Universidad Autónoma de San Luis Potosí
Facultad de Ingeniería
Centro de Investigación y Estudios de Posgrado

Social Media Data and Machine Learning to Predict Acute Respiratory Infections

Para obtener el grado de:
Doctorado en Ciencias de la Computación

Presenta:
José Manuel Ramos Varela

Asesor:
Dr. Juan Carlos Cuevas Tello
Co-Asesor:
Dr. Daniel E. Noyola Cherpitel

San Luis Potosí, S. L. P. Junio 2025

Universidad Autónoma de San Luis Potosí
Facultad de Ingeniería
Centro de Investigación y Estudios de Posgrado

Redes Sociales y Aprendizaje Automático para predecir Infecciones Respiratorias Agudas

Para obtener el grado de:
Doctorado en Ciencias de la Computación

Presenta:
José Manuel Ramos Varela

Asesor:
Dr. Juan Carlos Cuevas Tello
Co-Asesor:
Dr. Daniel E. Noyola Cherpitel

San Luis Potosí, S. L. P. Junio de 2025

UASLP
Universidad Autónoma de San Luis Potosí

FACULTAD DE INGENIERÍA

10 de abril de 2025

M.E.M. JOSÉ MANUEL RAMOS VARELA
P R E S E N T E.

En atención a su solicitud de Temario, presentada por los **Dres. Juan Carlos Cuevas Tello y Dr. Daniel Ernesto Noyola Cherpitel,** Asesor y Coasesor de la Tesis que desarrollará Usted, con el objeto de obtener el Grado de **Doctor en Ciencias de la Computación**, me es grato comunicarle que en la sesión del H. Consejo Técnico Consultivo celebrada el día 10 de abril del presente año, fue aprobado el Temario propuesto:

TEMARIO:

"Redes Sociales y Aprendizaje Automático para predecir Infecciones Respiratorias Agudas"

1. Introducción.
2. Estado del Arte: Revisión Sistemática.
3. Datos de Twitter e Infecciones Respiratorias Agudas.
4. Metodología Computacional Basada en Aprendizaje Automático.
5. Resultados.
6. Conclusiones.
   Referencias.

"MODOS ET CUNCTARUM RERUM MENSURAS AUDEBO"

A T E N T A M E N T E

DR. EMILIO JORGE GONZÁLEZ GALVÁN
DIRECTOR.

UNIVERSIDAD AUTONOMA
DE SAN LUIS POTOSI
FACULTAD DE INGENIERIA
DIRECCION

Copia. Archivo.
*etn.

www.uaslp.mx

Av. Manuel Nava 8
Zona Universitaria • CP 78290
San Luis Potosí, S.L.P.
tel. (444) 826 2330 al39
fax (444) 826 2336

UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ
FACULTAD DE INGENIERÍA
Área de Investigación y Estudios de Posgrado

D E C L A R A C I Ó N

El presente trabajo que lleva por título:

"Redes Sociales y Aprendizaje Automático para predecir Infecciones Respiratorias Agudas"

se realizó en el periodo enero de 2024 a junio de 2025 bajo la dirección del Dr. Juan Carlos Cuevas Tello como asesor y del Dr. Daniel Ernesto Noyola Cherpitel como co-asesor.

Originalidad

Por este medio aseguro que he realizado el trabajo reportado, y la escritura de este documento de tesis para fines académicos sin ayuda indebida de terceros y sin utilizar otros medios más que los indicados.

Las referencias e información tomadas directa o indirectamente de otras fuentes se han definido en el texto como tales y se ha dado el debido crédito a las mismas.

El autor exime a la UASLP de las opiniones vertidas en este trabajo escrito y asume la responsabilidad total del mismo.

Este trabajo no ha sido sometido como tesis o trabajo terminal a ninguna otra institución nacional o internacional en forma parcial o total, exceptuando el caso cuando existe un convenio específico de doble titulación celebrado entre ambas instituciones.

Se autoriza a la UASLP para que divulgue este documento para fines académicos.

El autor del trabajo escrito, José Manuel Ramos Varela

First, I want to thank God for giving me life and the desire to pursue my dreams, as well as the health and opportunities to achieve my goals. I want to thank my life partner, Nayely Rivero, for encouraging and motivating me to pursue this academic goal, for being by my side and pushing me to achieve our dreams. I want to thank my parents for their emotional support and motivation to pursue what I want with effort and dedication. I thank Dr. Cuevas and Dr. Noyola for their motivation to achieve this goal, their patience in teaching, and their guidance in research.

Abstract

We studied the relationship between tweets referencing Acute Respiratory Infections (ARI) or COVID-19 symptoms and confirmed cases of these diseases. Additionally, we propose a computational methodology for selecting and applying Machine Learning (ML) algorithms to predict public health indicators using social media data. To achieve this, a novel computational methodology was developed, integrating three distinct models to predict confirmed cases of ARI and COVID-19. The dataset contains tweets related to respiratory diseases, published between 2020 and 2022 in the state of San Luis Potosí, Mexico, obtained via the Twitter API (now X). The methodology is composed of three stages, and it involves tools such as Dataiku and Python with ML libraries. The first two stages focus on identifying the best-performing predictive models, while the third stage includes Natural Language Processing (NLP) algorithms for tweet selection. One of our key findings is that tweets contributed to improved predictions of ARI confirmed cases, but did not enhance COVID-19 time series predictions. The best-performing NLP approach was the combination of Word2Vec algorithm with the KMeans model for tweet selection. Furthermore, predictions for both time series improved by 3% in the second half of 2020 when tweets were included as a feature, where the best prediction algorithm was DeepAR.

Resumen

Estudiamos la relación entre los tweets que hacen referencia a Infecciones Respiratorias Agudas (IRA) o síntomas de COVID-19 y los casos confirmados de estas enfermedades. Además, proponemos una metodología computacional para seleccionar y aplicar algoritmos de Aprendizaje Automático (ML) para predecir indicadores de salud pública utilizando datos de redes sociales. Para lograr esto, se desarrolló una novedosa metodología computacional, que integra tres modelos distintos para predecir casos confirmados de IRA y COVID-19. El conjunto de datos contiene tweets relacionados con enfermedades respiratorias, publicados entre 2020 y 2022 en el estado de San Luis Potosí, México, obtenidos a través de la API de Twitter (ahora X). La metodología se compone de tres etapas e involucra herramientas como Dataiku y Python con librerías de ML. Las dos primeras etapas se centran en identificar los modelos predictivos de mejor rendimiento, mientras que la tercera etapa incluye algoritmos de Procesamiento de Lenguaje Natural (PLN) para la selección de tweets. Uno de nuestros hallazgos clave es que los tweets contribuyeron a mejorar las predicciones de casos confirmados de IRA, pero no mejoraron las predicciones de las series temporales de COVID-19. El enfoque de PLN con mejor rendimiento fue la combinación del algoritmo Word2Vec con el modelo KMeans para la selección de tweets. Además, las predicciones para ambas series temporales mejoraron un 3% en el segundo semestre de 2020 al incluir los tweets como una característica, donde el mejor algoritmo de predicción fue DeepAR.

Contents

Índice general

Chapter 1

Introduction

Information from social networks has become an important resource for assessing public health emegencies, offering a powerful tool for real-time disease monitoring and the prediction of communicable diseases [1]. The growing trend in the use of social networks to share personal situations from users' daily lives generates large amounts of data that describe the behavior of the population, demographic data, geolocation data, images, text and interactions between users [2]. If it were possible to identify users' posts related to the presence of a disease in a community prior to an official confirmation (reported by a hospital or laboratory) of an acute respiratory infection (ARI), the collective information of an area or region would help to identify a potential outbreak. Traditionally, disease monitoring relies on Indicator-Based Surveillance Systems, which collect and report information to government health agencies. However, these processes often involve slow data flow and adherence to protocols that delay the publication of critical information [3]. A second type of surveillance systems, known as Event-Based Surveillance Systems, is gaining prominence with the availability of new information sources like social networks. The World Health Organization (WHO) defines these systems as the organized and rapid collection of information from sources such as social media, news outlets, and public health networks about events that may pose risks to public health [4]. An example of the importance of the use of these Event-Based Surveillance Systems is the report generated on December 30, 2019, by the network-based tool ProMED (Programme for Monitoring Emerging Infectious Diseases), which identified posts on the Chinese social network Weibo regarding pneumonia cases in Wuhan, prior to the recognition of the COVID-19 virus as a potential pandemic [5]. Another example is the use of data from WeChat, a social network from China, to predict COVID-19 trends, through the "Pandemic Forecast and Warning WeChat Mini Program," which enabled early warnings of new outbreaks across 31 provinces in China [2].

During the pandemic, WeChat was also used to distribute surveys for gathering information on COVID-19-related symptoms and to generate epidemic time series based on the collected data [6].

Another concept that has gained attention in recent research is Digital Epidemiology, an emerging field that leverages big data—including information generated through social networks—and digital technologies to analyze disease-related patterns. Its primary goal is the early detection and monitoring of viral and other infectious diseases outbreaks. Fallatah and Adekola [7] suggest that integrating this approach into existing surveillance systems could enhance global health infrastructure and help save lives during viral epidemics.

Based on the aforementioned concepts, this research focuses on social media data; however, one of the significant challenges in utilizing data from these platforms is the limited access relative to the large amount of information that is generated on a daily basis. Charles et al. [8] carried out a literature review on the use of social network information in the context of public health monitoring, and the main social network was Twitter (now X) with 81% (from 33 studies reviewed). Gupta and Katarya [9] performed a similar review, observing that Twitter data was used in 64% of a total of 26 selected investigations. The reason for the frequent use of this social network to carry out research is precisely the availability and access to current and past publications. However, due to recent changes in its administration, the access policies for the publications of this social network have become more restrictive. In addition, use of data from social networks for surveillance presents some challenges. One of the main concern when working with such data to monitor public health is misinformation; Wang et al. [10] conclude that misinformation is more popular than accurate information, and misinformation frequently induces fear, anxiety, and mistrust in institutions. Therefore, the use of NLP to identify those publications related to the diseases that are of interest for monitoring is a quite complex task, which requires the application of a diverse set of procedures and algorithms. Giancotti et al. [11] highlight additional concerns when working with data from social networks, including the absence of an ethical framework, the lack of internationally recognized privacy protection policies, the risk of spreading conspiracy theories, and the potential for false-positive events due to geographic or cultural variations in language.

