electronics

MDPI

Article

COVIDSensing: Social Sensing Strategy for the Management of the COVID-19 Crisis

Alicia Sepúlveda [1], Carlos Periñán-Pascual [1], Andrés Muñoz [2], Raquel Martínez-España [3], Enrique Hernández-Orallo [4,*] and José M. Cecilia [4]

1 Department of Applied Linguistics, Universitat Politècnica de València, 46022 Valencia, Spain; alsemuo@idm.upv.es (A.S.); jopepas3@upvnet.upv.es (C.P.-P.)
2 Department of Computer Science, Universidad de Cádiz, 11003 Cádiz, Spain; andres.munoz@uca.es
3 Department of Information and Communication Engineering, University of Murcia, 30100 Murcia, Spain; raquel.m.e@um.es
4 Department of Computer Engineering (DISCA), Universitat Politècnica de València, 46022 Valencia, Spain; jmcecilia@disca.upv.es
* Correspondence: ehernandez@disca.upv.es

Abstract: The management of the COVID-19 pandemic has been shown to be critical for reducing its dramatic effects. Social sensing can analyse user-contributed data posted daily in social-media services, where participants are seen as Social Sensors. Individually, social sensors may provide noisy information. However, collectively, such opinion holders constitute a large critical mass dispersed everywhere and with an immediate capacity for information transfer. The main goal of this article is to present a novel methodological tool based on social sensing, called COVIDSensing. In particular, this application serves to provide actionable information in real time for the management of the socio-economic and health crisis caused by COVID-19. This tool dynamically identifies socio-economic problems of general interest through the analysis of people's opinions on social networks. Moreover, it tracks and predicts the evolution of the COVID-19 pandemic based on epidemiological figures together with the social perceptions towards the disease. This article presents the case study of Spain to illustrate the tool.

Keywords: social sensing; COVID-19; Natural Language Processing; Machine Learning; data analysis

Citation: Sepúlveda, A.; Periñán-Pascual, C.; Muñoz, A.; Martínez-España, R.; Hernández-Orallo, E.; Cecilia, J.M. COVIDSensing: Social Sensing Strategy for the Management of the COVID-19 Crisis. *Electronics* **2021**, *10*, 3157. https://doi.org/10.3390/electronics10243157

Academic Editor: Arturo de la Escalera Hueso

Received: 29 November 2021
Accepted: 15 December 2021
Published: 18 December 2021

1. Introduction

Social-streaming services, such as Twitter, enable citizens to express their opinions about events around the world in real time. In this way, citizens become social sensors [1]. The information extracted from social-media postings can be collected and analysed in real time thanks to the disciplines of big data and the Internet of Things. In this context, the COVID-19 (Coronavirus disease 2019) pandemic that is ravaging the world is a serious event about which we can express our opinions on social media. The analysis of such information can help make thoughtful decisions to improve the public health of citizens. This pandemic is the greatest global challenge in our recent history, which is putting the welfare state of today's societies at risk. Governments have taken drastic measures to reduce its spread, such as social distancing, contact tracing, perimeter closures, and even quarantines [2]. Recently, with the introduction of vaccines and new treatments, there seem to be high hopes of beating this pandemic. However, the socio-economic effects of these measures have a dramatic impact on all sectors of the population. The early discovery and understanding of social concerns can enable authorities to take actions to prevent possible scenarios that could increase the toll of the COVID-19 pandemic.

Social media has always been a useful source of information for the management of natural disasters and crises, such as earthquakes [3], floods [4], pandemic Zika and Ebola [5,6] or terrorist attacks [7,8], to name a few. Traditional hard sensors only offer quantitative information that is not valid in many scenarios where human-interactions are still mandatory.

Additional sources of information may provide valid knowledge to manage the COVID-19 crisis [9]. A new paradigm, called social sensing, emerged in the last decade to address a variety of problems by leveraging information from online social media. Indeed, social sensing is considered to be the new information age. This paradigm encompasses the set of elements that enable data sensing, where data are collected from humans or devices on their behalf [10]. Social sensing has resulted in new research lines aimed at analysing and interpreting the tremendous amount of information published daily in social-media tools [1]. These tools, which provide ubiquitous real-time information, have become a global phenomenon, where users post content to report facts or show situations of interest. Thus, social sensing considers these users as "social sensors" (a.k.a., soft sensors), that is, people that provide on-the-spot information about a particular fact through online social media such as social networks. Individually, social sensors generate "noise", as citizens may misinterpret a certain situation through their subjective perspective of reality. However, collectively, such opinion holders constitute a large critical mass dispersed everywhere and with an immediate capacity for information transfer.

In this article, we introduce COVIDSensing (https://covidsensing.com, last accessed 17 February 2021), a novel methodological tool based on social sensing that provides actionable information to deal with the crisis caused by COVID-19. This application constructs a systemic understanding of social welfare during the COVID-19 pandemic by combining different sources of information, including the real-time data obtained from different social networks and media sources. This information is processed using Natural Language Processing (NLP) techniques to identify real-time socio-economic issues. Moreover, clinical and epidemiological variables are also considered to predict the evolution of the COVID-19 pandemic through Machine Learning (ML) methods. We validate *COVIDSensing* for the particular case of Spain, which is one of the most affected countries in the world in terms of both socio-economic and health impacts. The results obtained by *COVIDSensing* show that the application serves to identify some critical issues that coincide with the timing of the main events of the COVID-19 pandemic.

The tool proposes a new methodology that goes beyond COVID-19 incidence data and news related to deaths. Indeed, this methodology has been devised to provide a comprehensive view of any problem that may be directly or indirectly related to the COVID-19 pandemic. The tool not only allows performing a global analysis of any socio-economic or health problem but can also be used for early warnings, thus supporting short-term decisions to address the given problem. Moreover, unlike other tools, it is intended not only to predict the evolution of COVID-19 cases and deaths but also to detect COVID-19-related problems long before they develop into serious problems, whose solutions are more costly both humanly and economically. Therefore, the main contributions of this study are as follows:

- Developing an AI-driven analytics framework capable of collecting and analysing any relevant social, economic or health information related to the COVID-19 pandemic; in this way, COVIDSensing can be regarded as an early warning system for detecting COVID-19 pandemic issues;
- Analysing the information according to its typology through novel NLP techniques;
- Integrating ML techniques to make the evolution of the events or situations detected correlate with the evolution of the pandemic;
- Designing the integration and combination of analyses from various perspectives, for example, figuring out how social inequalities unfold during the pandemic;
- Offering a simple and intuitive tool for public use, capable of being a source of reliable information for citizens;
- Implementing an application that can warn of the risks and dangers related to COVID-19, so that citizens and emergency responders can increase situational awareness.

The remainder of this article is organised as follows. In Section 2, some related work on social sensing and COVID-19 applications is discussed. In Section 3, the main building blocks of the COVIDSensing tool are briefly introduced. In Section 4, we provide a detailed

account of the Spanish socio-economic context as a case study to be analysed. Finally, Section 5 graphically shows the main results achieved by COVIDSensing for the early identification of social-risk situations in the transmission of SARS-COV-2.

2. Background

Throughout the COVID-19 pandemic, new lines of research have emerged with the aim of analysing the predictions and relationships of the SARS-CoV-2 virus around the world. These emergent research lines have been based on adapting new living conditions to the pandemic, for example, by trying to improve processes in the industry [11], monitoring patients' health [12,13], adapting work and workplaces [14,15], and so forth. Many researchers have focused on developing mathematical and ML models to predict the evolution of the pandemic, the rise and fall of detected cases as well as mortality [16–19] using official data and conducting individual city or country studies.

AI and ML techniques have been extensively used for fighting against COVID-19 in many different areas. For example, the authors of [20] showed that the distance learning applied in the COVID-19 confinement changed not only teaching strategies but also students' strategies when learning autonomously. In [21], the authors applied ML to analyse the spread of COVID-19 cases in the USA, obtaining highly accurate models focused on two different regimes, that is, lockdown and reopen, modelling each regime separately.

A different approach was proposed by [22] to evaluate the reliability of the factors that influenced the condition of COVID-19 patients. Their methodology is based on the calculation of the structure function as the classification task, using the approach of the classifier induction for building the structure function. In [23], the authors presented a screening support system that can accurately identify the patients that do not carry the disease, based also on the generation of classifiers.

