word *in*, which is useful in the context of location extraction from a text [8]. However, this action requires additional processing with a dictionary, elaborated with frequent accounts in the corpus and others specific to the region, so that this substitution provides value to the classification.

An example of the result of applying the first three techniques is shown in Table 3.

1. Processing of Abbreviations, Acronyms, Slang, and Establishment Names. The most recurrent abbreviations and slangs within the corpus are identified and stored for the construction of a dictionary. The names of establishments, institutions, and acronyms identified from Google Maps are added to this list, which includes bars, pharmacies, hotels, universities, etc. located within the Historical Center of Arequipa. The resulting dictionary is used to identify these words in the text of the publication and replace them with the related word [27]. The result is shown in Table 4.

Table 3. Text output after noise removal, translation, punctuation removal, and replacement of hashtags and mentions.

Original Text	Pre-Processed Text
I'm at Mallplaza Bellavista—@mallplazaperu in Bellavista, Callao https://t.co/brtyxSe8CY	**Estoy en** Mallplaza Bellavista **en mallplazaperu en** Bellavista Callao
work breakfast ! #friends #meeting en Universidad Jorge Tadeo Lozano https://t.co/sNYJhxG6cw	**Trabajo desayuno amigos reunión** en Universidad Jorge Tadeo Lozano
Un dia cualquiera en Cevicheria Karloncho Oficia https://t.co/f9kdEEwdMx	Un dia cualquiera en Cevicheria Karloncho Oficia
He venido a que mami me atiborre de comidaaaaaa (@Residencial Parque Central in Lima) https://t.co/dCCBbEgvZm	He venido a que mami me atiborre de comidaaaaaa **en Residencial** Parque Central **en** Lima

Table 4. Text output after processing abbreviations, acronyms, slang, and establishment names.

Original Text	Abbreviation Processing
I'm at Mallplaza Bellavista—@mallplazaperu in Bellavista, Callao https://t.co/brtyxSe8CY	estoy en **centro comercial** bellavista en **centro comercial** en bellavista callao
work breakfast! #friends #meeting en Universidad Jorge Tadeo Lozano https://t.co/sNYJhxG6cw	trabajo desayuno amigos reunión en universidad jorge tadeo lozano
Un dia cualquiera en Cevicheria Karloncho Oficia https://t.co/f9kdEEwdMx	un dia cualquiera en **restaurante** karloncho oficia
He venido a que mami me atiborre de comidaaaaaa (@ Residencial Parque Central in Lima) https://t.co/dCCBbEgvZm	he venido a que **mamá** me atiborre de comidaaaaaa en residencial parque central en lima

2. Spell Checking. The open-source tool Hunspell is used (embedded in LibreOffice, Mozilla Firefox, Thunderbird, and Google Chrome, and some proprietary software packages). It is a spell checker designed for languages with complex morphology, such as Spanish [34]. The concept of *Longest Common Subsequence* (LCS) is applied to each of the options proposed by the spell checker to improve the result. Hunspell checks that each token in a publication is a valid word in the language; otherwise, it is replaced if the LCS value of the word proposed by the proofreader is not less than 71%; if it is less, the word is deleted (see Table 5).

Table 5. Spell-checking process.

Word to be Corrected	Hunspell Correction Options	Choosing the Best Option Using LCS
sapato	['apasto', 'zapato', 'patoso', 'topatopa', 'sato', 'pato']	apasto = 50.0% **zapato = 83.33%** patoso = 66.66% topatopa = 66.66% sato = 50.0% pato = 66.66%
ClubMilita	['Club Militar', 'Club-militar', 'Militarizar']	**Club Militar = 100.0%** Club-militar = 54.545% Militarizar = 63.636%
casiita	['casinita', 'casiterita', 'marcasita', 'canastita']	casinita = 57.142% casiterita = 57.142% marcasita = 57.142% canastita = 42.857%
Munays	['Ayunas']	Ayunas = 50.0%

3. Stopwords. All stopwords are removed, except the words "en" (transl. *in*) and "de" (transl. *from*) because these are spatial indicators that help identify that a user is posting the tweet from a location.
4. Lemmatization and PoS Tagging. All publications are pre-checked and cleaned before lemmatization for better results. We used the Freeling tool, which is a library providing language analysis functionalities (morphological analysis, named entity detection, PoS-tagging, parsing, etc.) for a variety of languages [36]. The results are shown in Table 6.