The use of data from social networks to predict respiratory diseases is not a new research topic, and the complexity of the algorithms varies depending on their architecture and the degree of analysis of the messages posted by platform users. Dai et al. [12] describes a classification of four types of analyses to identify the degree of assessment of the text obtained from social media posts and the complexity of the algorithms used: 1) Keywords-based Approaches, 2) Learning-

based Approaches, 3) Lexicon-Based Approaches, and 4) Word Embedding Based Approach. The architecture of the algorithms used in this research includes elements of types 1), 2), and 4), and the reason for using different approaches was to compare the results obtained. Previous investigations have described the selection of tweets through Keywords-based Approach. For example, words like “sore throat", “cough" and “fever" found in tweets, which are related to respiratory diseases have been used [13]. Subsequently, the quantity or frequency of tweets related to these words can be compared with the number of persons who present with certain illnesses, such as those reported in national respiratory disease statistics. Talvis et al. [13] calculated the linear correlation coefficient between selected Tweets and Google Flu Trends queries. Hirose and Wang [14] used multiple linear regression models to relate the dynamics of tweets with influenza keywords, and Influenza-Like Illness data [15]. One of the disadvantages of using keywords in the selection of tweets is that they may not be related to events associated directly with the person who is publishing a social media post; for example, if the selected tweets are those which contain the word “influenza", in some instances these could be part of a public health campaign, and not related to people who have had or have a respiratory illness. So, one of the main challenges is the selection of social media posts. In this research, different methods were compared to identify an increase in the accuracy of the predictions depending on the filters used to select the tweets and the algorithms used for this. The second type of algorithms described by Dai et al. [12], require labeled data for training. Then, a ML classifier is used to predict if a tweet is related or not to acute respiratory infections. Some of the most widely used ML techniques are Naïve Bayes, Random Forest, Decision Tree, k-Nearest Neighbors (KNN), XgBoost, Logistic Regression and Support Vector Machines (SVM). Similar approaches were found in the investigations carried out by Zuccon et al. [16], Santos and Matos [17], and Prieto et al. [18], using different tools such as WEKA[1] and scikit-learn toolkit[2] to compare different ML models, and the models with the highest classification accuracy were SVM and Naïve Bayes. Jiang et al. [19] analyzed help-seeking messages on the Chinese social network Weibo during different stages of the COVID-19 pandemic. They used Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to approach the task as a binary topic classification problem.

Of note, not all works that analyze social media data associated with respiratory infections aim to use this information for surveillance purposes. Agrawal et al. [20] used different ML models to classify people’s feelings and analyze the direction of polarity in tweets, with

[1]https://ml.cms.waikato.ac.nz/weka/index.html
[2]https://scikit-learn.org/

support vector machine model being the one that resulted in the best performance; Ueda et al. [21] identified the changes in feelings before and after the declaration of a state of emergency in Japan due to the COVID-19 pandemic, relating tweets to different emotions such as anger, sadness, surprise, disgust, etc; and Aldosery et al. [22] evaluated the reaction sentiment (positive, neutral and negative) towards the COVID-19 pandemic in the UK, with a sentiment analysis model that combines a recurrent neural network and an embedded topic model. Guo et al. [23] examined Twitter users' beliefs about vaccination mandates during the COVID-19 pandemic in the USA, with classification ML models; Kalanjati et al. [24] did similar research to classify sentiments and opinions regarding COVID-19 and COVID-19 vaccination on Indonesian-language. Hamedani et al. [25] analyzed Reddit posts during the COVID-19 pandemic to identify major discussion topics using the Latent Dirichlet Allocation (LDA) algorithm. They then classified the sentiment of each topic using the Syuzhet package for the R programming language.

As mentioned above, data labeling presents one of the main challenges for training ML models. Manual classification of tweets has been the most frequently used mechanism to build the training data set of the models reported in previous works, and this requires classification of tweets by declaring which of these are related to respiratory disease and which are not. In the dataset used by Kagashe et al. [26], if a tweet was related to influenza infection, it was classified as "relevant", otherwise it was classified as "irrelevant". Zuccon et al. [16] used a scale from 0 to 100 (0 being no flu, 100 being certain of a flu), in which the evaluators classified the tweets related to influenza-like illness (ILI). The mechanism used by Santos and Matos [17] to reduce errors consisted in a classification carried out by three evaluators, labeling messages according to the opinion of the majority.

The unsupervised ML Algorithms in the fourth type (Word Embedding Based Approach) of Dai et al. [12] do not require labeled data, which makes human intervention in classification minimal. There are different processes to achieve this objective; one of them is the conversion of sentences (tweets) to numerical vectors. Word2Vec is one of the most commonly used tools for this purpose, and contains two types of algorithms: continuous bag-of-words and continuous skip-gram [27]. Other researchers use this algorithm to discover relationships between words through context [28, 29]; the comparison can be made using measures such as the cosine similarity, and the results can be used for different purposes, such as finding a relationship between named entities in text corpora, or simply with keywords. In this study, we use the results of Word2Vec to train the Kmeans clustering model, with the aim of grouping tweets according to their context. Chen et al. [30] used the Kmeans model to identify the relationship

of viral publications (Twitter and Weibo) and 77 features related to different topics (violence, international, vaccination, misinformation, etc.).

Didi et al. [31] integrate sequences of ML algorithms for different purposes; they used algorithms such as FastText and Glove for feature extraction, classification models such as SVM, XgBoost and LR; to later make predictions with Prophet, LSTM and SVR models. In the literature analyzed, we found some articles that used deep learning models, Long Short Term Memory (LSTM) [32, 31, 23] and Convolutional Neural Networks (CNN) [33]. In recent years, these algorithms have been developed and are more accurate than traditional ML models, e.g. Naive Bayes, Support Vector Machine, Random Forest, etc. Part of our approach is to study deep learning algorithms for prediction: i) DeepAR is a forecasting method based on autoregressive recurrent neural networks, which learns a global model from historical data [34] and ii) Transformer is an architecture based on an attention mechanism, embeddings and feed forward neural networks [35]. And in the same way, we integrate a prior analysis and selection of tweets using Word Embedding techniques, such as Word2Vec [12]. Other models that we use for classifying and grouping tweets are Bidirectional Encoder Representations from Transformers (BERT) [36] and Kmeans [37] respectively. Although these models have been used previously [38, 30], the combination with deep learning prediction models in a pipeline is innovative in this research. Lande et al. [39] used BERT models in tweet classification that provided better results compared to LDA models. To et al. [40] used this same model to identify anti-vaccination tweets comparing with bidirectional long short-term memory networks with pre-trained GLoVe embeddings (Bi-LSTM) and more traditional ML models such as SMV; they obtained the best results with the BERT model. And Yang et al. [41] compared BERT model with SVM, classifying tweets to identify users who have self-reported chronic stress experiences, obtaining the best results with BERT. Chen et al. [42] utilized a Chinese pre-trained BERT model to extract 20 key topics from posts on the Chinese social network Weibo, analyzing public perception of active aging after the COVID-19 pandemic. They suggest that these topics can serve as a guideline for long-term government planning, helping integrate public concerns into policy development.

The processes for selecting which algorithms to work with can be complicated and confusing. So, we introduce a computational methodology for this task, which contains three phases. In each phase, tests and comparisons are carried out among models to facilitate their selection and development. We use ML platforms such as Dataiku and the programming language Python. The main objective of this research is to investigate the relationship between tweets and information reported by health authorities related to ARI and COVID-19. Thus, we compare

different ML prediction models using confirmed cases of ARI and COVID-19 as a target; and as a feature, tweets related to possible cases of these diseases (independent variable) [43].

For each phase of the proposed computational methodology, we follow a data science workflow [44], which contains a workflow consisting in three steps: data, analysis and communication. The Data step involves data collection, cleaning, storage and sharing; Analysis step is composed by training and forecasting models; and the last step corresponds to the communication of the results, that is the ability to share knowledge with interested parties (visualization).

1.1 Research Questions

1. What is the relationship between Twitter information and acute respiratory infection diseases?

2. Using tweets as the independent variable, how are the results in predictions of positive ARI and COVID-19 cases?

3. What is the appropriate process and combination of machine learning models to obtain the most accurate prediction of acute respiratory infectious disease outbreaks?

1.2 Objetives

1.2.1 Main objective

To develop a machine-learning-based model with social media data to predict acute respiratory infectious disease outbreaks.

1.2.2 Specific objectives

1. To analyze information available from Twitter to identify contents related to acute respiratory infectious diseases.

2. To analyze respiratory infectious disease outbreak patterns in Mexico using publicly available surveillance databases against Twitter information.

3. To develop a machine-learning-based model to predict acute respiratory infectious disease outbreaks.

1.3 Publications from this Thesis

- Ramos-Varela, J. M., Cuevas-Tello, J. C., & Noyola, D. E. (2025). A Machine Learning-Based Computational Methodology for Predicting Acute Respiratory Infections Using Social Media Data. Computation, 13(4), 86. https://doi.org/10.3390/computation1304008 ISSN: 2079-3197

- Ramos-Varela, J. M., Cuevas-Tello, J. C., & Noyola, D. E. (2025). Acute Respiratory Infections Prediction through Social Media Data and Machine Learning: A Systematic Review, BMC Medical Informatics and Decision Making, Springer, ISSN: 1472-6947, In review process.

1.4 Thesis organization

This research is organized as follows. First, we start in Chapter 2 with the literature review with the related work to this research. Then, following the data science workflow: Chapter 3 presents the data, the first step; Chapter 4 as part of the analysis step the proposed computational methodology is described; Chapter 5 is the visualization of results, which corresponds to the communication step. At the end comes the conclusions and further research.

Chapter 2

State of the Art: Systematic Review

2.1 PRISMA Statement for Reporting Systematic Reviews

The methodology used in the development of this systematic literature review was the PRISMA Statement for Reporting Systematic Reviews, developed by Moher et al. [45]. This methodology allows structuring a literature review in an orderly manner, and presenting the results under certain predefined criteria. It is based on research questions. The results from the broad search are not only statistically (quantitative). The detailed analysis of the selected papers (qualitative) is the main contribution of a systematic review, i.e. extracting the information from each publication: algorithms, metrics, results, etc. In summary, the PRISMA methodology is important for systematic reviews because it promotes standardization, transparency, comprehensive reporting, bias reduction, enhanced quality, ease of peer review, increased credibility, and supports evidence-based decision-making. It has ample acceptance in studies related to health issues [8, 46, 47]. The goal of this systematic review is to examine recent research that uses ML techniques as a method to identify and quantify messages posted on Twitter related to infectious respiratory diseases. One of the challenges posed in this review is the differences in use of social networks from one country to another, due to social and economic factors. Also, the period of time used to collect data varies between studies, and this has an influence on the precision of ML techniques, since the period of time identifies when an epidemic occurs. The key features analyzed in the Systematic Review are the following:

1. Analysis methodologies used in studies related to the use of social networks for the prediction of acute respiratory infections (ARI).

2. ML techniques used in investigations related to the extraction of information on Twitter.

3. Metrics used to measure the performance of the ML techniques.
4. Main countries related to the extraction of information on Twitter with ML techniques.
5. The time series periods analyzed in studies.

2.1.1 Search Strategy

The EBSCO Discovery Service (EDS) database for scientific literature was searched for records published between January 2006 and December 2020; the criterion to select the start date was the Twitter´s creation date (2006). Additionally, a filter for peer-reviewed publications was applied. Table 2.1 shows the search terms used, divided into three sections; each selected record had to contain at least one item from each section. The attributes provided by EDS platform for each article are the following:

1. Article Title
2. Author
3. Journal Title
4. ISSN
5. ISBN
6. Publication Date
7. Volume
8. Issue
9. First Page
10. Page Count
11. Accession Number
12. DOI

Table 2.1 Search terms applied on the EDS database, including title and abstract during the search.