Despite this increasing number of studies, very few of them have proposed the use of the social-sensing paradigm to analyse and/or evaluate the coronavirus pandemic. In [24], the authors proposed the implementation of an interactive visualisation through an application of human mobility using a low-cost passive Wi-Fi community tracking infrastructure. The network deployed proved to be appropriate to collect relevant information related to people's mobility during the COVID-19 outbreak and to show changes in mobility patterns in the Madeira Islands (Portugal), where the authors conducted their particular case study. In [25], a social-sensing detection methodology based on images of social indicators was proposed. The methodology uses information from social networks that is processed with ML techniques and crowdsourcing to discover and validate observations of social behaviour. Then, such observations are aggregated to estimate statistical indicators of social behaviour that are useful for policymakers. The methodology is grounded on a semi-automated social-detection pipeline that combines automatic image classification techniques with crowd-based validation techniques, enabling the reliable estimation of social indicators. A sentiment-analysis model of the COVID-19 pandemic using social sensing was presented in [26]. This model has two stages. In the first stage, the unsupervised Bidirectional Encoder Representations from Transformers model (BERT) is used to classify categories of sentiment. In the second stage, a term frequency-inverse document frequency model is used to determine trends, summarise posts, and analyse topics to identify the features of negative sentiment. The model was applied to a case study in China using posts from the social network Weibo. A review of different risk warning systems based on social sensing was presented in [27], which collected and studied social data to infer the evolution of the spread of COVID-19. This review described different models and stand-alone applications that had been used for other similar problems and focused on the way they could be applied to COVID-19. To the best of our knowledge, there is no application similar to *COVIDSensing*, which tracks and alerts about socio-economic issues related to the COVID-19 pandemic in real time from a social-sensing approach.

3. The *COVIDSensing* Tool

COVIDSensing is a tool that serves to provide a global perception of the evolution of the different socio-economic and health problems that are affecting our society due to the COVID-19 pandemic.

Figure 1 shows the main stages of the methodology of COVIDSensing. The first stage consists in defining the problem to be addressed. Then, the data-acquisition stage crawls information from different data sources, mainly from social networks and official sources. In the next stage, data are processed in several steps. On the one hand, data are geolocated. On the other hand, the NLP module analyses and categorises the micro-texts obtained from social networks. Moreover, the ML module analyses the official data about the COVID-19 pandemic to predict the evolution of different variables, such as the number of active cases or deaths. Finally, the outcomes are visualised, and conclusions can be drawn from them, for which we can even combine different problems by either joining or intersecting the data.

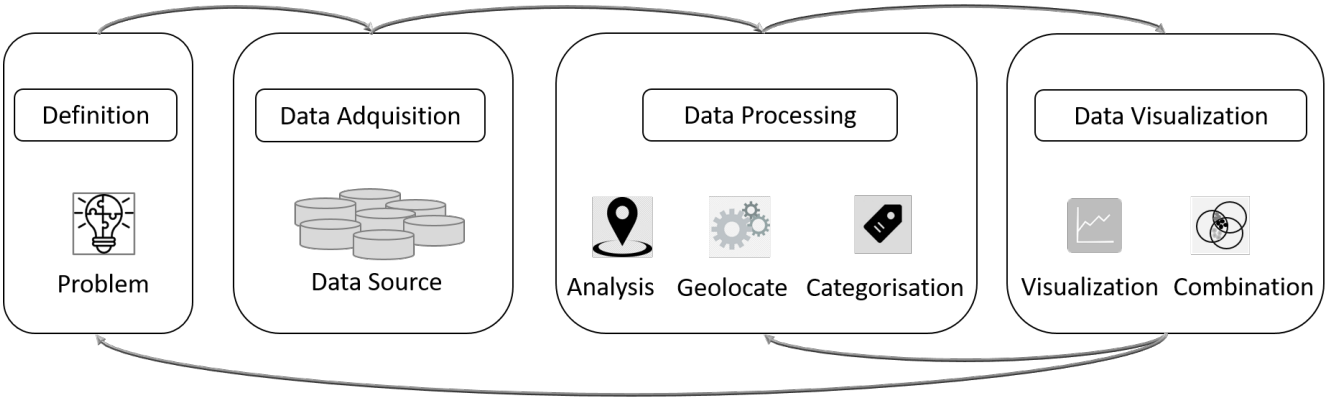

Figure 1. Description of the phases of the intrinsic methodology of the COVIDSensing tool.

Figure 2 shows the internal implementation details of COVIDSensing. The execution flow starts on the left-hand side of this figure, at the front end of the application. Based on Google's Firebase, it communicates with the web API, where different micro-services are available. The Apache Kafka cluster, which is a distributed streaming platform not only to publish and subscribe to streams of records but also to store and process these streams in real time, distributes the information received from social-media sources. In this regard, there are three main streams of records in our application: the first two are generated by the Twitter and Telegram crawlers, and the latter by the RSS crawler. The data derived from these crawlers are sent to the NLP module that is explained below. Finally, the information is persisted in an Elastic Search database.

It should be noted that, in the last phase, the tool allows you to go either to the first phase to define a new problem or to the analysis phase to re-model the data in combination with other problems. This will be discussed further below.

3.1. Definition

The definition of the problem begins by describing semantic spaces that help us in the search for information. COVIDSensing was designed to identify a range of problems related to the pandemic domain in general, and to COVID-19 in particular. Therefore, the first task is to define several pandemic-related topic categories (e.g., QUARANTINE and VACCINE, among many others) in the form of semantic spaces, which must be constructed before text processing. Constructing the semantic space of a given topic category involves a three-step process. First, the user decides seed terms that are characteristic of the semantic space, typically the name of the category itself, for example, *cuarentena* [quarantine]. Second, the user is automatically presented with the meanings associated to each seed term and selects those meanings that are semantically related to the topic category. Third,

the system lexically expands the seed terms based on the choice of meanings and suggests a set of additional topical words and phrases semantically related to these seed terms, where named names (e.g., people, places or organisations) are not considered. Therefore, every topic category is associated with a set of topical words, where every topical word is represented as an object with attributes such as the part of speech (POS), that is, noun, verb or adjective, and the sentiment (i.e., 1, −1 or 0 for positive, negative or neutral, respectively).

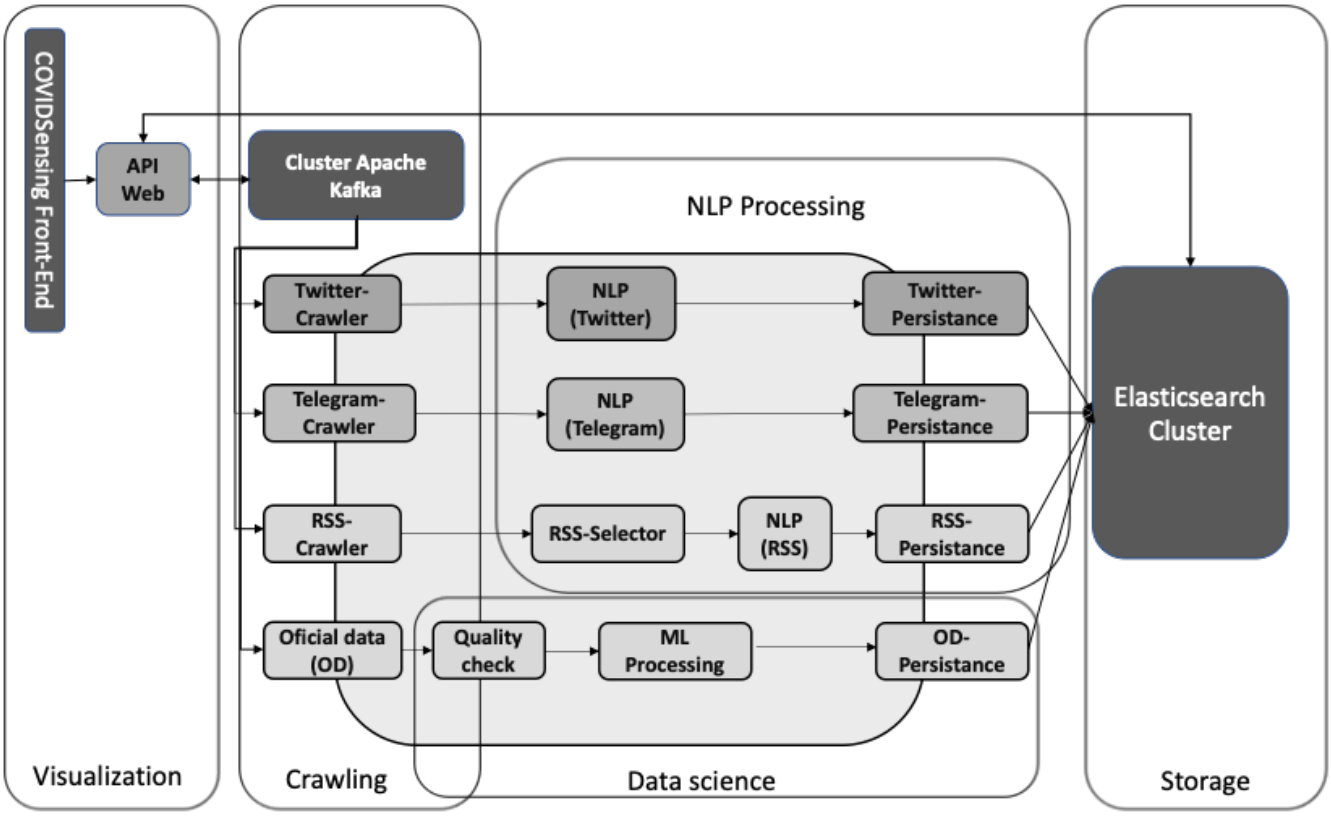

Figure 2. *COVIDSensing* software architecture.