Table 6. Text output after lemmatization and PoS tag with Freeling.

Original Text	Lemmatized Text	Part of Speech
I'm at Mallplaza Bellavista - @mallplazaperu in Bellavista, Callao https://t.co/brtyxSe8CY	estar en centro comercial bellavista en centro comercial en bellavista_callao	estar/VMI en/SP centro/NC comercial/AQ bellavista/NP en/ SP centro/NC comercial/AQ en/SP bellavista_callao/NP
work breakfast! #friends #meeting en Universidad Jorge Tadeo Lozano https://t.co/sNYJhxG6cw	trabajo desayuno amigo reunión en universidad jorge_tadeo_lozano	trabajo/NC desayuno/NC amigo/AQ reunión/NC en/SP universidad/NC jorge_tadeo_lozano/NP
Un dia cualquiera en Cevicheria Karloncho Oficia https://t.co/f9kdEEwdMx	día cualquiera en restaurante pescado oficia	día/NC cualquiera/PI en/SP restaurante/NC pescado/NC oficiar/VMI
Terminando de cantar la Santa Misa dominical #santamisa #musicacatolica #singer #catholic #church en Capilla Jesus Hostia https://t.co/pLpvHwRNYh	terminar cantar santo misa dominical cantante iglesia en capilla jesús_hostia	terminar/VMG cantar/VMI santo/NC misa/NC dominical/AQ cantante/NC iglesia/NC en/SP capilla/NC jesús_hostia/NP
He venido a que mami me atiborre de comidaaaaaa (@ Residencial Parque Central in Lima) https://t.co/dCCBbEgvZm	venir mamá atiborrar en residencial parque central	venir/VMP mamá/NC atiborrar/VMS en/SP residencial/NC parque/NC central/AQ

After pre-processing, the tweets belonging to the application data are selected according to their geolocation (South America). Each tweet positioned in the radius of the city of Arequipa goes through two conditions: the first condition selects the tweets that are located within the polygon that represents the historical center, and the second condition selects the tweets that are located in a polygon that represents a block. To perform the first filtering, the *shapefile* containing the representation of the historical center is imported into the PostgreSQL database and the tweets geo-positioned in Arequipa (ID, latitude, and longitude) are loaded into a *tweet_gps* table. To select the tweets that are located in the polygon, the coordinates are transformed into *geometry* type data with the function *st_geomfromtext*. For instance, the SQL query used returned a total of 2343 tweets distributed in the plane as shown in Figure 3. For the second filter, the polygons of the blocks are stored in the PostgreSQL database as *geometry* type data. Tweets that are located in the middle of the street or outside the historic center are removed, providing 924 tweets (see in Figure 4a zoomed area).
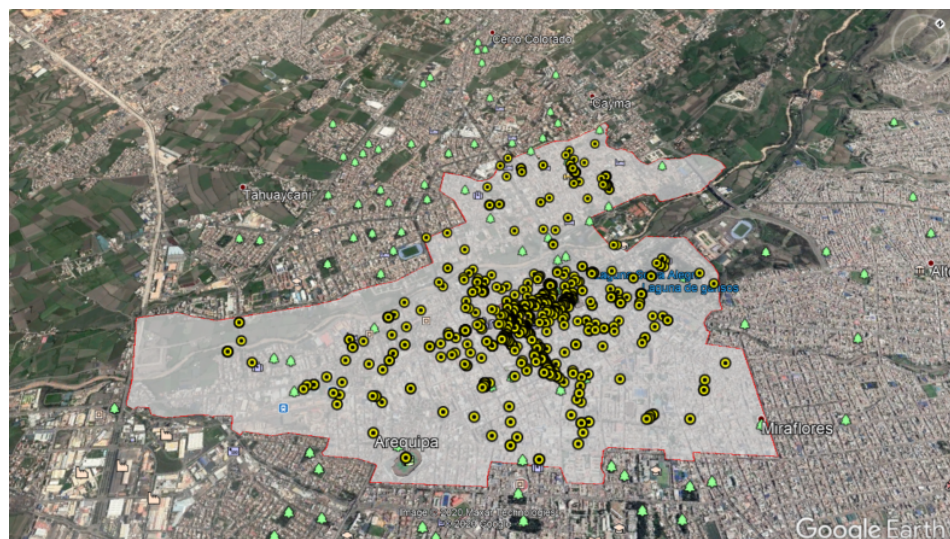