Terms related to diseases	**Twitter related terms**	**Terms related to prediction models**
influenza OR coronavirus OR SARS OR epidemic OR healthcare OR disease OR outbreak OR surveillance OR epidemiology OR pandemic OR influenza-like illness OR pneumonia OR acute respiratory infection OR COVID OR public health OR surveillance systems OR flu OR ILI OR SARS-CoV-2 OR COVID-19	social media OR twitter OR tweets	forecasting OR machine learning OR predict OR deep learning OR predict OR predicting OR classification OR recognition OR artificial intelligence OR classifiers

The databases consulted by EDS are listed in Table 2.2.

2.1.2 Study Selection

For the study selection, first, the article's title was analyzed, and the following were excluded: i) duplicated articles, ii) those published in a language other than English, and iii) systematic reviews. Once the pertinent articles were identified, the abstract was analyzed and excluded those who are not related to the following subjects: i) human subjects, ii) social networks, and iii) respiratory diseases or prediction models. Additionally, we only considered peer-reviewed publications, such as conferences and journals; therefore, informal articles were excluded, i.e. technical reports, handouts, etc. Because the measurement of people's feelings does not correspond to the purpose of this research, any article related to this concept was also removed from the selection.

2.1.3 Data collection process

From each selected article, the information of the authors, the year of publication, the models used for prediction, metrics and results, as well as the period of time, the region and the respiratory disease analyzed in the study were extracted.

In order to analyze the first two features, we followed the division proposed by Dai et al. [12]. So, the methodologies that use social media data with ARI, particularly ILI, are divided into four categories [12]:

1. *Learning-based Approaches.* This methods require labeled data for training. Then, a ML classifier is used to predict if a tweet is related or not with ARI. Some of the most widely used ML techniques are Naive Bayes, KNN and SVM.

2. *Lexicon-Based Approaches.* It is a knowledge-based unsupervised learning method; does not need annotated labels for training; instead uses dictionaries knowledge to incorporate labels.

3. *Word Embedding Based Approach.* It is an unsupervised learning method and does not require annotated data; it learns the optimal vectors from surrounding words. A tweet can be represented as a few vectors and divided into clusters of similar words; and this vectors can represent the semantic information of words.

4. *Keywords-based Approaches.* This methodology correlates the number of respiratory infection-related keywords identified with the number of real cases registered.

2.2 Results found in the literature review

The results are divided into Study Selection and Study Characteristics. The Study Selection describes clearly the process of how the articles were selected in this systematic review, and Study Characteristics provides the answer for each research question formulated in §2.1.

2.2.1 Study Selection

The search generated 2,032 records; the distribution by content provider can be seen in the Table 2.2. The duplicate records were 1034; only the titles and abstracts of the remaining 998 articles were reviewed, this includes an analysis of each article where the objective is to

corroborate that the research corresponded to: *i)* the social network Twitter, *ii)* respiratory diseases, and *iii)* ML techniques or prediction models. Additionally, during this analysis it was verified that the article was not related to the tweet's sentiment analysis (or emotions). So, 883 articles that did not meet these criteria were eliminated, as well as 9 written in a language other than English and 12 articles classified as systematic literature reviews. After this verification, 94 articles were selected for full text review to corroborate if they contained the three necessary characteristics: *i)* the social network Twitter, *ii)* respiratory diseases, and *iii)* ML techniques or prediction models. After this process, it was determined that only 46 articles met the eligibility criteria. Figure 2.1 summarizes the whole process using the PRISMA flow diagram.

Table 2.2 Total articles ($n = 2,032$) identified in the search and classified by content provider.

Content Provider	**Number of Records**
MEDLINE	527
Complementary Index	395
Academic Search Ultimate	282
Directory of Open Access Journals	176
IEEE Xplore Digital Library	163
Supplemental Index	125
ScienceDirect	100
Library, Information Science & Technology Abstracts with Full Text	98
Springer Nature Journals	80
Business Source Complete	55
Emerald Insight	8
ERIC	6
JSTOR Journals	5
GreenFILE	4
Dentistry & Oral Sciences Source	4
SciELO	2
SciTech Connect	1
Academic Source	1

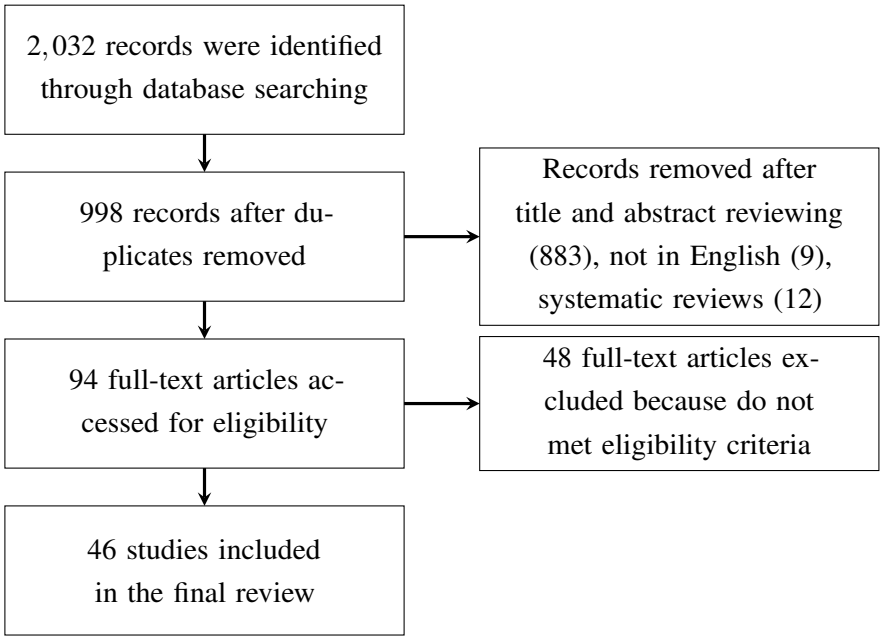

Figure 2.1 PRISMA flow diagram for the selection of literature reviewed.

2.2.2 Study characteristics

Based on the information collected from each article, tables and graphs are created to analyze each feature.

Analysis methodologies used in studies related to the use of social networks for the prediction of acute respiratory infections (ARI)

To analized the first feature, four categories were used to group the methodologies [12]. The category with the highest number of studies is Learning-based Approach with 43.48% followed by Keywords-based Approach with 31.61%, see Figure 2.2. The years in which the highest number of Learning-based studies were published are 2015 and 2018 (four studies each year); for the Keywords-based category, the year with the largest number of published studies is 2018 (five studies).

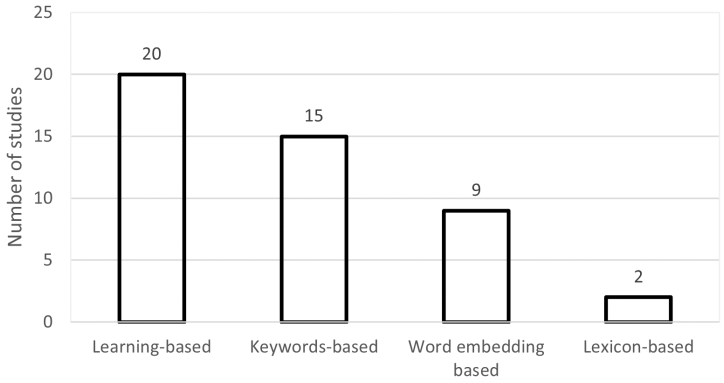

Figure 2.2 Analysis methodologies in social networks.

ML techniques used in investigations related to the extraction of information on Twitter

To investigate our second feature, the models used in each of the four categories were analyzed. The most frequently ML techniques used were those with a Learning-based Approach. The category with the least commonly used techniques was Keywords-based Approach, because the simple structure of these models. Correlation methods can be included in this category, but they do not classify as a ML techniques.

Learning-based Approaches. Tables 2.3 and 2.4 show the characteristics of studies carried out with Learning-based Approaches. Table 2.3 shows the ML technique and the respiratory disease. The respiratory disease is classified as Flu, Influenza, H1N1, Middle East Respiratory Syndrome (MERS) and COVID-19. Tables 2.4 shows the geographical region and the period analyzed in each study presented in Table 2.3. The second column in Table 2.3 shows the ML techniques used in the study. In general, more than one ML technique was tested in the selected investigations, and compared the results obtained in different models between each other. Figure 2.3 displays the most commonly employed machine learning techniques in the studies reviewed. Of the total models, the most used with 29.73%, corresponds to Support Vector Machine (SVM) also referred as Support Vector Regression (SVR). The second place corresponds to Naive Bayes with 21.62%, which is followed by Decision Trees including AdaBoost and XGBoost.

Table 2.3 Learning-based Approaches: Techniques and respiratory diseases.

Study	**Technique**	**Respiratory Disease**	**Ref.**
1	SVR	H1N1	[48]
2	Linear Regression and SVR	ILI	[49]
3	SVM, Naive Bayes, Random Forest, Decision Trees and K-Nearest Neighbour	ILI	[17]
4	SVM, Naive Bayes, Decision Trees and K-Nearest Neighbour	Flu	[18]
5	Multilayer Perceptron	Flu	[50]
6	SVM, C4.5 Decision Trees and Naive Bayes	Influenza	[16]
7	Stacked Linear Regression, SVM, and AdaBoost with Decision Trees Regression	Influenza	[51]
8	Naive Bayes	Influenza	[52]
9	SVM and Naive Bayes	ILI	[53]
10	SVM	ILI	[54]
11	Least Absolute Shrinkage and Selection Operator	Influenza	[55]
12	SVM, AdaBoost and Long Short Term Memory (LSTM)	ILI	[32]
13	SVM, Naive Bayes, Random Forest and Decision Trees	H1N1	[56]
14	SVM, Logistic Regression and Naive Bayes	MERS	[57]
15	Backpropagation Neural Network	ILI	[58]
16	SVM	Influenza	[59]
17	Autoregressive Models, Deep Multilayer Perceptron and Convolutional Neural Network (CNN)	ILI	[33]
18	FastText, Random Forest, Naive Bayes, SVM, C 4.5 Decision Trees, k-Nearest Neighbors (KNN) and AdaBoost	ILI	[60]
19	SVR	ILI	[61]
20	XGBoost, Decision Trees	COVID-19	[62]

Table 2.4 Learning-based Approaches: Region and period.