3.2. Data Acquisition

After defining the semantic spaces, the tool starts the data-acquisition process. Many governments rely on public health surveillance for COVID-19 based on main principles provided by the World Health Organization (WHO) [28]. These systems are based on clinical and epidemiological criteria such as the definition of confirmed, suspected and probable cases. This information is usually provided by governments on a daily basis. Each government sets different strategies for communicating and using this information [29]. In the case of Spain, the information is provided by regional governments, reporting daily (except on weekends and holidays) to the Spanish Ministry of Health, which ultimately develops a report about the current situation of the pandemic in Spain [30]. It is important to note that the information is updated backwards when new notifications arrive from previous days, mainly due to delays, error detection, and so forth. COVIDSensing obtains clinical and epidemiological information from this official source, updating historical information when necessary.

Public health surveillance systems miss a large number of positive cases due to multiple reasons, for example, the existence of asymptomatic cases and limited access to diagnostic tests. Spain has developed up to four serological surveys to assess the extent of the epidemic, reported on July and December, respectively. This population-based study, developed in Spain and other countries, aims to determine the seroprevalence of SARS-CoV-2 infection at the national level [31]. The last serological study estimated

that approximately 4.7 million people have been affected by COVID-19 (approximately 9.9% of the Spanish population), against the 1,762,212 cases officially confirmed on that date. The identification of undetected cases and even the early detection of potentially risky social attitudes that could increase the number of infections would shed light on the magnitude and characteristics of the outbreak and reduce subsequent transmission [32]. In this regard, NLP and ML techniques applied to open data from news, social networks or web searches have been extensively used to provide epidemiological insights, being integrated into the formal surveillance strategies [33,34]. Some of them even claimed to have detected the first cases of the disease for COVID-19, before the WHO released a statement on the outbreak [35]. Indeed, there are different ways to create a direct communication channel with the population. Crowdsourcing is one of these alternatives, where the population can send their opinions off-line. This approach, however, requires a public awareness campaign through a citizen science project that gives you enough critical mass for decision-making. Another alternative is to access online social networks, where people freely and continuously express their opinions. COVIDSensing relies on the latter approach because the COVID-19 pandemic is a popular topic nowadays, so there is a large number of people expressing their opinions in these channels. Therefore, COVIDSensing crawls a collection of micro-texts from different Spanish social networks, including Twitter, Telegram, and Really Simple Syndication (RSS) feeds.

3.3. Data Processing

Once the data have been collected, they need to be processed to geolocate where the information comes from and to analyse the information to gain insights into the pandemic situation. The sub-processes involved in this stage are described below.

3.3.1. Geolocation and Categorisation

The geolocation and data-categorisation stages are made up of various tasks to identify the origin of the message and to categorise the content of the message using NLP techniques. The categorisation of information involves not only topic classification and sentiment analysis but also metrics that aim to make a first approximation to the detection of fake news. Such categorisation is grounded on a knowledge-based approach. In this regard, apart from the Spanish WordNet [36], from which we leveraged knowledge about semantic relations occurring between words, *COVIDSensing* employs several other resources that were constructed in [37], for example, the polarity lexicon and the lexicon of valence shifters. On the one hand, the polarity lexicon includes positively- and negatively-marked words in terms of sentiment analysis. On the other hand, the lexicon of valence shifters includes words and phrases that can neutralize the values of the topic and sentiment attributes (e.g., *no*, *sin* [without]), or increase or decrease the value of the sentiment attribute (e.g., *bastante* [enough] or *poco* [little], respectively).

The first task of this stage is processing natural language. In this task, *COVIDSensing* proceeds as follows. Each micro-text is divided into different phrases, which are in turn divided into tokens. Each token is represented as an object with different attributes, for example, the lexeme, the POS, the position in the micro-text, the topic, and the sentiment, where the default values of these last two attributes are zero. After obtaining the tokens, the second task is finding out words related to the topic. In this task, *COVIDSensing* identifies significant words and phrases in the micro-text that are related to each topic category. In particular, the value 1 is set to the topic attribute of every word or phrase in the micro-text that is found in the inventory of topical words and phrases for a given topic category, which were semi-automatically defined in the first task. In this case, the POS is also taken into consideration. Then, it proceeds with finding out words related to sentiment. This task consists in detecting significant words and phrases in the micro-text with respect to the sentiment. In particular, the values 1 or -1 can be assigned to the sentiment attribute of every word or phrase in the micro-text according to our polarity lexicon or, in the case that the topic attribute is 1, to the sentiment attribute of the corresponding topical

word. The next task consists in handling valence-shifters. This task is aimed at modifying the values of the topic and sentiment attributes of the words according to our lexicon of valence shifters. In this regard, determining the direction and length of the scope of the valence shifters is a critical factor. Finally, *COVIDSensing* applies a technique for detecting problems. For this task, we devised a metric to determine if a problem about a given topic category is likely to happen. This can also be used to set alerts in the system (e.g., minor, moderate or critical) and as an initial process of detecting fake news. A problem-relatedness perception index (PPI) defined in [37] is used for this purpose. The PPI calculation is based on three main steps. On the one hand, cosine similarity is used to assess the degree of relatedness between topic categories and micro-texts. Therefore, a micro-text is related to a given topic category if the similarity score is greater than zero. Regarding the sentiment score, on the other hand, it is worth mentioning that we are only concerned with problem detection. Therefore, only negative words and phrases are taken into consideration, which are assigned a negative value by the sentiment-relatedness function. Positive words and phrases are set to zero. Finally, the PPI is computed as the geometric mean of the topic and sentiment relatedness scores, so that a proportional compromise between topic categorisation and sentiment analysis can be reached.

Therefore, any micro-text is classified as a problem only if the PPI is a positive value. Accordingly, alerts can be set based on the PPI metrics. Indeed, it is very likely that there is a potential problem when the system discovers a sufficiently large number of messages (N) whose PPI is very close to 1. This means that there is a high number of people reporting a certain situation or demanding a certain action. In this way, the administrator can automatically inform via email about any risk or situation of interest. This can be addressed automatically by setting a PPI threshold limit and a number of messages above which a potential problem about a given topic is reported. For example, if five messages about a given topic have a PPI of 0.75, the administrator could be alerted to the occurrence of a potential problem. Both the number of messages and the PPI threshold is configurable by the administrator.

Once the information has been geolocated and categorised, it is analysed in the next stage. It should be noted that, although this article focuses on the processing of micro-texts in Spanish, *COVIDSensing* has already been expanded to deal with English. Since the construction of the semantic spaces associated with topic categories is based on WordNet synsets, each one representing a set of synonymous words in a variety of languages, adapting this tool to other languages only requires a few further resources. In particular, *COVIDSensing* makes use of two types of language-dependent resources, that is, text-processing resources (e.g., POS tagger) and lexical resources (e.g., polarity lexicon, lexicon of valence shifters). In this regard, since the former are readily available for many European languages, the main effort should only be placed on the latter.

3.3.2. Data Analysis

Focusing only on the clinical and epidemiological variables, most of the predicting models in the COVID-19 scenario have failed to predict new outcomes or even short-term evolution. The problem is that models are just a simplification of a messy, complex reality to help us understand what might happen in a given situation. Two main sets of models to predict the evolution of COVID-19 have been proposed in the literature. On the one hand, epidemiological compartmental models distribute the population into different compartments; for example, simple SIR models have three compartments: Susceptible, Infectious, and Recovered. These models can predict the dynamics and evolution of an infectious disease. In these models, the prediction is severely affected by small differences in the fitting parameters, limiting the reliability of the forecast [38]. On the other hand, statistical-based models, for example, time series analysis and forecasting, only use data from the past to predict the immediate future. Among these models, we may highlight the AutoRegressive Integrated Moving Average (ARIMA) model [39] and the Long Short-Term Memory Network (LSTM) model [40,41].

On the one hand, the ARIMA model is derived from the ARMA model, which combines the autoregressive (AR) model and a moving average (MA) model. The ARMA model is only considered for a univariate stationary times series. If the series is not stationary, as in the COVID-19 series, we need to differentiate it to make it stationary (i.e., the integrated part of the ARIMA model). An ARIMA model has three main parameters (p, d, q). These parameters correspond to the order of the autoregressive model (p), the degree of differentiation (d), and the order of the moving average model (q).

On the other hand, the LSTM model, which is a type of recurrent neural network (RNN), solves the difficulties of RNNs when learning long-range dependencies. LSTMs have a chain-like structure, where the replay module has four layers that interact with each other, instead of having a single layer. LSTMs are deep-learning networks that perform satisfactorily in time series problems.