Figure 3. Data located in the historical center.

Figure 4. Detailed example from the second cleaning process.

3.3.1. Feature Extraction

Feature extraction deals with the transformation or addition of features to obtain a different, possibly larger and enriched feature space. Using only one feature extraction method does not guarantee the best results according to [10]. As a consequence, this research explores the use of TF-IDF values associated with a multidimensional vector, the *n-grams* (unigrams, bigrams, and trigrams), which have shown to work well in the clas-

sification of documents [37], and *Bag-of-PoS* that works similarly to the *n-grams*, but based on the sequence patterns of the Part-of-Speech labels.

For example, according to [38], the *"Verb-Preposition"* pairs are commonly followed by the name of the place where a user is located, which allows it to be used as an indicator to identify whether the user is talking about an establishment or location. Once the data is numerical, the categorization of tweets becomes a standard machine learning classification problem.

3.3.2. Classification

The Multinomial Naïve Bayes Classifier (MNB) is one of the most popular algorithms in social media data and text categorization [24,37]. It is based on the evaluation of the probabilities of each class. Each tweet in the corpus has associated vectors with the extracted characteristics and their respective label of the land use type; therefore, the classifier is trained with each one of the characteristics, evaluating them individually, and later different combinations are examined.

Unlike applying the classifier to the corpus data which was filtered in the labeling process ensuring that they all refer to one location, when using the classifier with the application data, there is a need to previously identify which tweets refer to the user being in one location and which refer to other topics. For that purpose, with the tagged and untagged corpus data, we selected the PoS sequences with Bag-Of-PoS and identified the *i-most frequent* sequences that represent each data set, as it is done in [39]. Depending on the presence or not of the sequences of each class in the text ([it is/it is not] in a location), the tweet is classified by the MNB approach.

3.4. Results
3.4.1. Land Use Categorization

Class imbalance and expected results are some of the considerations taken into account in the selection of metrics to measure the performance of a classifier. There is no one metric that measures well the performance of a classifier in every scenario. In this research, the classifier is of the multi-class type (many possible results that are mutually exclusive) [40], so it is also important to calculate the metrics for each class $c_1 \ldots c_n$ individually, and then calculate the classifier metrics generally, as proposed in [27]: accuracy, precision, recall, and F1-score. Since the classes are unbalanced (some classes have more data, so they are more likely to appear than others) the prioritized indicator is the F1-score [37]. Results for each model are shown in Table 7.

Table 7. Metrics with different combinations of characteristics.

Feature	Accuracy	Precision	Recall	F1-Score
TF-IDF	0.830	0.911	0.702	0.744
TF-IDF/lemma	0.836	0.907	0.730	0.772
Unigram	0.884	0.896	0.810	0.838
Unigram/lemma	0.884	0.886	0.832	0.851
Bigram	0.837	0.887	0.770	0.805
Bigram/lemma	0.843	0.868	0.784	0.813
Trigram	0.615	0.843	0.492	0.559
Trigram/lemma	0.651	0.855	0.551	0.624
N-gram (1,2,3)	0.894	0.899	0.844	0.863
N-gram (1,2,3)/lemma	**0.904**	**0.900**	**0.870**	**0.880**

The corpus is used with two variations: one, with the text in its lemmatized form and the other with the original text, without lemmatization. The values achieved with these two variations are very close to each other, but the use of the lemmatized text always provides slightly better results (see Table 7, where two rows for each feature is shown, with and without lemma). Furthermore, comparing the results of the models it is clear that by using all n-grams (unigram + bigram + trigram) together the highest accuracy is achieved. Therefore, this model is chosen for the classification of application data resulting in tweets labeled to be positioned in a Land Use Map (a sample of the results achieved with real data is shown in Table 8).