Study	**Region**	**Period Analyzed**
1	USA	October 4, 2009 to May 16, 2010
2	USA	September, 2009 to May, 2010
3	Portugal	March, 2011 to February, 2012
4	Portugal and Spain	October 30, 2012 to November 30, 2012
5	USA	January, 2011 to April, 2015
6	Victoria, Australia	May, 2011 to August, 2011
7	USA	2011 to 2013
8	USA	Not Defined
9	England and Wales	February, 2014 to August, 2014
10	Cincinnati, USA	November 1, 2014 to May 1, 2015
11	USA, Brazil, Paraguay, Mexico, and Venezuela	July, 2012 to May, 2013
12	31 geolocations (25 in USA and 6 in other countries)	January, 2011 to December, 2014
13	India	Not Defined
14	Global	April 27, 2014 to July 16, 2014
15	USA	October, 2016 to Octuber, 2017
16	Japan	November, 2012 to May, 2013; November, 2013 to May, 2014; November, 2014 to May, 2015
17	USA	2009 to 2010 and 2011 to 2014
18	USA	January, 2018 to May, 2018
19	USA	October, 2016 to October, 2017
20	China	January 22, 2020 to April 13, 2020

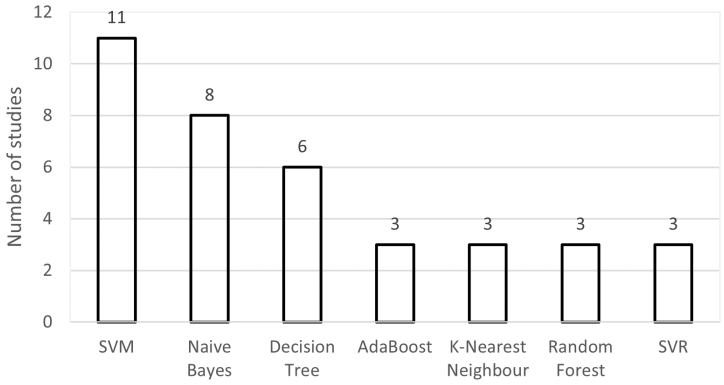

Figure 2.3 Main ML techniques for Learning-based Approaches.

Lexicon-Based Approaches. The analysis related to this classification is presented in Tables 2.5 and 2.6. The results include the technique, the respiratory disease, the region, and the period analysed. Only two studies related to this methodology were found [63, 64].

This approach corresponds to a knowledge-based unsupervised learning method because it does not need annotated labels for training and, instead, uses information available in dictionaries to incorporate labels [12]. The two investigations analyzed data from the United States of America (USA). Velardi et al. [64] created their own prediction model and Abraham et al. [63] used Conditional Random fields (CRFs) and a Log Linear Model.

Table 2.5 Lexicon Based Approach: Techniques and respiratory diseases.

Study	**Technique**	**Respiratory Disease**	**Ref.**
21	Algorithm developed by authors	ILI	[64]
22	Conditional Random fields and a Log Linear Model	Influenza, Common Cold and Listeria	[63]

Table 2.6 Lexicon Based Approach: Region and period.

Study	**Region**	**Period Analyzed**
21	USA	January 26,2013 to May 6, 2013
22	USA	Not defined

Word Embedding Based Approach. Results of studies related to this classification are summarized in Table 2.7 and Table 2.8, also showing the technique, the respiratory disease, the region and the time period. The most frequently used techniques for topic modeling in these studies are shown in Figure 2.4. The techniques described in this analysis were those used for converting the words to a continuous vector representation. Therefore, other techniques used in other parts of the model process were not taken into consideration. Word Embedding is one of the strongest trends in Natural Language Processing (NLP) [12]. The two most used techniques in this category were Latent Dirichlet Allocation (LDA) with 27.27% and Word2Vec with 27.27%. The Word2Vec tool allows the calculation of vector representations of words, and has two learning algorithms for this, continuous bag-of-words and continuous skip-gram; continuous bag-of-words predicts the target word based on previous and subsequent words (in a context window), and continuous skip-gram maximizes the classification of a target word based on another word in the same sentence [27]. In Table 2.7 there are only three studies referring to Word2Vec, but only one specifies the learning algorithm used, continuous bag-of-words Kuang and Davison [28].

Table 2.7 Word Embedding Based Approach: Techniques and respiratory diseases.

Study	**Technique**	**Respiratory Disease**	**Ref.**
23	LDA	Flu	[65]
24	LDA	ILI	[66]
25	LDA	Flu	[67]
26	Word2Vec	Influenza	[12]
27	Word2Vec (continuous bag-of-words)	Swine and Avian Influenza	[28]
28	Word2Vec, GloVe	Not Defined	[29]
29	LaBSE, BERT	COVID-19	[38]
30	Biterm Topic Model	COVID-19	[68]
31	K-Means Clustering	COVID-19	[30]

Table 2.8 Word Embedding Based Approach: Region and period.

Study	Region	Period Analyzed
23	15 Countries in South America	December, 2012 to January, 2014
24	USA	May, 2009 to October, 2010
25	South America	December, 2012 to August, 2014
26	USA	Not Defined
27	Global	2009 to 2012
28	Not Defined	July 12, 2018 to July 12, 2019
29	Global	January 4, 2020 to April 5, 2020
30	Global	March 3, 2020 to March 20, 2020
31	Global	January 6, 2020 to April 15, 2020

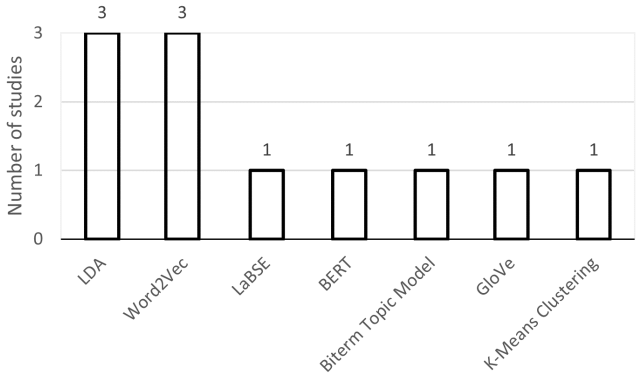

Figure 2.4 Main techniques for Word Embedding Based Approaches.

Keywords-based Approaches. The studies related to Keywords-based Approaches are described in Tables 2.9 and 2.10. Similar to the other approaches, we summarized the technique of the model used, respiratory disease involved in the study, study region, and the time period analyzed. Figure 2.5 shows the main techniques used in the Keywords-based Approaches; Autoregressive Model (35.71%) and Linear Regression Model (35.71%) are the main ones. Keywords-based Approach was the category with the least complex algorithms. The main idea of this category is to evaluate if there is a relationship between the volume of mentions of certain respiratory disease-associated words in tweets and the actual incidence of those diseases.

Linear regression models (autoregressive and non-autoregressive) were the most commonly used analyses in articles classified within this category. These include simple regressions with one independent variable and multiple regressions with more than one. The latter can include data from Twitter as well as from other sources.

Table 2.9 Keywords-based Approaches: Techniques and respiratory diseases.

Study	**Technique**	**Respiratory Disease**	**Ref.**
32	Autoregressive Model	ILI	[69]
33	Linear Regression Model	ILI	[14]
34	Linear Regression Model	ILI	[70]
35	Linear Regression Model	ILI	[15]
36	Correlation Coefficient	Influenza	[13]
37	Correlation Coefficient	ILI	[71]
38	Autoregressive Model	ILI	[72]
39	Not Defined	Not Defined	[73]
40	Autoregressive Model	ILI	[74]
41	Autoregressive Model	Influenza	[75]
42	Linear Regression Model	Influenza	[76]
43	Correlation Coefficient	ILI	[77]
44	Autoregressive Model	Influenza	[78]
45	Nonlinear Gaussian Process	ILI	[79]
46	Linear Regression Model	COVID-19	[80]

Table 2.10 Keywords-based Approaches: Region and period.

Study	Region	Period Analyzed
32	USA	October 18, 2009 to October 31, 2010
33	USA	December, 2011 to April, 2012
34	Korea	October, 2011 and September, 2012
35	New York, USA	October 15, 2012 to May 10, 2013
36	USA	December 2, 2012 to April 7, 2013
37	11 cities in USA	September 29, 2013 to March 1, 2014
38	USA	November 27, 2011 to April 5, 2014
39	Global	June, 2014 and December, 2014
40	Maryland, USA	November 20, 2011 to March 16, 2014
41	Boston metropolis, USA	September 6, 2009, to May 15, 2016, and May 22, 2016, to May 7, 2017
42	United Arab Emirates	November 2016 and January 2017
43	Toronto, Canada	June 26, 2015 to September 10, 2015
44	Italy	October 2016 to April 2017 and October 2017 to April 2018
45	England	March, 2012 to August, 2015
46	USA	January 12, 2020 to April 5, 2020

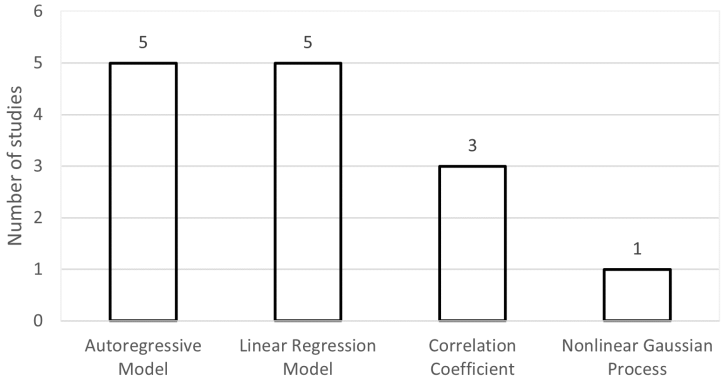

Figure 2.5 Main prediction techniques for Keywords-based Approaches.

Metrics used to measure the performance of the ML techniques

The models and algorithms used in the selected research vary depending on their proposed objectives. Dai et al. [12] helps us segment the models according to their complexity, but even within the same category there may be different objectives. For example, in the Word Embedding Based Approach, we find research that focuses on classifying tweets according to their content, but in addition, time series prediction models are used, with the count of classified tweets. The models found in the different investigations are related to three types of solutions: i) Classification, ii) Regression and iii) Clustering. These three categories were used to classify the algorithms found and their respective metrics. As mentioned before, there are researches with various objectives and different types of ML techniques, so metrics from more than one category were found.

The metrics found associated with *Classification* algorithms are Accuracy, Precision, Recall, F-Measure and Area Under Curve (AUC) from the Receiver Operating Characteristic (ROC). The Correlation metric was considered within the Classification category, given that the objective of this metric is to classify a relationship between two or more variables (existence of a positive, negative or neutral relationship). The metrics found associated with *Regression* algorithms are Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Maximum Absolute Percent Error (MAPE), Root Mean Squared Percent Error (RMSPE) and Mean Absolute Error (MAE). *Clustering* algorithms were found in seven of the nine articles classified as Word Embedding Based Approach, nevertheless, the comparison between metrics results to evaluate the quality of the models was not found in any of them except Chen et al. [30], who describe and graph the obtained values of the sum of squares within a group. In Paul et al. [66] and Mackey et al. [68], the use of the coherence score to optimize the quality of the clusters is mentioned, but not its result. The rest of the articles that were found do not mention any metric to compare the quality of the clusters.

To analyze the performance of the models (related to the selected articles), the methods and performance metrics were compared. Four tables were created, divided by the classification previously used, to describe the metrics used. Comparing the values of the metrics is not a simple task, since the contexts of the selected articles are different. That is, the datasets used for the models are different, the calibration of the models is not the same, and the temporality and spatiality of the data used differs in each experiment. However, we can descriptively identify which models appear most promising by comparing the results from various articles.

Table 2.11 corresponds to Learning-based Approach; within this category, eleven articles were found related to classification algorithms and nine related to Regression algorithms. The

most used metrics for the classification algorithms were accuracy, precision and recall. The models related to the highest value of each of these metrics are: accuracy (84.2%) - A Hybrid Naive Bayesian Classifier with NLP; precision (94.38%) - SVM; and Recall (90%) - Naive Bayes. Regarding regression algorithms, the most used metrics are RMSE and MAPE. The algorithm with the lowest RMSE (0.01%) was LSTM; and with the lowest MAPE (23.6%) was SVM.