These two models use data from the above processes to predict the evolution of the pandemic. The tool calculates both models and displays the one with the smaller mean squared error based on the available data. The predictions model active cases, confirmed cases on a single day, deaths, and recovered patients. Currently, the models can predict the upward or downward trend of the pandemic. As shown in the case study, the system clearly visualises not only the waves that Spain has suffered but also how the fourth wave is beginning to rebound.

3.4. Data Visualisation

At the visualisation stage, the tool offers trend graphs that can be customised by date or by the default options of month, week, day or hour. Current active data come from Twitter, Telegram, and RSS. Figure 3 shows an example of a search on the problem of "Coronavirus in Spain" between the dates from 30 December 2020 and 6 January 2021. As can be seen, most of the messages came from the social network, while no information was obtained from Telegram.

The bottom graph in Figure 3 analyses the messages using sentiment analysis, ranging from very negative through neutral to very positive. In the legend of the sentiment types, we can activate or deactivate the different items, and the graph is dynamically updated. The messages influx graph represents the total number of tweets, RSS posts and Telegram messages by date. This date will depend on the previously selected filter, for example, hour, day, week, month, etc. The graph focuses on representing the differences in information in a qualitative way from various information sources. In Figure 4, a bar chart showing only positive or very positive sentiments is displayed. As can be seen by comparing Figures 3 and 4, positive sentiments with the term coronavirus are much lower than negative ones.

In addition, Figure 4 shows a map where the messages obtained were geolocated, as well as a graph displaying the most frequent words in the messages according to their size. This form of visualisation and use of the tool allows for flexibility and adaptability to any type of public, so that information can reach everyone from various perspectives.

One of the last features of data visualisation in COVIDSensing is the possibility for the tool to combine information, so that users can find out how a given problem can affect or be detected from other problems by joining or intersecting information. When the selected option is "union", all messages from the selected topics are displayed with their corresponding analysis. When the selected option is "intersection", an attempt is made to find out whether one topic is related to the other or whether they are independent topics.

Figure 3. Example of visualisation considering the dates from 30 December 2020 to 6 January 2021.

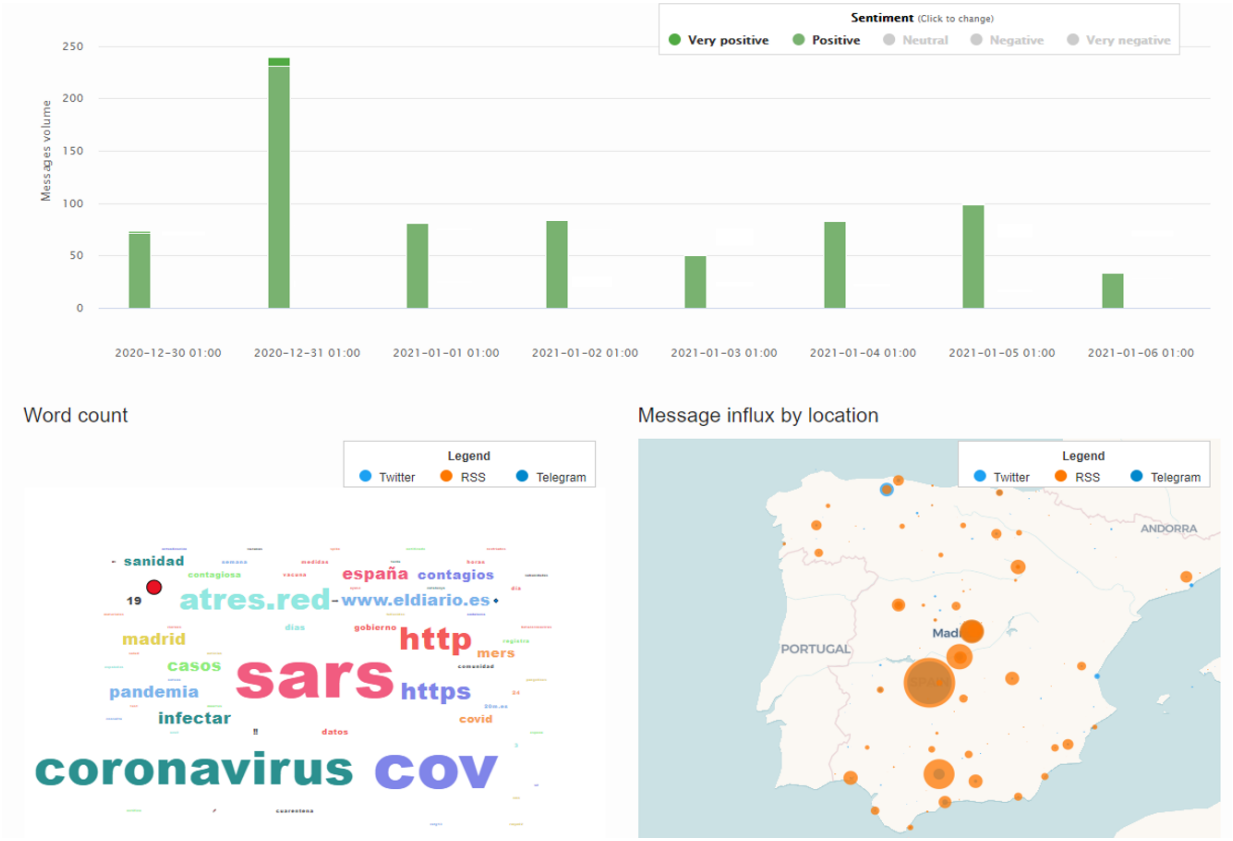

Figure 4. Example of geolocation visualisation considering the dates from 30 December 2020 to 6 January 2021.

In the example shown in Figure 5, we wanted to make an intersection between coronavirus in Spain, racial inequality, and education. In this regard, when users want to perform both the intersection and the merging of topics, the tool moves back from the visualisation stage to the data-processing stage to obtain the results from which we can derive useful information.

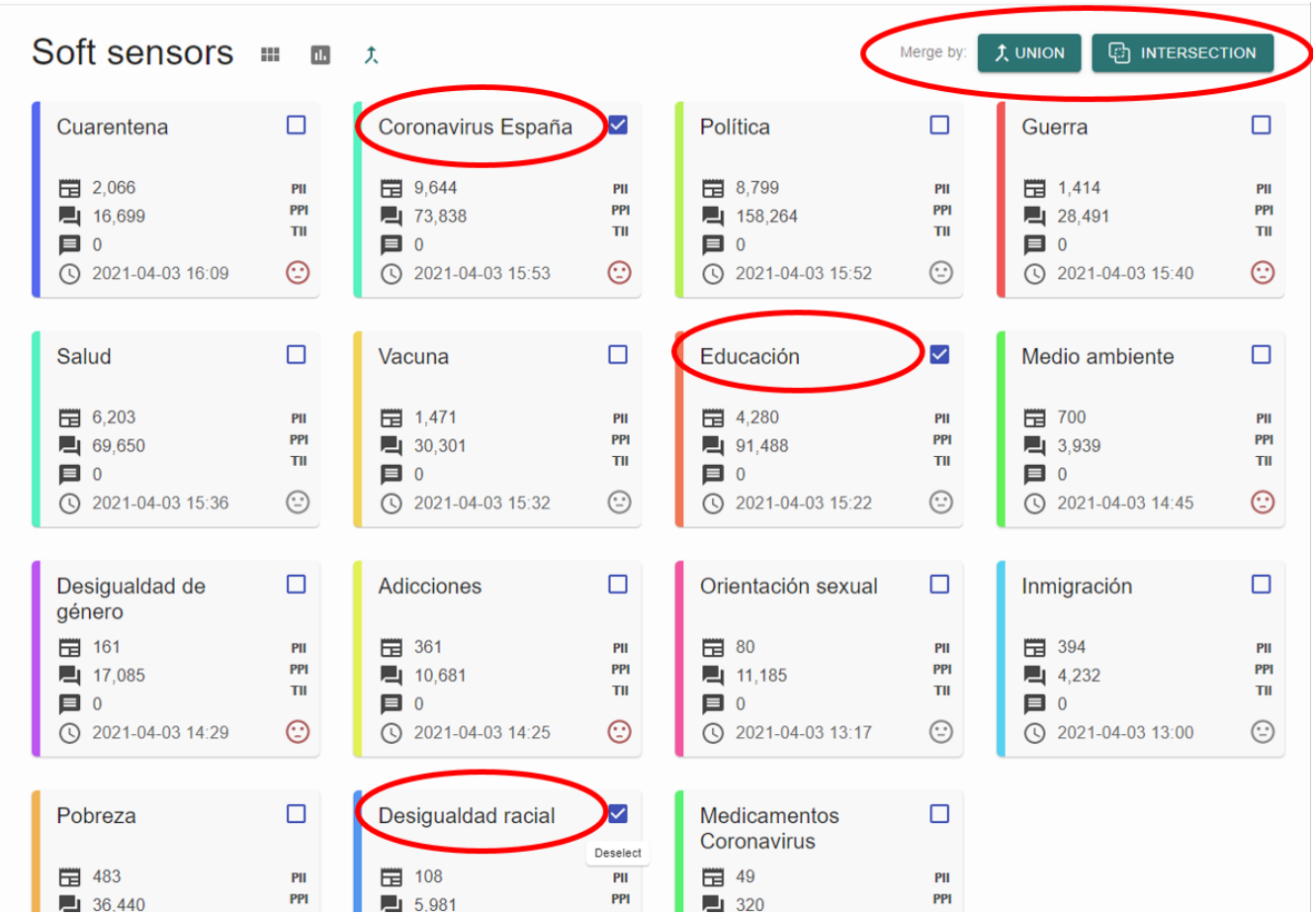

Figure 5. Combination of topics from *CovidSensing.com* to obtain information from different problems.