Table 8. Classification results with real data.

Original Text	Latitude	Longitude	Label
Buenas noches Arequipa!! (@ Zig Zag in Arequipa) https://t.co/SY97QrTo9e https://t.co/v4UNcnljT0	−16.39525055	−71.53541831	Commercial
Amando conocer este pais [?] en Plaza de Armas de Arequipa https://t.co/rBZ4Dmw0iI	−16.39869651	−71.53693914	Unbuilt land
#Arequipa #Arte #Concierto de bienvenida en la inauguracion de la #Exposicion #Raices @ Centro De Las Artes De La Ucsp https://t.co/f76n4QdHks	−16.3998186	−71.5393672	Institutional
Tarde de películas en casa #amor #dulcehogar #movie time	−16.399428	−71.539881	Residential
Saliendo de la oficina #work en Galeria San Jose https://t.co/BvkDpu2zi6	−16.3988135	−71.5318563	Industrial
#Plus135: Que son las condiciones objetivas de punibilidad? https://t.co/64doI8zoLI	−16.3902511	−71.5360128	Non-classified

Out of the 924 tweets used by the classifier, 327 were classified in the commercial category, 248 in the institutional category, 35 in residential, 14 in unbuilt land, and 8 in industrial. There were 292 tweets left that did not reach the cut-off point established for the classifier, so they were not labeled (non-classified) assuming that the text does not refer to a location.

3.4.2. Results Visualization

The tweets were also presented on the cadastre map of the historic city center. In this way, it is possible to compare the land uses identified by the classifier against the land use category registered in the cadastre by the local municipality.

The map of the historic center is shown in Figure 5 (on the left) with the classification of land uses according to the city's cadastre. On the map, each category of land use was associated with a color. Thus, the use of residential, commercial, industrial-office, institutional-governmental, and unbuilt land have the colors yellow, red, blue, light blue, and green, respectively. The second map in Figure 5 (on the right) shows the tweets labeled with the categories of land use. According to the information shown on the map, a large number of land uses corresponding to commercial, industrial, unbuilt land, and institutional-governmental were identified. However, a small number of tweets with the residential label is observed, although, according to the cadastre, a significant percentage of land uses in the historic center corresponds to residential. Therefore, there is no correspondence on the residential category between the cadastre and the automatic classification of land use.
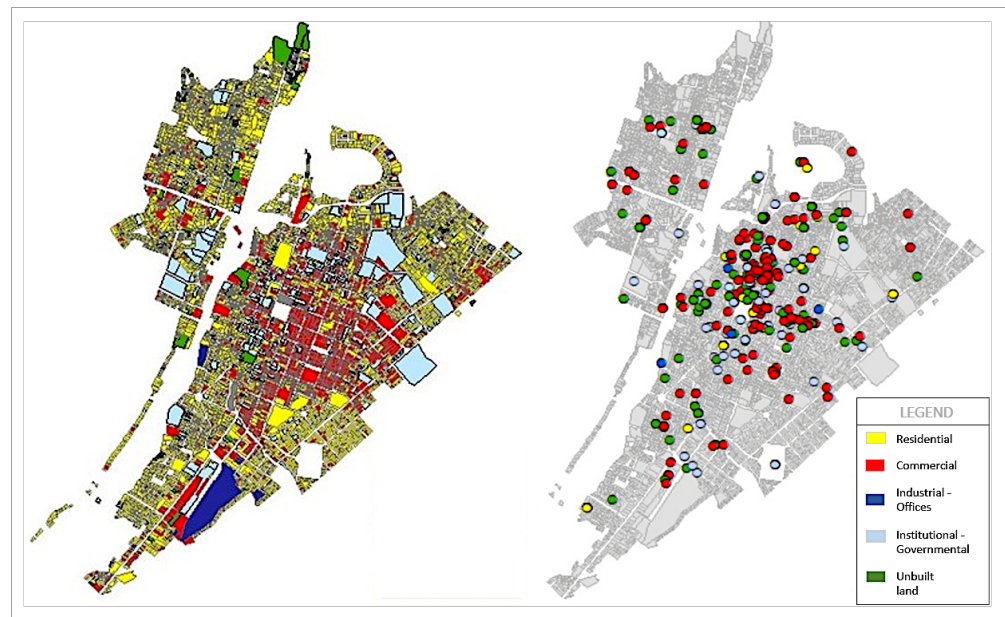
Figure 5. The image on the left shows the land use map of the historical center (cadastre); the image on the right shows the tweets labeled on the map by our approach.