Table 2.11 Learning-based Approaches, best model and metrics. The third column contains the Model Category (Cat.): Regression (R) or Classification (C).

Study	**Best Model**	**Cat.**	**Metrics**	**Metrics values**
1	SVR	R	Average error	0.37%
			Standard Deviation	0.26%
2	SVM	C	Accuracy / F1	83.98% / 90.01%
			Precision / Recall	94.38% / 86.63%
3	Naive Bayes + Queries Model	C	F-Mesuare / Precision	83% / 78%
			Recall / AUC	90% / 94.1%
4	Naive Bayes	C	F-Mesuare, Precision, Recall, AUC	Best overall[1]
5	Regularized Multi-task Feature Learning Model, Paraguay data	C	AUC	83.07%
6	SVM with linear kernel and stochastic gradient descent	C	F-Mesuare, Precision, Recall	Best overall
7	SVM with Radial Basis Function (RBF)	R	Pearson Correlation / RMSE	0.989 / 0.176%
			RMSPE / MAPE	8.27% / 23.6%
			Hit Rate	69.4
8	Hybrid Approach with NLP	C	Accuracy	84.2%
9	Autoregressive Integrated Moving Average combined with tweet count regression	R	MAE	8.20
10	SVM based on a linear kernel	C	Accuracy	78%
11	Social Media Nested Epidemic Simulation (SimNest)	R	MSE / Pearson Correlation P-value	Best overall

Study	Best Model	Cat.	Metrics	Metrics values
12	LSTM model only Social	R	Pearson Correlation / RMSE	0.79 / 0.01%
	Media predictors		RMSPE / MAPE	29.52% / 69.54%
13	Naive Bayes	C	F-Measure / Precision	77% / 70%
			Recall	86%
14	SVM classifier with RBF kernel bag-of-words features, MERS data	C	Accuracy	88.26%
15	IAT-BPNN	R	MSE / RMSE / MAPE	Best overall
16	TRAP[2] Model with NLP and High-population areas data	C	Accuracy	70%
17	CNN (Twitter data)	R	RMSE / MAE	3.12 / 4.43
18	Random Forest	C	Precision / F-mesuare	90.5% / 90.1%
			Recall	90.2%
19	SVR with improved particle swarm	R	MSE / RMSE	0.8401 / 0.8225
	optimization (data Region 10)		MAPE	0.7082
20	XGBoost, Feature Sets: Original + Event + SEIR (Susceptible Exposed Infected Recovered)	R	MSE[3]	5862

For the Lexicon Based Approach category, there were only two studies and both of them in the classification category, see Table 2.12. The only metric that was used in both studies is Precision. The highest value is 79%, corresponding to a Conditional Random Fields algorithm that uses a Log Linear Model.

[1]Best overall performance, mentioned by the authors.
[2]TRAP = It is the name of a model that detects disease epidemics.
[3]The MSE values correspond to the scale of the predicted variable.

Table 2.12 Lexicon Based Approaches, best model and metrics. The third column contains the Model Category (Cat.): Classification (C).

Study	Best Model	Cat.	Metrics	Metrics values
21	Model with a threshold $\beta = 0.2$	C	Precision / MRR	69% / 82%
			Coverage	60%
22	Algorithm Conditional Random Fields,	C	Precision / Recall	79% / 58%
	(Measles + pneumonia + mumps) data		F1 score	67%

In category Word Embedding Based Approach, we found combined algorithms, see Table 2.13; Regression models together with clustering models, or Classification models in combination with clustering models. Regression Algorithms were found in two articles, classification algorithms in six articles and clustering algorithms in seven of them. The metrics used in this category are very different, and only one metric is repeated three times, Accuracy, with a maximum value of 87.1% in the word embedding model Word2Vec, Dai et al. [12]. Although this metric does not evaluate the quality of a clustering model, it does allow evaluating the classifications of the clusters, and the creation of new ones. Dai et al. [12] use the cosine similarity to define the number of clusters in its model; this value compares the closeness between two numerical vectors. Each word in a tweet is converted to a numerical vector and compared to vectors of keywords like “influenza", and the objective is to classify each tweet based on its proximity to the keywords.

Table 2.13 Word Embedding Based Approach, best model and metrics. The third column contains the Model Category (Cat.): Regression (R), Classification (C) or Grouping/Clustering (G).

Study	Best Model	Cat.	Metrics	Metrics values
23	HFSTM[1] model	R	RMSE	Best overall
		G	Not specified	Not specified
24	Ailment Topic Aspect Model	C	Pearson Correlation	Best overall
		G	Coherence score	Not specified
25	HFSTM-A model	R	RMSE	Best overall
		G	Not specified	Not specified
26	Word Embedding Clustering,	C	Precision / Recall	96.2% / 75.6%
	similarity threshold = 0.6		F1 / Accuracy	84.6% / 87.1%
		G	Cosine Similarity	Not specified
27	Algorithm II + CNN, Influenza Dataset	C	Accuracy	72.84%
28	BiLSTM-CRF (word + char + POS[1],	C	Precision / Recall	94.93%/81.98%
	word embedding method)		F-score	87.52%
29	SVM classifier applied on LaBSE	C	Accuracy / F1 (Micro)	86.92% / 87.6%
			F1 (Macro)	88.1%
		G	Not specified	Not specified
30	Model with tweets that included	C	Spearman	0.45
	users self-reporting symptoms		Correlation	
	and self-reported recovery	G	Coherence score	Not specified
31	Twitter data: Kmeans(6),	G	Sum of squares	Range: 1000
	Weibo data: Kmeans(5)		within a group	to 1100

Most models in the category Keywords-based Approach correspond to either regression or classification (correlations), eleven and four articles respectively; see Table 2.14. The most used metric for regression algorithms is RMSE, and its lowest value is 0.001897, related to a multiple linear regression model. The comparison between the different models that use the

[1]POS tagging is an additional word feature.

RMSE is not reliable, since its value depends on the scale on which the predicted variable is. Corresponding to the classification algorithms, the most used metric is the Correlation Coefficient with a maximum value of 0.79.

Table 2.14 Keywords-based Approaches, best model and metrics. The third column contains the Model Category (Cat.): Regression (R) or Classification (C).

Study	**Best Model**	**Cat.**	**Metrics**	**Metrics values**
32	Model without retweets, Syndrome	R	Correlation Coefficient	0.9846
	Elapse (time = one week)		RMSE	0.318
33	Multiple linear regression model with ridge regularization unweighted	R	RMSE	0.001897
34	LASSO with a marker 'novel flu'	R	Coefficient of Determination	0.853
35	Daily infection tweets model	R	Correlation Coefficient	0.763
36	Google Flu Trends and USA tweets model	C	Pearson Correlation	0.79
37	Correlation model (San Diego data)	C	Correlation Coefficient	0.93
38	Model with Twitter data	R	MAE	0.1866
39	Not Defined	C		
40	Model with ILI and Google Flu Trends data	R	Log-likelihood	Not Defined
41	ARGO Model (athena+Google+Flu-Near-You data)	R	RMSE / MAE / MAPE / Pearson Correlation	Best overall
42	Experiment 3 (January data)	R	RMSE	0.14
43	Model with Twitter data related with heat syndrome and temperature	C	Cross-Correlation Coefficient	0.5

Study	Best Model	Cat.	Metrics	Metrics values
44	Model M4 (nowcast)	R	RMSE / MAE	0.1972 / 0.1191
			MAPE / Correlation Coefficient	10.75 / 0.9867
			Coefficient of Determination	0.9704
45	Google model	R	Pearson Correlation / MSE	0.96 / 3.86
	(national level: England data)		RMSE / MAE	1.96 / 1.47
			MAPE / Mean Error	14.10% / 0.54
46	Google Trends model	R	Coefficient of Determination	0.9209

Of the 46 articles selected, classification algorithms were found in 23 of them, regression algorithms in 22, and clustering algorithms in 7. The most used metric for the classification algorithms was precision, for the regression algorithms was RMSE, and for the clustering algorithms was the coherence score.

Main countries related to the extraction of information on Twitter with ML techniques

Analysis of countries where studies were carried out is a relevant issue, as it allows to understand the regions in which Twitter has been used for epidemiological surveillance studies and, therefore, to what extent the results of these studies can be generalized. It also helps identify areas that require further research, such as adaptation of text analysis in diverse languages. In addition, it can allow to assess if research in this area is consistent with the intensity of Twitter use in a region, or specific areas of interest (such as health issues) in certain populations.

To answer this, the countries analyzed in the studies were counted. To summarize the studies, only articles that reported analyses carried out in a single country were taken into consideration; studies that made reference to more than one country were not considered. In total, 32 studies were analyzed to determine the geographical distribution of studies. The country with the

highest number of studies was the USA, 66% of the total (21 studies). The second country in which more studies were carried out was England with 7% of the total (two studies). Only one study was carried out in each of all other countries (3%), see Figure 2.6.

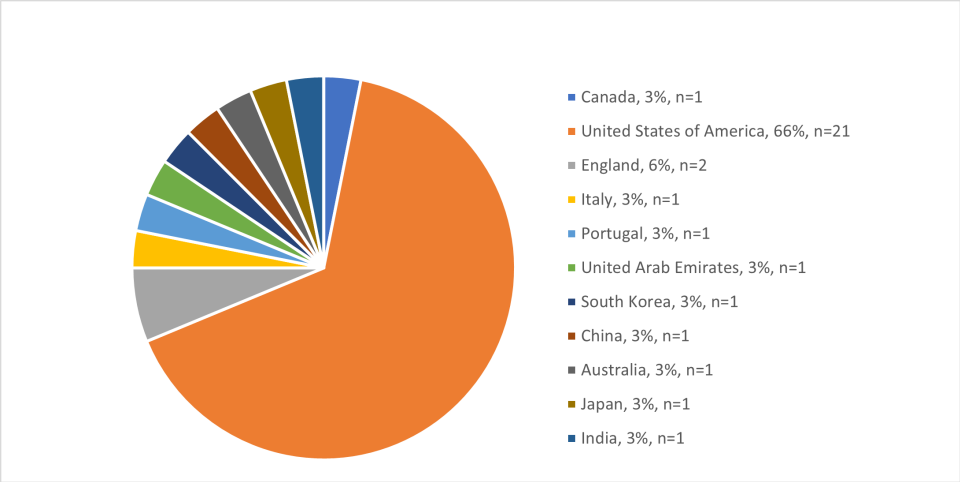

Figure 2.6 Distribution of the selected studies worldwide. Only studies focusing on a single country are shown; n= number of studies.

The time series periods analyzed in studies

As for analysis of time periods included in the studies, this is of relevance, as it allows to understand if findings are consistent through time, and if observations are stable or require adjustments over time. Overall, most studies were carried out over relatively short periods of time (one year or less), which indicates the need for further research regarding the potential need of model adjustments, as language use, and terms may change over time. A clear example of this is the “creation" of new terms (such as COVID-19 and SARS-CoV-2) during epidemic periods, which would require adjustments in text analysis in order to adapt to changing epidemiological circumstances.