4. Case Study: The COVID-19 Pandemic in Spain

Spain has been one of the most affected countries by the COVID-19 pandemic. *The Lancet* published an editorial [42] highlighting the already fragile situation of the four main pillars of the Spanish health system (i.e., financing, delivery, governance, and workforce) after a decade of austerity. Despite the efforts made during the pandemic [43], at the time of writing this article (end of March 2021), Spain is the ninth country in terms of COVID-19 incidence worldwide, with 3,300,965 diagnosed cases, and the tenth according to the number of deaths, with up to 75,698 deaths. The number of cases increased significantly in the fall despite restrictions in the summer, such as nationwide shutdown of nightclubs and bars. Thus, the Spanish government declared a new state of emergency on 26 October 2020, for six months. This state of emergency was decentralised, that is, each regional government decided, in consensus with the central government, the measures to control the incidence of the pandemic. These measures were based on the early action plan, developed by the Spanish Ministry of Health, which specified metrics based on the evolution of the pandemic such as cumulative incidence, positive rate or hospital occupancy levels.

According to the latest reports from the Ministry of Health, the main cause of the spread of COVID-19 in Spain is social gatherings. Indeed, the cases of COVID-19 associated with family and social outbreaks (e.g., meetings with friends) have increased by 54% in the last seven days. As a matter of fact, the number of social events increases dramatically

during vacation periods. Since the first Spanish wave ended on 20 July 2020, Spanish society has had different vacation periods (see Figure 6). The longest one was the summer, particularly in August, when mobility was fully allowed to promote tourism and not to harm the economy. It is worth mentioning that the tourism sector represents 12.4% of Spanish GDP. This social relaxation resulted in a steep increase of 185 in the 14-day cumulative incidence (i.e., CI(14 days))—July 20's CI(14 days) was 27.387 and September 1's CI(14 days) was 212. This second wave started to gradually decrease on 23 September until 9 October, when the CI started to increase again. There were two vacation periods since 9 October at the national level, from 9–12 October and from 5–8 December. During the previous vacation period, there was a steep increase in the 14-day CI that ended with the declaration of a state of emergency on 26 October by the Spanish government. These measures managed to reduce the curve until 11 December when the drop in CI slowed down and began to grow again. The third wave appeared in mid-December and grew rapidly during late 2020 and early 2021. The relaxation of restrictions at Christmas resulted in a large number of infections and deaths compared to the second wave.

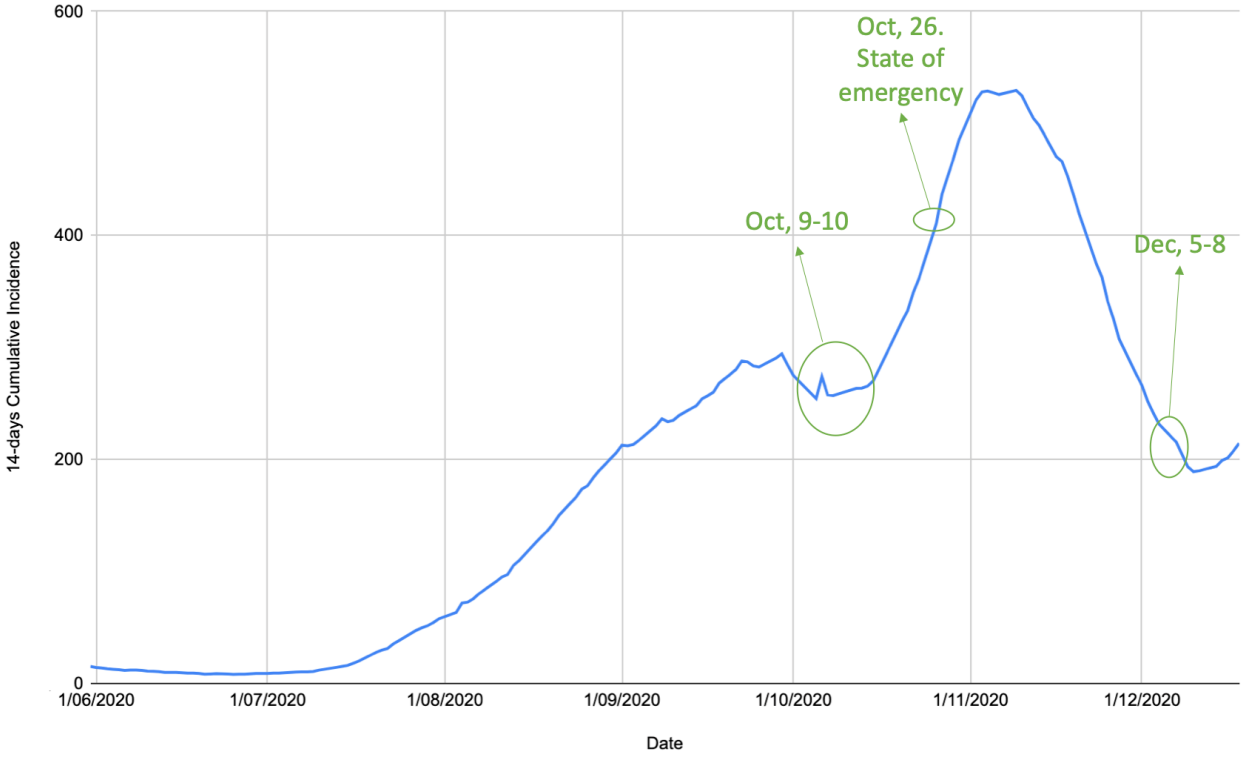

Figure 6. 14-day cumulative incidence (CI) in Spain. Holidays and the declaration of the emergency state are highlighted.

In this context, the last projections developed by the OECD (Organisation for Economic Co-operation and Development) reduced the growth of the Spanish economy in 2021 (5%) and 2022 (4%) after the steep decline in 2020. Therefore, the recovery phase will be slower than expected, and the level of GDP will remain below pre-crisis levels at the end of 2022 [44]. The following section analyses several particular cases in Spain, starting in October 2020, when the second wave began to increase. In addition, we discuss how pandemic-related problems can be identified and show how ML algorithms can predict changes in the spread of the infection.

5. Results and Discussion

Early action plans to deal with the evolution of the COVID-19 pandemic are based on the identification of positive (or confirmed) cases. However, these figures provide an ex-post view of what is happening in society. In fact, they are at least 7–10 days late, that

is, the incubation period of the COVID-19 disease. Figure 6 shows that it took seven days from the declaration of the state of emergency on 26 October until the curve began to bend on 3 November. Therefore, the main question of this research is whether the analysis of people's opinions on social networks can anticipate this scenario, identifying reckless behaviours that may lead to future infections. Eventually, the increase in the number of cases occurs because social relations are intensified and prevention measures are relaxed. Thus, the initial hypothesis with *COVIDSensing* is that this social relaxation is transferred to social networks. When people are aware of the loosening of social distancing and prevention measures, they tend to express their concerns and complaints on social media.

Figure 7 reflects an example of this situation, showing the number of positive messages on social media (i.e., Twitter, Telegram, and RSS) during October in the semantic space "Coronavirus in Spain". This semantic space is composed of the following words: *coronavirus*, *covid-19*, *sars*, *pandemia* [pandemic], *sars-cov-2*, *epidemia* [epidemic], *contagio* [contagion], *infección* [infection], *cuarentena* [quarentine], and *incubación* [incubation]. Indeed, these are the main keywords typed by users when posting messages about the COVID-19 pandemic. The number of posts about this semantic space decreased significantly only during the vacation period in October (10–12 October). After that period, a lower number of posts compared to the days before holidays was reported. Therefore, social interactions usually increase during holiday periods, reducing the number of social network messages about the coronavirus; when people have fun, they forget about the pandemic. Moreover, security measures against the coronavirus are relaxed (e.g., the use of masks, social distancing, etc.), which means that there is a post-vacation period where these measures continue to be relaxed. The announcement of the state of emergency on October 25 increased once again the number of messages about Coronavirus on social networks.