In order to deepen the analysis of the inconsistencies revealed between the two methods, Figures 6 and 7 clearly show interesting differences. Color scale for land use is the same for both, as depicted in the legend. Figure 6 illustrates areas (light blue circle) labeled as residential in the cadastre, but identified as vacant land or institutional. On the other hand, Figure 7 shows a block (yellow) in which is located the Santa Catalina Convent, and it is cataloged as residential in the cadastre. However, tweets indicate the potential use as unbuilt or institutional-government land (religious is a subcategory within the last one). Multiple cases of inconsistency were detected when comparing cadastre and classifier outcome, many of which were later verified by a site visit. In general, the results provided by our approach were much more accurate than those registered in the cadastre.

Figure 6. Example of inconsistency between the categorization of the cadastre and the result obtained by the classifier.

Figure 7. The highlighted area corresponds to the convent of Santa Catalina (in the cadastre it is wrongly classified as residential land use).

4. Conclusions

Social networks provide valuable data on urban dynamics, offering new opportunities for research in the field. In this study, we used Twitter data to analyze land use in the historical center of the city of Arequipa by capturing tweets from the area with the text of the publication, time, date, and coordinates.

This research proposes a complete methodology of NLP for the analysis of tweet texts and coordinates, together with the Naïve Bayes Multinomial algorithm for the classification of spaces within land use categories. The evaluation of the model shows that the approach provides excellent results, with accuracy of about 90%, and F1-score of about 88%. The information of the area obtained from the project "Height for Culture" is used as a basis for the interpretation of the results, verifying that, as expected due to the knowledge of the area, a large percentage of the properties belong to the commercial category because these are in the historical city center with high presence of tourism, followed by the buildings of the institutional-cultural category due to the historical character of the area. However, the methodology detected that many residential spaces registered in the cadastre have currently other activities or uses.

We conclude that the Twitter data provides useful information to identify land uses in the geographic area where it is captured. Tweets are gathered in a simple and inexpensive way and provide information that can be used as an additional method by urban planning professionals and organizations interested in that area. The advantage of this model over the traditional ones resides in its dynamism since it uses data that is constantly updated by the users and allows reflecting the inconsistencies that exist in the maps generated by the cadastre due to the constant change of the environment. Moreover, the methodology is easily transferred to other geographical areas by means of the use of specific dictionaries to the region under study. This knowledge might be used as a recommendation system for short-term supervision or updating of the cadastre.

Finally, this method depends on the amount of geo-located data available at the time of classification, so the capture of publications from other social networks should be considered for future work to maximize the effectiveness and usefulness of the results.