To analyzed this feature, we based the analysis on the publication year of the selected studies. Figure 2.7 shows the number of studies published during each year. The studies included in our review were published between 2011 and 2020. The years in which more studies were published were 2014 and 2018 (9 studies in each year). Keywords based approach were more frequently use during early years (2011-2014) and Learning based approach have been used most frequently since 2015; Word embedding based approach appears to be increasing in use in recent years, while Lexicon based approach is seldom used.

The highest number of studies found in the Keywords-based approach category were in 2018, for Learning based approach it was 2015 and 2018, for Lexicon based approach it was 2014 and 2017, and for Word embedding based approach it was 2020 (with 60% of the total for that year), so this category is one of the most recently used.

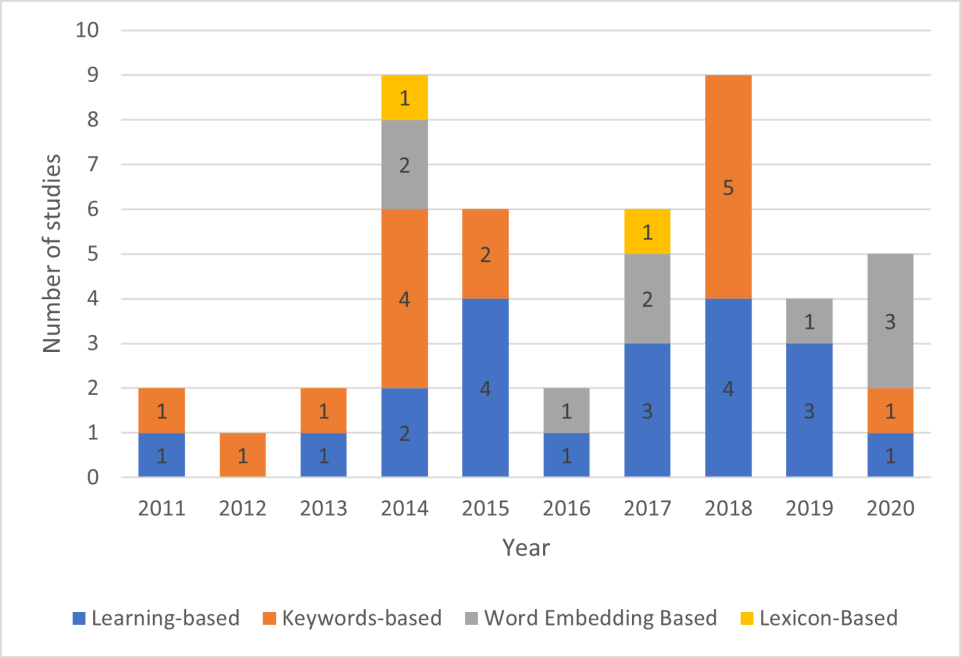

Figure 2.7 Studies classified by publication year and type of analysis methodology.

2.3 Discussion

Respiratory infectious diseases continue to pose significant threats to human health. Traditional surveillance systems are labor intensive and frequently require laboratory testing for disease confirmation. Information derived from social media is increasingly considered as a potential source of data that could contribute to epidemiological surveillance. Twitter has been identified as the most attractive platform to provide information for this purpose [9, 81]. Our results support this notion and provide information regarding areas of interest for future research.

Processing data from social networks to extract knowledge and selecting the most appropriate ML techniques is not a simple task. In this systematic review, we classified ML techniques according to the objectives of the analysis and the level of natural language processing on social media posts, as well as the most used metrics depending on the problem that the models solve.

Our results expand on finding from previous literature reviews. During the literature search for the present work, we identified twelve Systematic Reviews, and seven fulfilled the thematic criteria for our study. However, since these publication are considered secondary research (since they do not provide primary data), they were analyzed separately to avoid counting two or more times the results reported by primary research publications. A summary and main findings of these Systematic Reviews are presented in Tables 2.15 and 2.16.

Table 2.15 Selected systematic reviews. The first column is the identifier (Id); the third column contains the number of studies analyzed (#) and the fourth column the methodology used.

Id	Title	#	Methodology	Ref.
a	Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review.	32	Not referenced	[3]
b	Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review	60	PRISMA	[8]
c	Social media based surveillance systems for healthcare using machine learning: A systematic review	26	Not referenced	[9]
d	Twitter as a Tool for Health Research: A Systematic Review	137	PRISMA	[81]
e	Adoption of Digital Technologies in Health Care During the COVID-19 Pandemic: Systematic Review of Early Scientific Literature.	124	PRISMA	[82]
f	Big data analytics as a tool for fighting pandemics: a systematic review of literature	45	Methodi Ordinatio	[83]
g	Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review.	28	PRISMA	[84]

Table 2.16 Selected systematic reviews (databases, analysis period and notable findings). The first column is the identifier (Id).

Id	Databases	Period	Notable findings
a	PubMed,Scopus and Scirus	1990-2011	There is a need for technologies that monitor health-related issues on the Internet. Although the acceptance of the scientific use of information from social networks generates diverse opinions.
b	PubMed, Embase, Scopus, and Ichushi-Web	-Feb. 2013	The analysis of the literature demonstrates that information from social networks improves public health mechanisms and helps identify vulnerable populations.
c	IEEE, ACM Digital Library, ScienceDirect and PubMed	2010–2018	Most of the studies found focus on flu or influenza-like illness (ILI). Twitter is the social network with the largest number of studies developed, and the most used ML technique is Support Vector Machine.
d	PubMed, Embase,Web of Science, Google Scholar and CINAHL	-Sep. 2015	Health research based on Twitter data is a growing field. To describe the use of Twitter in health areas, a taxonomy with six categories was created.
e	MEDLINE and medRxiv	Jan. 2020 -Apr. 2020	Artificial Intelligence algorithms based on image recognition and clinical data are a promising field. User health tracking applications are effective, but there is a debate about data privacy.
f	Web of Science and Scopus	2014-2020	The two main sources of Big data information are internet search engines and social networks; and some common techniques for analyzing this data are correlation and regression.
g	ScienceDirect, PubMed, Web of Science, IEEE Xplore and Scopus	Jan. 2010 - Jun. 2020	The use of technologies such as artificial intelligence and machine leaning, in the field of sentiment analysis on social networks, contributes to an improvement in results.

In addition, we conducted a review of the literature to identify additional Systematic Reviews published between 2021 and 2024 regarding the topic of interest of this review. The search was conducted using the same key words as described in the Methods section, but limiting results to systematic reviews only. There were seven new reviews identified; a summary of these is presented in Table 2.17 and Table 2.18.

Table 2.17 Systematic Reviews published between 2021 and 2024. The first column is the identifier (Id), the third column contains the number of studies analyzed (#), and the fourth column the methodology used.

Id	Title	#	Methodology	Ref.
h	The application of artificial intelligence and data integration in COVID-19 studies: a scoping review	794	PRISMA	[85]
i	Surveillance of communicable diseases using social media: A systematic review	23	PRISMA	[1]
j	Applications of machine learning for COVID-19 misinformation: a systematic review	43	PRISMA	[86]
k	Classification of Covid-19 misinformation on social media based on neuro-fuzzy and neural network: A systematic review	34	Kitchenham [87]	[88]
l	The Impact and Applications of Social Media Platforms for Public Health Responses Before and During the COVID-19 Pandemic: Systematic Literature Review	678	Not referenced	[89]
m	Promoter or barrier? Assessing how social media predicts Covid-19 vaccine acceptance and hesitancy: A systematic review of primary series and booster vaccine investigations	113	PRISMA	[90]
n	Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review	15	PRISMA	[91]

Table 2.18 Systematic Reviews published between 2021 and 2024 (databases, analysis period and notable findings). The first column is the identifier (Id).

Id	Databases	Period	Notable findings
h	National Institutes of Health (NIH) LitCovid (part of PubMed) and the World Health Organization (WHO) COVID-19 database	-Mar. 2021	Identification of research areas related to COVID-19 in which AI is applicable. In the AI applications found, there is a lack of integration of heterogeneous data.
i	ACM Digital Library, IEEE Xplore, PubMed, and Web of Science	-Mar. 2020	Data mining of health-related content on social networks can represent a tool for monitoring public health and predicting contagious diseases. Another advantage of NLP analysis of content on social networks is remote detection of public health through the Internet.
j	Scopus, Web of Science, and Google Scholar	-July 2021	Deep learning-based classification methods are more effective in classifying COVID-19 misinformation than ML methods. The lack of reference data sets to compare research was found, as well as the lack of multilingual and multimodal information services.
k	IEEE Xplore, SpringerLink, ScienceDirect, Scopus, Taylor and Francis, Wiley, Google Scholar	2018–2021	Studies on the classification of misinformation on social networks have increased since the COVID-19 pandemic. Methods like Adaptive Neural-based Fuzzy Inference System (ANFIS) and Deep Neural Network were shown to be effective for misinformation classification. The use of algorithms combining both methods is recommended.

Id	Databases	Period	Notable findings
l	PubMed, Medline and Institute of Electrical and Electronics Engineers Xplore	Dec. 2015 - Dec. 2020	A growing number of studies were found related to the classification of misinformation since the COVID-19 pandemic. Existing applications of social network information contribute to the monitoring and surveillance of public health.
m	PubMed, Scopus, and Web of Science	Jan. 2020 - Feb. 2023	More negative associations with vaccination perception were identified than positive ones; however, predictions about positive perceptions show an increasing trend. This positive perception was also identified in publications related to booster vaccination.
n	PubMed, Embase, Web of Science, and Scopus	-Dec. 2023	Very promising results were found in research related to the use of LLM (Large Language Model) such as GPT-4 and BERT in social media data, and the detection of urinary tract infections or Lyme disease surveillance. Research requires greater depth in the use of LLM and AI models in areas of diagnosis, surveillance and prediction of disease progression; as NLP and LLM are expected to contribute to the management of infectious diseases.

While Systematic Reviews published up to now are within the same broad interest area of our review, several differences were observed, including the main focus of interest (for instance, sentiment analysis [84] or vaccine acceptance [90]) and the time period of studies included in the analysis. In this context, our systematic review provides a contribution in some areas which have previously not been analyzed completely, such as the classification of the studies according to the algorithms used in the mathematical models, either based on the objective they want to achieve (classification, regression or clustering) or on the complexity of the models according to the analysis of information from social networks. The systematic review of Gupta

& Katarya [9] identified the ML techniques from its selected research, but does not delve into the results obtained from them. In our research we analyzed the algorithms used based on the problem they solve and we also analyzed the metrics used, comparing them with other research. Alamoodi et al. [84] performed a classification on the selected studies, but do not provide information about the algorithms used per study; in addition, their research focused exclusively on Sentiment Analysis associated with diverse infectious diseases, and not on potential use on epidemiological surveillance. More recent reviews pay particular attention to COVID-19 [84, 86, 88]. Of note, in general, they do not address disease surveillance or forecasting, but assess other aspects (such as social challenges, misinformation, and vaccine acceptance).

In the present work, we used the methodology outlined in the PRISMA statement. As noted in Tables 2.15 and 2.17, this methodology is frequently used in health-related research. PRISMA was the methodology used in nine of the 14 related reviews, while other methodologies were less frequently reported.