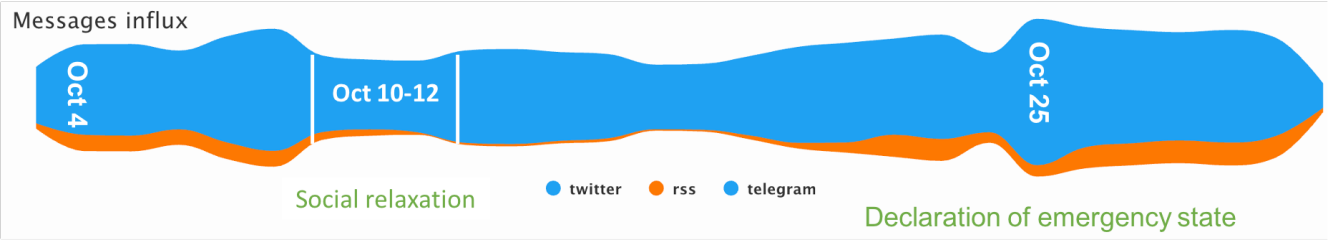

Figure 7. The number of messages on social networks (i.e., Twitter, RSS, and Telegram) about Coronavirus in Spain during the vacation period in October 2020.

Comparing Figures 6 and 7 with respect to the evolution of the pandemic, the cases of COVID-19 in Spain were decreasing at the beginning of October. This scenario changed suddenly from 12 October onwards, when again the number of people infected started to increase, even at a higher rate from October 16. Several days after the declaration of the state of emergency (5 November), Spain drastically reduced the 14-day cumulative incidence.

Similar conclusions can be drawn from the analysis of social media in December (see Figure 8). During the first vacation period in this month (5–8 December), social relaxation was identified. This pattern occurred again the next weekend (12–13 December). On 17 December, the region of Valencia reinforced the measures against the coronavirus already planned for Christmas, preventing family members and friends from entering that region to meet their relatives. This again had repercussions on social networks at a national level by increasing the number of posts about Coronavirus.

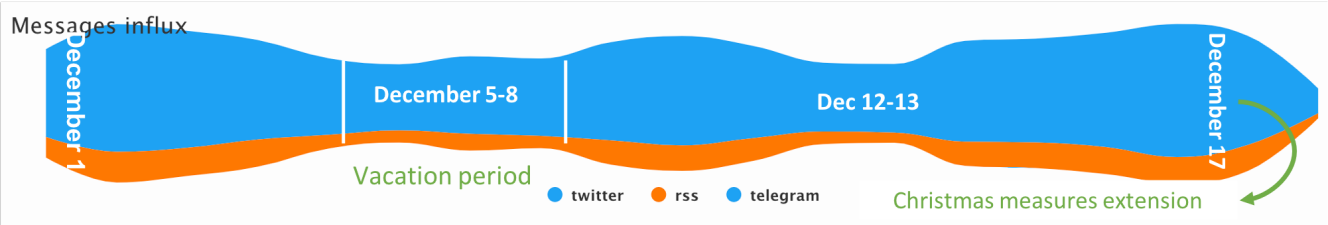

Figure 8. The number of messages on social networks (i.e., Twitter, RSS, and Telegram) about Coronavirus in Spain during the vacation period in December 2020.

In addition to the analysis performed on the evolution of the pandemic through social sensing, the methodology employed in this work also allows for the analysis of socio-economic problems. An example is shown in Figure 9, which reflects whether poverty and the coronavirus are problems that occur together, aggravating each other. This figure shows the evolution of the messages as well as the sentiment analysis of the messages in the Christmas period from 27 December 2020 to 6 January 2021. In this example, the hypothesis is that poverty is a cross-cutting problem for the coronavirus and it has a negative influence on its perception. Figure 9 shows the results of this intersection, indicating that it is a hot issue at Christmas. It is also observed that most of the information received is negative or very negative; this information increased on the days around the Epiphany (6 January), as the inequalities of gifts and the poverty of children are often commented on. This analysis of the data confirms our hypothesis.

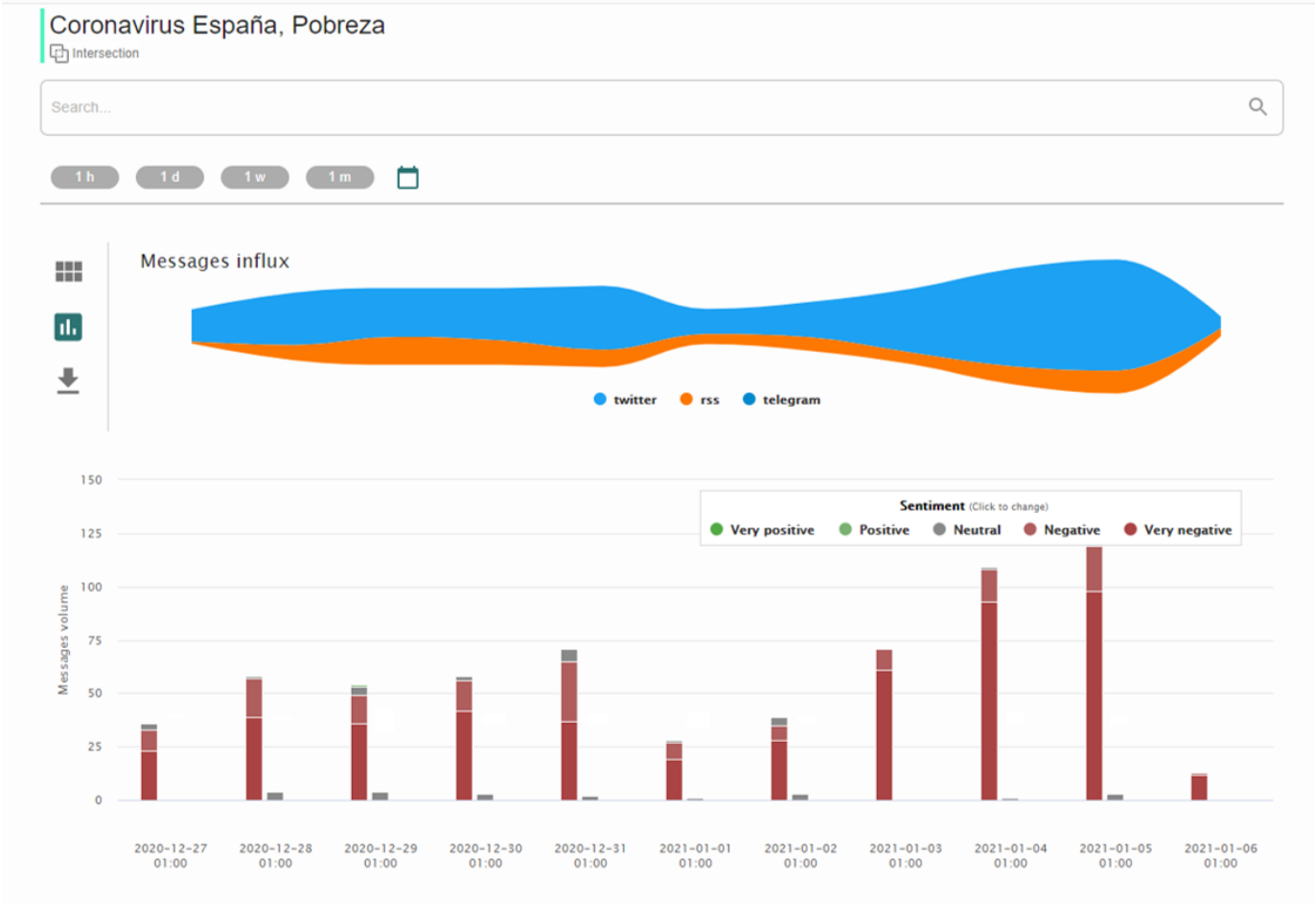

Figure 9. Combining problems through the intersection of Coronavirus and poverty with a focus on Spain.

To conclude this section of experiments, Figure 10 represents the predictions made by the ARIMA and LSTM techniques integrated into the tool and executed in the data-processing methodology phase. After running both techniques, the coefficient of determination (R^2) and root mean square error (RMSE) metrics are calculated, the predictions for the best values are jointly selected, and they are finally displayed to the user. We refer the reader to [45] for insights on the evaluation of these procedures. Fitting the ARIMA model is not straightforward, as the best combination of parameters has to be found to minimise the RMSE. From our analysis of the cases in Spain, we obtain an ARIMA(6,1,2) model. Nevertheless, these models can only predict short-time behaviour since the confidence interval grows extremely fast as time elapses [38]. For the LSTM model, the best parameters indicate a set of 50 input neurons and training between 1000 and 1500 epochs. The techniques obtain a fit between 80% and 90%, as indicated by the value of R^2.