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

References

1. Liu, X.; Kang, C.; Gong, L.; Liu, Y. Incorporating spatial interaction patterns in classifying and understanding urban land use. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 334–350. [CrossRef]
2. Yeh, A.G. Urban planning and gis. *Geogr. Inf. Syst.* **1999**, *2*, 1.
3. Anugraha, A.S.; Chu, H.-J. Land use classification from combined use of remote sensing and social sensing data. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2018**, *XLII-4*, 33–39. [CrossRef]
4. Ye, Y.; An, Y.; Chen, B.; Wang, J.; Zhong, Y. Land use classification from social media data and satellite imagery. *J. Supercomput.* **2020**, *76*, 777–792. [CrossRef]
5. Sendra, J.B.; García, R.C. El uso de los sistemas de Información Geográfica en la planificación territorial. *An. Geogr. Univ. Complut.* **2000**, *20*, 49.
6. Lin, Y.; Geertman, S. Can social media play a role in urban planning? A literature review. In *Computational Urban Planning and Management for Smart Cities, Lecture Notes in Geoinformation and Cartography*; Springer: Cham, Switzerland, 2019; pp. 69–84.
7. Lei, C.; Zhang, A.; Qi, Q.; Su, H.; Wang, J. Spatial-Temporal Analysis of Human Dynamics on Urban Land Use Patterns Using Social Media Data by Gender. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 358. [CrossRef]
8. Thakur, G.; Sims, K.; Mao, H.; Piburn, J.; Sparks, K.; Urban, M.; Stewart, R.; Weber, E.; Bhaduri, B. Utilizing geo-located sensors and social media for studying population dynamics and land classification. In *Human Dynamics Research in Smart and Connected Communities*; Springer: Cham, Switzerland, 2018; pp. 13–40.
9. Mora, H.; Pérez-delHoyo, R.; Paredes-Pérez, J.; Mollá-Sirvent, R. Analysis of Social Networking Service Data for Smart Urban Planning. *Sustainability* **2018**, *10*, 4732. [CrossRef]
10. Tessore, J.P.; Esnaola, L.M.; Russo, C.C.; Baldassarri, S. Comparative analysis of preprocessing tasks over social media texts in Spanish. In Proceedings of the XX International Conference on Human Computer Interaction, Gipuzkoa, Spain, 25–28 June 2019; pp. 1–8.
11. Stock, K. Mining location from social media: A systematic review. *Comput. Environ. Urban Syst.* **2018**, *71*, 209–240. [CrossRef]
12. Iglesias, J.A.; García-Cuerva, A.; Ledezma, A.; Sanchis, A. Social network analysis: Evolving twitter mining. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics, Budapest, Hungary, 9–12 October 2016; pp. 1809–1814.
13. Hajna, S.; Dasgupta, K.; Joseph, L.; Ross, N.A. A call for caution and transparency in the calculation of land use mix: Measurement bias in the estimation of associations between land use mix and physical activity. *Health Place* **2014**, *29*, 79–83. [CrossRef]
14. Terroso-Saenz, F.; Muñoz, A. Land use discovery based on volunteer geographic information classification. *Expert Syst. Appl.* **2020**, *140*, 112892. [CrossRef]
15. Da Silva, C. Usos del suelo: Distribución, análisis y clasificación con sistemas de información geográfica (SIG). *Rev. Digit. Grupo Estud. Sobre Geogr. Análisis Espac. Con Sist. Inf. Geográfica (GESIG)* **2013**, *5*, 142–152.
16. Silva, T.; Viana, A.; Benevenuto, F.; Villas, L.; Salles, J.; Loureiro, A.; Quercia, D. Urban computing leveraging location-based social network data: A survey. *ACM Comput. Surv.* **2018**, *52*, 1–39. [CrossRef]
17. Frias-Martinez, V.; Soto, V.; Hohwald, H.; Frias-Martinez, E. Characterizing urban landscapes using geolocated tweets. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; pp. 239–248.
18. Zhan, X.; Ukkusuri, S.V.; Zhu, F. Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. *Netw. Spat. Econ.* **2014**, *14*, 647–667. [CrossRef]
19. García-Palomares, J.C.; Salas-Olmedo, M.H.; Moya-Gómez, B.; Condeço-Melhorado, A.; Gutiérrez, J. City dynamics through twitter: Relationships between land use and spatiotemporal demographics. *Cities* **2018**, *72*, 310–319. [CrossRef]
20. Ajao, O.; Hong, J.; Liu, W. A survey of location inference techniques on twitter. *J. Inf. Sci.* **2015**, *41*, 855–864. [CrossRef]
21. Tandel, S.; Jamadar, A.; Dudugu, S. A survey on text mining techniques. In Proceedings of the 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS), Coimbatore, India, 15–16 March 2019; pp. 1022–1026.
22. Wabula, Y.; Nuzir, F.; Dewancker, B. Dynamic land-use map based on twitter data. *Sustainability* **2017**, *9*, 2158.
23. Ragini, J.R.; Anand, P.R.; Bhaskar, V. Big data analytics for disaster response and recovery through sentiment analysis. *Int. J. Inf. Manag.* **2018**, *42*, 13–24. [CrossRef]
24. Kateb, F.; Kalita, J. Classifying short text in social media: Twitter as case study. *Int. J. Comput. Appl.* **2015**, *111*, 1–12. [CrossRef]
25. Kulkarni, A.; Shivananda, A. *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning Using Python*; Apress: New York, NY, USA, 2019.
26. Lansley, G.; Longley, P. The geography of twitter topics in london. Computers, *Environ. Urban Syst.* **2016**, *58*, 85–96. [CrossRef]
27. Tellez, E.; Miranda-Jiménez, S.; Graff, M.; Moctezuma, D.; Siordia, O.S.; Villaseñor García, E. A case study of spanish text transformations for twitter sentiment analysis. *Expert Syst. Appl.* **2017**, *81*, 457–471. [CrossRef]
28. Varma, R.; Ahmad, S. Mass violence detection using data mining techniques. *World Sci. News* **2018**, *113*, 218–225.