2.3.1 Summary of evidence and perspectives

In our review, which was centered in studies that assess the correlation between publications on Twitter and the actual number or rate of ARI, we identified two areas of major importance during the conduction of these studies: 1) identification of relevant messages and extraction of information; and 2) comparison of social media information with disease data. Regarding data extraction methods, a Learning-based approach is the most frequently used methodology. While the categorization approach proposed by Dai et al. [12] provides very precise divisions to differentiate studies corresponding to each category, the description of the methodology in some studies is unclear or ambiguous. In other studies there is a mix of characteristics of two different methodologies. As such, future research should provide clear information regarding the methodological approach that is selected by the authors to allow to assess the advantages and drawbacks of each method. In addition, patterns of Twitter usage across time need to be considered in message analysis methods, including trends in language use, language contagion, amplification patterns, as well as misinformation contents [92, 93]. Since these factors have been reported to vary in different countries and different languages, text analysis methods that may be applied across several cultures would be useful in order to avoid the need to tailor these procedures to specific settings.

Another aspect that is noticeable in our study is that temporal trends in the use of different text analysis methodologies do not allow to clearly establish a preferred method due to the relative short time frame (2011-2020) in which studies were carried out, as well as the overall number of publications. However, it is of note that the Word Embedding based approach has been used preferentially in most recently years; due to the use of unsupervised models and non-manual classification (in training datasets), it is foreseeable that the preference for models of this category might increase in the coming years.

With respect to the regions analyzed by study, the largest number of studies was carried out in the USA (66%). As discussed in the limitations section, this might be explained in part due to the inclusion of only studies written in English in our review; this reduces studies performed in countries with other official languages. In addition, the intent for Twitter use may vary in different countries and, therefore, the predictive ability of message data analysis requires to be assessed in diverse regions in order to generalize its use. In this regard, another important issue that requires to be addressed is the convenience of using surveillance data based on case-definitions agreed upon internationally. This is not always an easy task, as even when case-definitions are similar, financial, social, and health system organization characteristics affect the implementation of surveillance activities.

Overall, most studies reported good results when data from Twitter was used in models to assess the occurrence of respiratory infection outbreaks. In general, analyses focusing on correlations showed better results than predictive models. Of note, most studies assessed several models, showing that even with the same data, different models can provide diverse results. This highlights the fact that comparing models between studies may be difficult, and should be interpreted with caution. As such, studies that assess the performance of models in diverse regions and through different time frames would be valuable to determine whether the proposed approaches can be applied in different scenarios. This will be of particular interest, since most studies that were identified in this review were carried out in a limited geographical area, and during a relatively short period of time.

Unlike previous systematic reviews, this review focuses on studies related to Twitter posts on ARI. In addition, the diversity and temporal trends in the use of different ML techniques has not been previously performed. Learning-based Approach was identified as the most widely used methodology to understand the relationship between information provided by Twitter users and the actual number of infections. Within this category the most used ML technique is SVM, but we expect that in the following years models based on deep learning may become

more popular, for example models based on LSTM and CNN. According to the objective of the algorithms, the most used category is classification, and its most used metric is precision.

Chapter 3

Twitter and Acute Respiratory Infections Data

This section presents the collection and cleaning steps. For the storage and sharing steps, we use Comma Separated Files (CSV) among the different phases of the proposed computational methodology.

3.1 Twitter API and the limitation on downloading social media data

Our first data source is Twitter (currently known as X). The advantage of Twitter over other social media platforms is the availability of the data; although these policies have recently changed due to the new company's management. Many researchers have the same difficulty to obtain data from social networks, and find in Twitter, the ideal information provider for their research [8]. Other data sources are reports published by the Mexican government (ARI and COVID-19 time series) and posts published on [43].

3.1.1 Twitter Data

The following are some tools and applications used in the researches analyzed to extract tweets from the Twitter platform. Some of them are applications that range from the collection, filtering and selection of tweets to the publication of the information obtained from them. ESA and Twitcident fall into this category.

1. ScraperWiki. This platform served as a mechanism through which researchers could share preconstructed segments of code designed to extract and analyze public Twitter data [94]. This platform it is no longer available.

2. Tweepy. It is a Python library for accessing the Twitter API. The tweepy library methods are diverse, from authentication, downloading twetts, to manipulating and modifying the downloaded information [95].

3. ESA. Emergency Situation Awareness (ESA) is a platform that monitors the Tweet stream across Australia and New Zealand; this system mines microblogs in realtime to extract and visualise useful information about incidents and their impact on the community. The information that microblogging media users can generate regarding a natural disaster, an explosion, massive contagion, among others, is vitally important to prevent and notify the authorities. For that reason, ESA use language processing and data mining techniques to analyse Twitter data in real-time. The main advantage of ESA is the transmission in real time of the information shared by citizens; for example, a witness in a local fight, transmits the information through a tweet and the system captures this type of non-common events reported on microblogging media [96]. Zuccon et al. [16] use ESA as a method of downloading tweets. As additional information, ESA includes google maps and Yahoo services.

4. Twitcident. It is a framework and web based system for filtering, searching and analyzing information about real-world incidents or crises. It is automatically connected to emergency services. By detecting unusual incidents in Twitter posts, it generates an alarm. The main objective of this application is to use the information that Twitter users publish on the platform to detect incidents such as fires, earthquakes, storms, etc. This application constantly monitors Twitter posts and filters the information it considers relevant for the detection of incidents [97]. One of the main components of Twitcident is semantic enrichment, since it helps to identify tweets with content related to incidents, in addition, it summarizes the information of each publication by finding entities within the text; these entities represent places, people, organizations, products or events. Semantic enrichment determines how important an entity is to the meaning of a text. The Twitcident interface allows users to access the real-time analysis panel, where current incidents are displayed; in addition, the interface allows the filtering of information to search for a particular incident. To complement the information, diagrams and graphs are shown about the events and related publications. In the Netherlands, registered incidents are reported

through the P2000 communication network, indicating when, where and what has happened. Twitcident uses the Twitter API to collect messages, images and videos from this platform. The application has the Named Entity Recognition (NER) module based on semantic enrichment. The first action of this module is the detection of entities (people, Places); later it makes use of the concepts of DBpedia to map the entities [97]. DBpedia is a project that builds a large-scale multilingual knowledge based by extracting structured data from Wikipedia editions in 111 languages. Some of the DBpedia applications are data integration, named entity recognition, topic detection, and document ranking. The DBpedia user community creates mappings from Wikipedia information representation structures to the DBpedia ontology [98].

The method used in this research to access tweets, was Twitter API and its Python library Tweepy. This library contains various methods, some examples: authentication, download data (tweets), manipulate data and modify downloaded information [95]. The programming language used was Python; with its library Tweepy used to configure the Twitter API and query parameters. The main function used was “search full archive". This function searches for tweets in the entire Twitter base from 2006 to date. In addition, it works under Boolean searches, which delimit the searches. Boolean searches help specify selection criteria for downloading tweets; for example, if it is declared in the query that only tweets that contain the word “COVID-19" are downloaded, only these will be downloaded. We downloaded tweets that contain any of the following words: asthma, bronchitis, respiratory, cough, and flu (in Spanish correspond to ‘asma’, ‘bronquitis’ ‘respiratorias’ and ‘tos’ or ‘gripe’, respectively), as suggested by Gónzalez-Bandala et al. [99]. Additionally, we added the words ‘COVID19’ and ‘COVID’; with the limitation of a maximum of 128 characters per query set by Twitter for our type of Twitter developer account. Additionally, two more filters were declared; the first one was only tweets in Spanish language, and the second, tweets generated in the state of San Luis Potosí, Mexico. The account given by Twitter for downloads has a level of “Elevated". This level allows fifty queries per month, and five thousand downloaded tweets per month. These limits correspond to the use of the “Search Tweets: Full Archive" function, for historical queries in the entire Twitter database. The total number of tweets downloaded in a time period between 2020-01 and 2022-12 (yyyy-mm), was 5,759. It was done through requests for a maximum of 1,000 tweets per query. Table 3.1 shows an example of downloaded tweets [43].

Table 3.1 Example of the result of a query; three tweets, with tweet information, user, location, publication date and detected tags.

Id	Text	User name	Loc	Date	Tags
1	Nadie debe pagar IVA en las pruebas COVID 19 PruebasCovidSinIva	Konishi Ramirez		2022-02-11 14:36:34+00:00	Pruebas CovidSinIva
2	Una parte si es COVID19 pero la mayor parte el mal manejo de las autoridades	zazuetarturo	San Luis Potosi	2022-02-08 20:28:17+00:00	COVID19
3	La vacuna contra el Covid19 es esencial para salvar vidas Les comparto las sedes	TonoLorcaSLP	San Luis Potosi	2022-02-08 14:15:58+00:00	Covid19

3.1.2 Government Data: ARI Data and COVID-19 Data

The General Directorate of Epidemiology is an organism that belongs to the Mexican Health Ministry (known as Secretaría de Salud). Its functions include monitoring epidemics within Mexican territory and publishing periodic epidemiological data. The information reported by this agency includes ARI [100] and COVID-19 cases [101]. The information is represented as time series. The time series downloaded correspond to a time period between 2020-01-01 and 2022-12-31 (yyyy-mm-dd). The year 2020 was chosen as the starting point due to the onset of the COVID-19 pandemic [43].

3.2 Cleaning social media data and COVID-19/ARI time series

There are numerous methods, tools, and applications for data cleaning. Table 3.2 shows a summary of the main ones analyzed in this work. Some of them contain several libraries for cleaning data, and others are just a tool for a particular process.

Natural Language Toolkit

Natural Language Toolkit (NLTK) is a platform for building Python programs to work with human language data. Some of the applications for text manipulation that can be carried out with this platform are the following: classification, tokenization, stemming, tagging, parsing, and semantic reasoning and wrappers for industrial-strength NLP libraries. It also allows

Table 3.2 Data cleaning tools.

Related Authors	Tools	Comments
[18]	Language detection library for java	Java library to detect 53 languages.
[12]	Word2Vec	It is an open source deep learning toolkit.
[102]	WordNet	English language lexical database.
[17]	Natural Language Toolkit (NLTK)	A platform for building Python programs to work with human language data.
[16]	Medtex	It is an application to identify SNOMED-CT concepts in texts.
[103]	SNOMED-CT	Systematized Nomenclature of Medicine– Clinical Terms.
[104]	UMLS	The Unified Medical Language System.

working with interfaces over 50 lexical resources, such as WordNet (lexical database of the English Language) [105]. Aside from tokenizing tweets and removing stopwords, Santos and Matos [17] use Natural Language Toolkit to generate bigrams for each word, used to measure the probability that one word is preceded by another. In total, 5106 bigrams were generated, although their application did not generate any improvement in the classification.

Medtex

Medtex (Medical Text Extraction) is an application to identify SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms) concepts in texts. Was created with GATE (General Architecture for Text Engineering), an open source architecture for natural language processing (NLP) [103]. SNOMED-CT is developed and maintained by College of American Pathologists, and it is a comprehensive clinical reference terminology (contains more than 360,000 concepts and over 1 million relationships). Another reference source for medical terminology is the standardised coding system UMLS (The Unified Medical Language System) [104]. Medtex uses tokenization to separate the words of a free text into tokens; they can be words, numbers, punctuations, or spaces. After that, Medtex maps the strings in the free text to the closest concepts in the UMLS medical terminology resource; and UMLS concepts are mapped to SNOMED-CT concepts [103]. Medtex uses the message architecture based on message queues, and allows the reception of messages in parallel from the same queue. Zuccon et al. [16] expand the Medtex platform to include information extraction capabilities for specific

entities; for example, to extract specific information from some user (e.g. @Username). In addition, they used this application to extract blank spaces, word roots, web page links (within the tweet), hashtags and @ symbol.