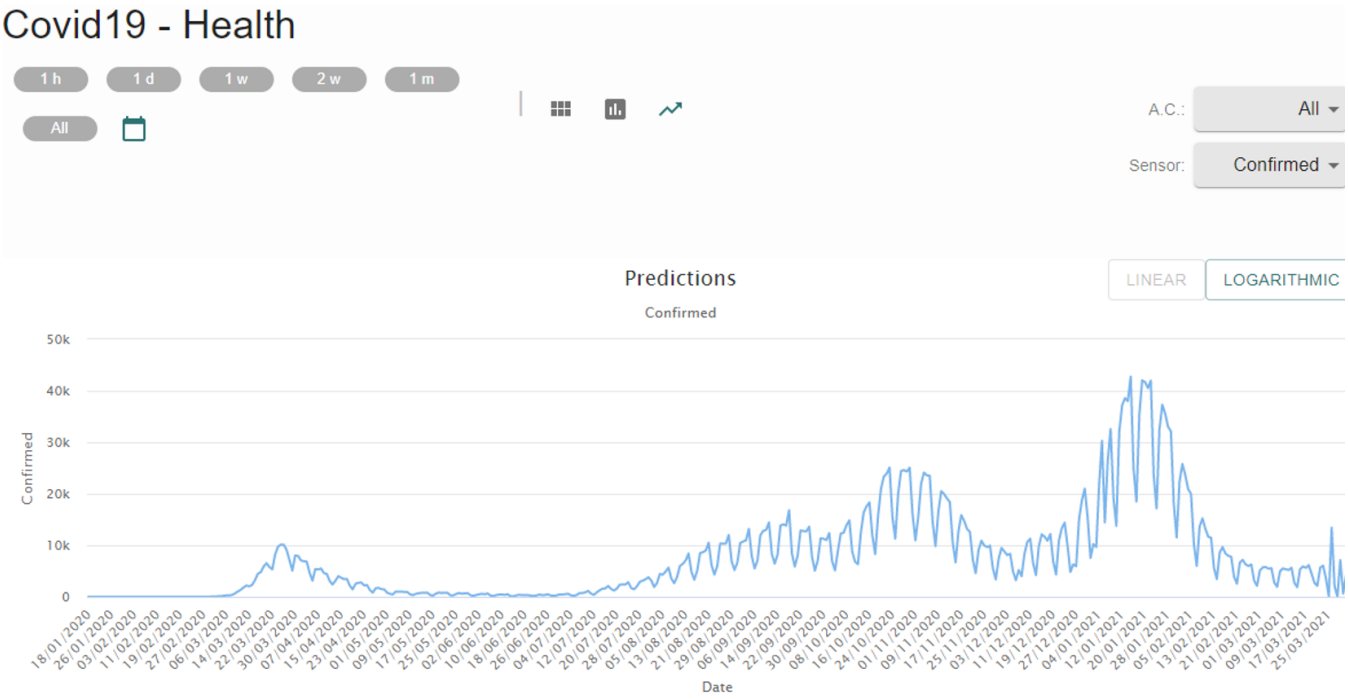

Figure 10. Prediction of confirmed cases in Spain since the beginning of the pandemic.

Figure 10 shows that confirmed cases stand out in the three waves that took place in Spain: the first wave from mid-March 2020 to May 2020, the second wave from August 2020 to the end of November 2020, and the third wave from the end of December 2020 to early March 2021. Focusing on the end of March, it can be seen that a new upswing is beginning, leading to the fourth wave that epidemiologists have already announced. Therefore, the proposed techniques are quite reliable in predicting trends in the increase of confirmed cases.

6. Conclusions and Future Work

This article presented COVIDSensing, a novel methodological application based on social detection for the socio-economic monitoring of the COVID-19 pandemic through people's opinions on social media. This application follows a working methodology that makes it extensible and adaptable to any problem related to the pandemic and allows the alerting of citizens and the helping of policymakers to make decisions. It is also a free tool available through its website to anyone who wants to be informed about current affairs and the problems related to the pandemic.

This article provided a detailed account of the stages involved in the methodology for developing COVIDSensing, from problem definition through data acquisition and

processing, to data visualisation. In this methodology, the data-processing stage proves crucial, where data are geolocated, categorised and analysed using ML models and NLP techniques. To evaluate our approach, we analysed several social networks in the two periods (or waves) in which a change in the spread trend was identified. Taking into consideration the incubation period of the disease, these periods coincided with two holiday periods in Spain (October and December), when a reduction in the number of social-media posts about the coronavirus in Spain was identified. Several days after both periods, the Spanish government took restrictive measures to reduce the infection curve, which also led to an increase in the number of posts on social media. Thus, it can be concluded that both the periods of COVID-19 spread risk and the measures issued by the Spanish government had an impact on people's opinions on social networks.

In addition, we evaluated the tool showing an example of a cross-cutting problem during the pandemic, that is, poverty, which was easily noticeable over the Christmas period. The intersection of poverty and COVID-19 demonstrated that both issues are closely connected—a situation that becomes further accentuated in the days leading up to the Epiphany.

Finally, we showed the predictions from the beginning of the pandemic until the end of March 2021 made by the ARIMA and LSTM models of the tool. In this manner, we could verify how the predictive models can correct the trend of unsuspected COVID-19 cases. Therefore, the results confirm that the perception of the problem on social media correlates perfectly with the evolution of the COVID-19 pandemic. This information can help government authorities and emergency managers in their decision-making processes to prevent major infection scenarios at an early stage.

Regarding the limitations of this work, it should be noted that the tool is based on opinions collected from different social-media sources, which cannot be considered a representative sample of the whole of Spanish society. However, as several authors argue (see for example [46]), data obtained from social media undoubtedly contribute to situation awareness. In this manner, our tool could be used to detect user-generated content that serves to raise situational awareness regarding COVID-19 related social problems. On the other hand, in this work we did not take into consideration any other factors apart from COVID-19 that could affect the perception of social problems. As future work, it would be possible to introduce new factors (e.g., political factors) into the tool to compare whether there are significant changes in the trends detected above.

Other future improvements involve the implementation of a more complex algorithm to detect fake news and the refinement of the predictions made by the models to achieve not only the prediction of trends but also greater accuracy when predicting confirmed cases.

Author Contributions: Conceptualization, A.S., J.M.C. and C.P.-P.; Formal analysis, A.S. and J.M.C.; Investigation, A.S., J.M.C., C.P.-P., A.M., R.M.-E., E.H.-O. and J.M.C.; Methodology, A.S., A.M. and R.M.-E.; Visualization, A.S., A.M., R.M.-E., J.M.C.; Project administration, J.M.C.; Resources, E.H.-O. and J.M.C.; Supervision, E.H.-O. and J.M.C.; Validation, A.S., C.P.-P., A.M., R.M.-E., E.H.-O.; Writing, All authors. All authors have read and agreed to the published version of the manuscript.

Funding: This work is derived from R&D project RTI2018-096384-B-I00, as well as the Ramon y Cajal Grant RYC2018-025580-I, funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe", by the Spanish "Agencia Estatal de Investigación" (grant number PID2020-112827GB-I00/ AEI/10.13039/501100011033), and by the "Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital, Proyectos AICO/2020", Spain, under Grant AICO/2020/302.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We also thank other members of the technical team; Julio Fernández, Pedro García, Juan Morales, José G. Giménez who have actively participated in the development of the tool.