29. Maheswari, M. Text mining: Survey on techniques and applications. *Int. J. Sci. Res.* **2017**, *6*, 1660–1664.

30. Thangaraj, M.; Sivakami, M. Text classification techniques: A literature review. *Interdisc. J. Inf. Knowl. Manag.* **2018**, *13*, 117–135. [CrossRef]

31. Mercado, S.K.A. Instrumento de Financiamiento Urbano para la Conservación del Patrimonio Arquitectónico de la Ciudad de Arequipa. Master's Thesis, Universidad Nacional de San Agustín, Arequipa, Peru, 2018.

32. Information Resources Management Association (Ed.). *Emergency and Disaster Management: Concepts, Methodologies, Tools, and Applications*; IGI Global: Pennsylvania, PA, USA, 2019.

33. Campan, A.; Atnafu, T.; Truta, T.; Nolan, J. Is data collection through twitter streaming api useful for academic research? In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 3638–3643.

34. Salas-Zárate, M.P.; Paredes-Valverde, M.A.; Ángel Rodriguez-García, M.; Valencia-García, R.; Alor-Hernández, G. Automatic detection of satire in twitter: A psycholinguistic–based approach. *Knowl.-Based Syst.* **2017**, *128*, 20–33. [CrossRef]

35. Asriadie, M.; Mubarok, M.; Adiwijaya, K. Classifying emotion in twitter using bayesian network. *J. Phys. Conf. Ser.* **2018**, *971*, 12041. [CrossRef]

36. Padró, L.; Stanilovsky, E. Freeling 3.0: Towards wider multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, 23–25 May 2012; pp. 2473–2479.

37. Anzovino, M.; Fersini, E.; Rosso, P. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems. Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2018; pp. 57–64.

38. Sakaki, T.; Matsuo, Y.; Yanagihara, T.; Chandrasiri, N.P.; Nawa, K. Real–time event extraction for driving information from social sensors. In Proceedings of the 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), Bangkok, Thailand, 27–31 May 2012; pp. 221–226.

39. Koto, F.; Adriani, M. The use of PoS sequence for analyzing sentence pattern in twitter sentiment analysis. In Proceedings of the 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, Gwangiu, Korea, 24–27 March 2015; pp. 547–551.

40. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]