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

References

1. Wang, D.; Szymanski, B.K.; Abdelzaher, T.; Ji, H.; Kaplan, L. The age of social sensing. *Computer* **2019**, *52*, 36–45.
2. Kissler, S.M.; Tedijanto, C.; Goldstein, E.; Grad, Y.H.; Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **2020**, *368*, 860–868.
3. Avvenuti, M.; Cresci, S.; La Polla, M.N.; Marchetti, A.; Tesconi, M. Earthquake emergency management by social sensing. In Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), Budapest, Hungary, 24–28 March 2014; pp. 587–592.
4. Arthur, R.; Boulton, C.A.; Shotton, H.; Williams, H.T. Social sensing of floods in the UK. *PLoS ONE* **2018**, *13*, e0189327.
5. Sharma, M.; Yadav, K.; Yadav, N.; Ferdinand, K.C. Zika virus pandemic—Analysis of Facebook as a social media health information platform. *Am. J. Infect. Control* **2017**, *45*, 301–302.
6. Fung, I.C.H.; Duke, C.H.; Finch, K.C.; Snook, K.R.; Tseng, P.L.; Hernandez, A.C.; Gambhir, M.; Fu, K.W.; Tse, Z.T.H. Ebola virus disease and social media: A systematic review. *Am. J. Infect. Control* **2016**, *44*, 1660–1671.
7. Zhao, X.; Zhan, M.M. Appealing to the Heart: How Social Media Communication Characteristics Affect Users' Liking Behavior During the Manchester Terrorist Attack. *Int. J. Commun.* **2019**, *13*, 22.
8. Bérubé, M.; Tang, T.U.; Fortin, F.; Ozalp, S.; Williams, M.L.; Burnap, P. Social media forensics applied to assessment of post–critical incident social reaction: The case of the 2017 Manchester Arena terrorist attack. *Forensic Sci. Int.* **2020**, *313*, 110364.
9. Liggins, M., II; Hall, D.; Llinas, J. *Handbook of Multisensor Data Fusion: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2017.
10. Wang, D.; Abdelzaher, T.; Kaplan, L. (Eds.) *The Future of Wireless Networks*; Morgan Kaufmann: Boston, MA, USA, 2015.
11. Udgata, S.K.; Suryadevara, N.K. COVID-19, Sensors, and Internet of Medical Things (IoMT). In *Internet of Things and Sensor Network for COVID-19*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 39–53.
12. Singh, R.P.; Javaid, M.; Haleem, A.; Suman, R. Internet of things (IoT) applications to fight against COVID-19 pandemic. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2020**, *14*, 521–524.
13. Otoom, M.; Otoum, N.; Alzubaidi, M.A.; Etoom, Y.; Banihani, R. An IoT-based framework for early identification and monitoring of COVID-19 cases. *Biomed. Signal Process. Control* **2020**, *62*, 102149.
14. Collins, C.; Landivar, L.C.; Ruppanner, L.; Scarborough, W.J. COVID-19 and the gender gap in work hours. *Gender Work Organ.* **2021**, *28*, 101–112.
15. Kaiser, M.S.; Mahmud, M.; Noor, M.B.T.; Zenia, N.Z.; Al Mamun, S.; Mahmud, K.A.; Azad, S.; Aradhya, V.M.; Stephan, P.; Stephan, T.; et al. iWorkSafe: Towards healthy workplaces during COVID-19 with an intelligent pHealth App for industrial settings. *IEEE Access* **2021**, *9*, 13814–13828.
16. Yan, L.; Zhang, H.T.; Goncalves, J.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jing, L.; Zhang, M.; et al. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2020**, *2*, 283–288.
17. Arora, P.; Kumar, H.; Panigrahi, B.K. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos Solitons Fractals* **2020**, *139*, 110017.
18. Bhardwaj, R. A Predictive Model for the Evolution of COVID-19. *Trans. Indian Natl. Acad. Eng.* **2020**, *5*, 133–140.
19. Guo, A.; Zhang, Q.; Zhao, X. A Graph-Based Approach Towards Risk Alerting for COVID-19 Spread. In *Software Foundations for Data Interoperability and Large Scale Graph Data Analytics*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 63–69.
20. Subirats, L.; Fort, S.; Atrio, S.; Sacha, G.M. Artificial Intelligence to Counterweight the Effect of COVID-19 on Learning in a Sustainable Environment. *Appl. Sci.* **2021**, *11*, 9923. https://doi.org/10.3390/app11219923.
21. Kamis, A.; Ding, Y.; Qu, Z.; Zhang, C. Machine Learning Models of COVID-19 Cases in the United States: A Study of Initial Lockdown and Reopen Regimes. *Appl. Sci.* **2021**, *11*, 11227. https://doi.org/10.3390/app112311227.
22. Levashenko, V.; Rabcan, J.; Zaitseva, E. Reliability Evaluation of the Factors That Influenced COVID-19 Patients' Condition. *Appl. Sci.* **2021**, *11*, 2589. https://doi.org/10.3390/app11062589.
23. Henzel, J.; Tobiasz, J.; Kozielski, M.; Bach, M.; Foszner, P.; Gruca, A.; Kania, M.; Mika, J.; Papiez, A.; Werner, A.; et al. Screening Support System Based on Patient Survey Data—Case Study on Classification of Initial, Locally Collected COVID-19 Data. *Appl. Sci.* **2021**, *11*, 10790. https://doi.org/10.3390/app112210790.
24. Ribeiro, M.; Nisi, V.; Prandi, C.; Nunes, N. A data visualization interactive exploration of human mobility data during the COVID-19 outbreak: A case study. In Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 7–10 July 2020; pp. 1–6.
25. Negri, V.; Scuratti, D.; Agresti, S.; Rooein, D.; Shankar, A.R.; Marquez, J.L.F.; Carman, M.J.; Pernici, B. Image-based Social Sensing: Combining AI and the Crowd to Mine Policy-Adherence Indicators from Twitter. *arXiv* **2020**, arXiv:2010.03021.
26. Wang, T.; Lu, K.; Chow, K.P.; Zhu, Q. COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model. *IEEE Access* **2020**, *8*, 138162–138169.
27. Rashid, M.T.; Wang, D. CovidSens: A vision on reliable social sensing for COVID-19. *Artif. Intell. Rev.* **2021**, *54*, 1–25.
28. WHO. *Critical Preparedness, Readiness and Response Actions for COVID-19: Interim Guidance, 4 Nov 2020*; Technical Report; World Health Organization: Geneva, Switzerland, 2020.
29. Han, E.; Tan, M.M.J.; Turk, E.; Sridhar, D.; Leung, G.M.; Shibuya, K.; Asgari, N.; Oh, J.; García-Basteiro, A.L.; Hanefeld, J.; et al. Lessons learnt from easing COVID-19 restrictions: An analysis of countries and regions in Asia Pacific and Europe. *Lancet* **2020**, *396*, 1525–1534.

30. Cecilia, J.M.; Cano, J.C.; Hernández-Orallo, E.; Calafate, C.T.; Manzoni, P. Mobile crowdsensing approaches to address the COVID-19 pandemic in Spain. *IET Smart Cities* **2020**, *2*, 58–63.
31. Pollán, M.; Pérez-Gómez, B.; Pastor-Barriuso, R.; Oteo, J.; Hernán, M.A.; Pérez-Olmeda, M.; Sanmartín, J.L.; Fernández-García, A.; Cruz, I.; de Larrea, N.F.; et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): A nationwide, population-based seroepidemiological study. *Lancet* **2020**, *396*, 535–544.
32. Budd, J.; Miller, B.S.; Manning, E.M.; Lampos, V.; Zhuang, M.; Edelstein, M.; Rees, G.; Emery, V.C.; Stevens, M.M.; Keegan, N.; et al. Digital technologies in the public-health response to COVID-19. *Nat. Med.* **2020**, *26*, 1183–1192.
33. Edelstein, M.; Lee, L.M.; Herten-Crabb, A.; Heymann, D.L.; Harper, D.R. Strengthening global public health surveillance through data and benefit sharing. *Emerg. Infect. Dis.* **2018**, *24*, 1324.
34. WHO. *Epidemic Intelligence from Open Sources (EIOS)*; Technical Report; World Health Organization: Geneva, Switzerland, 2020.
35. McCall, B. COVID-19 and artificial intelligence: Protecting health-care workers and curbing the spread. *Lancet Digit. Health* **2020**, *2*, e166–e167.
36. Agirre, A.G.; Laparra, E.; Rigau, G.; Donostia, B.C. Multilingual central repository version 3.0: Upgrading a very large lexical knowledge base. In Proceedings of the GWC 2012 6th International Global Wordnet Conference, Matsue, Japan, 9–13 January 2012; pp. 118–125.
37. Periñán-Pascual, C.; Arcas-Túnez, F. Detecting environmentally-related problems on Twitter. *Biosyst. Eng.* **2019**, *177*, 31–48.
38. Castro, M.; Ares, S.; Cuesta, J.A.; Manrubia, S. The turning point and end of an expanding epidemic cannot be precisely forecast. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 26190–26196.
39. Hernandez-Matamoros, A.; Fujita, H.; Hayashi, T.; Perez-Meana, H. Forecasting of COVID19 per regions using ARIMA models and polynomial functions. *Appl. Soft Comput.* **2020**, *96*, 106610.
40. Shahid, F.; Zameer, A.; Muneeb, M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* **2020**, *140*, 110212.
41. Chimmula, V.K.R.; Zhang, L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals* **2020**, *135*, 109864.
42. Health, T.L.P. COVID-19 in Spain: A predictable storm? *Lancet Public Health* **2020**, *5*, 568.
43. Moros, M.J.S.; Monge, S.; Rodríguez, B.S.; San Miguel, L.G.; Soria, F.S. COVID-19 in Spain: View from the eye of the storm. *Lancet Public Health* **2020**, *6*, 10.
44. OECD *OECD Economic Outlook*; OECD Publishing: Paris, France, 2020; Volume 2020.
45. García-Cremades, S.; Morales-García, J.; Hernández-Sanjaime, R.; Martínez-España, R.; Bueno-Crespo, A.; Hernández-Orallo, E.; López-Espín, J.J.; Cecilia, J.M. Improving prediction of COVID-19 evolution by fusing epidemiological and mobility data. *Sci. Rep.* **2021**, *11*, 15173.
46. Doran, D.; Severin, K.; Gokhale, S.; Dagnino, A. Social media enabled human sensing for smart cities. *AI Commun.* **2016**, *29*, 57–75. https://doi.org/10.3233/AIC-150683.