The following is a common nine-step process for data cleaning; Not all steps are completed in the reviewed investigations, but each investigation has its own data cleaning process.

1. Language detection. There are libraries created in different programming languages for the detection of languages in social media posts. [18] used a Language detection library for java to detect 53 languages. It has a precision of 0.99 and is based on Bayesian filters. In general, the researchers select the tweets of a single language, and the other tweets of the same query are eliminated (even if they have the same geographic area).

2. Remove Blank rows in Data. In this step, all blank spaces are removed from each tweet.

3. Change all the text to lower case. All letters are changed to lowercase. If two words are the same, but one is uppercase and other one is lowercase, can be identified as a different word.

4. Word Tokenization. It is a process of breaking a stream of text up into words or other meaningful elements called tokens. The disadvantage of this process is the loss of semantic relationships between words, [26]. See an example of word tokenization in the table 3.3.

Table 3.3 Example of word tokenization.

Text	Tokens
“The dog is running in the park.”	“The”, “dog”, “is”, “running”, “in”, “the”, “park”, “.”

5. Remove Stop words. Text may contain stop words like ‘the’, ‘is’, ‘are’. Stop words can be filtered from the text to be processed. There is no universal list of stop words in natural language processing research, however the python nltk module contains a list of english stop words [106]. In Nargund and Natarajan’s research, care was taken not to eliminate the following words: “I”, “me” and “you”, which are classified as stop words. The reason is to keep tweets where there is a possible declaration of illness, [107]. See an example of stop words in table 3.4.

6. Remove Non-alpha text. All characters that are not within the alphabet are removed.

Table 3.4 An example of ten stop words [106].

NLTK's list of english stopwords
i
me
my
myself
we
our
ours
ourselves
you
your

7. Word Lemmatization. The objective of this function is to reduce the inflectional forms of each word into a common base or root. See the lemmatization example in table 3.5.

Table 3.5 Lemmatization example.

Word	Lemmatization
stuying	study
studies	study
study	study

8. Encoding. It is used to transform Categorical data of string type in the data set to numerical labels which the model can understand, [108].

9. Word Vectorization. It is a general process of turning a collection of text documents into numerical feature vectors, [108]. Vectorization is the basis for the development of Word embedding methods.

Volkova et al. [32] use a data cleaning that consists in tokenizing tweets, transforming uppercase letters to lowercase, removing punctuation symbols, removing infrequent tokens (with a frequency less than five), deleting of hashtag symbols and mention symbols.

3.2.1 Data cleaning process

Data cleaning is a fundamental part of each phase of our computational methodology, since the data is collected in different formats. So, different procedures are necessary to standardize

these formats. The following steps are needed for each data source. It should be noted that each phase contains its own data cleaning process; for example in the first phase all data cleaning was done in Dataiku, and the last phase was done completely in Python code [43].

1. COVID-19 data.
    (a) Set the date format to the corresponding column.
    (b) Apply date filter according to period 2020-01 and 2023-12 (yyyy-mm).
    (c) Extraction of date components: year and month.
    (d) Weekly sum of COVID-19 confirmed cases.
2. ARI data.
    (a) Extraction of information from PDF files through the Python Tabula library. This script involves positioning functions on the PDF sheets to extract the corresponding tables. Additional cleaning is done by removing empty spaces, column renaming and construction of a Pandas dataframe for later export to CSV format.
    (b) Maintain rows exclusively from the state of San Luis Potosí.
    (c) Extraction of date components: year and month.
    (d) Set the date format to the corresponding column.
    (e) Union of the information of all the weeks in a single dataset.
    (f) Set the date filter according to period 2020-01 and 2023-12 (yyyy-mm).
3. Twitter data.
    (a) Set the date format to the corresponding column.
    (b) Removal of unknown characters (excluding letters and numbers). In this step, all emoticons, non-alphanumeric symbols, and pictorial characters are eliminated. Since the tweets downloaded from the API are exclusively in Spanish, the presence of characters from other languages is naturally avoided.
    (c) Elimination of line breaks and elimination of words made up of less than 2 letters and more than 21.
    (d) Normalization of words to lowercase.

Chapter 4

Machine Learning-Based Computational Methodology

Selecting appropriate algorithms and models can be a complex and confusing process. To address this, we propose a computational methodology structured in three phases. Each phase involves testing and comparing models to support their selection and refinement. This methodology leverages machine learning platforms such as Dataiku and the Python programming language. The primary objective of this research is to explore the relationship between tweets and official health authority reports on Acute Respiratory Infections (ARI) and COVID-19. To this end, we compare various machine learning prediction models, using confirmed ARI and COVID-19 cases as the target variable, and tweets referencing potential cases of these diseases as the independent variable [43].

For each phase of the proposed computational methodology, we follow a data science workflow [44], which contains a workflow consisting in three steps: data, analysis and communication, see Figure 4.1. The Data phase includes data collection, cleaning, storage, and sharing. The Analysis phase involves training and forecasting with predictive models. Finally, the Communication phase focuses on presenting and sharing results, emphasizing effective knowledge dissemination through visualization [43].

Figure 4.2 presents the proposed computational methodology, which consists of three phases. The first phase primarily involves using the Dataiku platform for data cleaning, transformation, and implementation of machine learning (ML) models. Dataiku, a collaborative platform for data science and ML, was chosen for its user-friendly interface, ability to standardize data formats, and extensive built-in functionalities for data processing and modeling. This phase enables rapid comparison of multiple ML prediction models, using confirmed cases of ARI

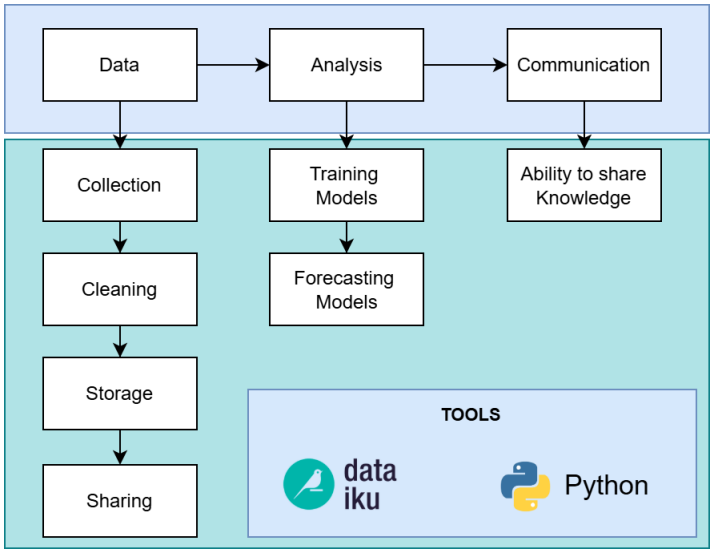

Figure 4.1 Data science workflow. There are three main steps: Data, Analysis and Communication [44, 43].

and COVID-19 as the target variable, and tweets related to potential cases as the independent variable. During this stage, basic data preparation is performed in Dataiku, excluding tweet classification and selection algorithms. The second phase focuses on comparing deep learning models and includes a tweet selection component using custom algorithms developed in Python. While most of this phase is executed in Dataiku to maintain speed and efficiency, Python code is integrated for tasks beyond Dataiku's native capabilities—such as implementing Word2Vec for word embedding. This hybrid approach, combining Python with Dataiku's tools, is a key feature of the methodology. The third phase is conducted entirely in Python, using the top-performing models identified in previous stages. Transitioning from Dataiku pipelines to Python code enables the use of recursive algorithms, allowing for a broader range of tests and more robust model validation [43].

4.1 Machine Learning Model Algorithm Comparison with Dataiku (Phase one)

As described in the "Twitter Data" section (see §3.1.1), the tweets were collected from the Twitter API using keywords such as asthma, bronchitis, respiratory, cough, and COVID-19. The data was filtered to include only posts from the state of San Luis Potosí and written in

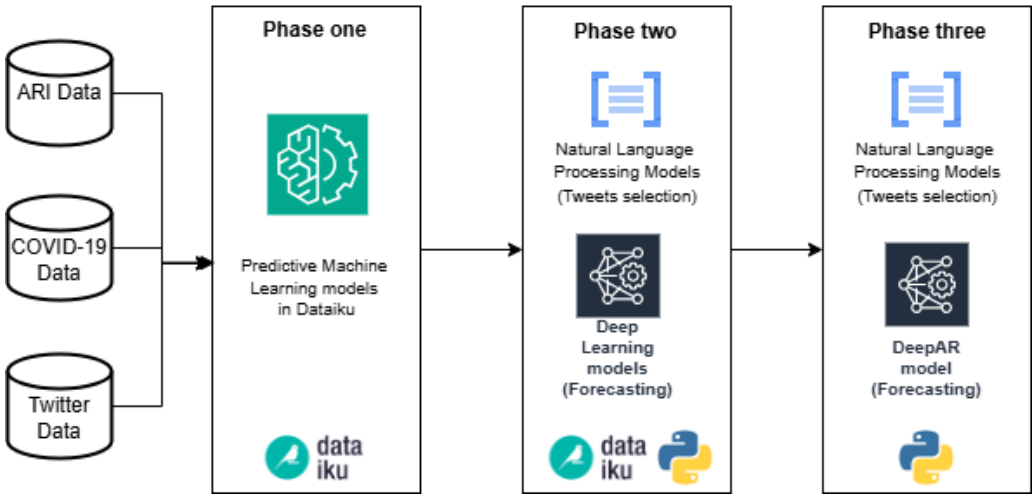

Figure 4.2 Computational Methodology to simplify the process of selecting ML algorithms for forecasting [43].

Spanish. The goal of this phase was to determine suitable machine learning (ML) algorithms for prediction. We employed models using continuous variables, with the target (dependent) variable being the weekly number of confirmed ARI and COVID-19 cases, and the independent variable being the weekly count of relevant tweets (see Figure 4.3). Data preprocessing (including cleaning and transformation), model training, and testing were conducted on the Dataiku platform. A key advantage of Dataiku is its ability to test and compare a wide variety of ML models efficiently. The models evaluated included: Random Forest, Ordinary Least Squares, Lasso Regression, Decision Trees, Support Vector Machines, K-Nearest Neighbors, and Artificial Neural Networks (hereafter referred to as Feed Forward). To evaluate model performance, we used Maximum Absolute Percent Error (MAPE) and Root Mean Squared Error (RMSE), as defined in Equations (5.1) and (5.2), respectively [43].

4.2 Natural Language Processing and Deep Learning models for time series forecasting (Phase two)

The second phase of our computational methodology involves the use of more complex algorithms and processes that are not feasible to implement within the Dataiku platform. This includes, for example, the application of word embedding techniques such as Word2Vec. De-

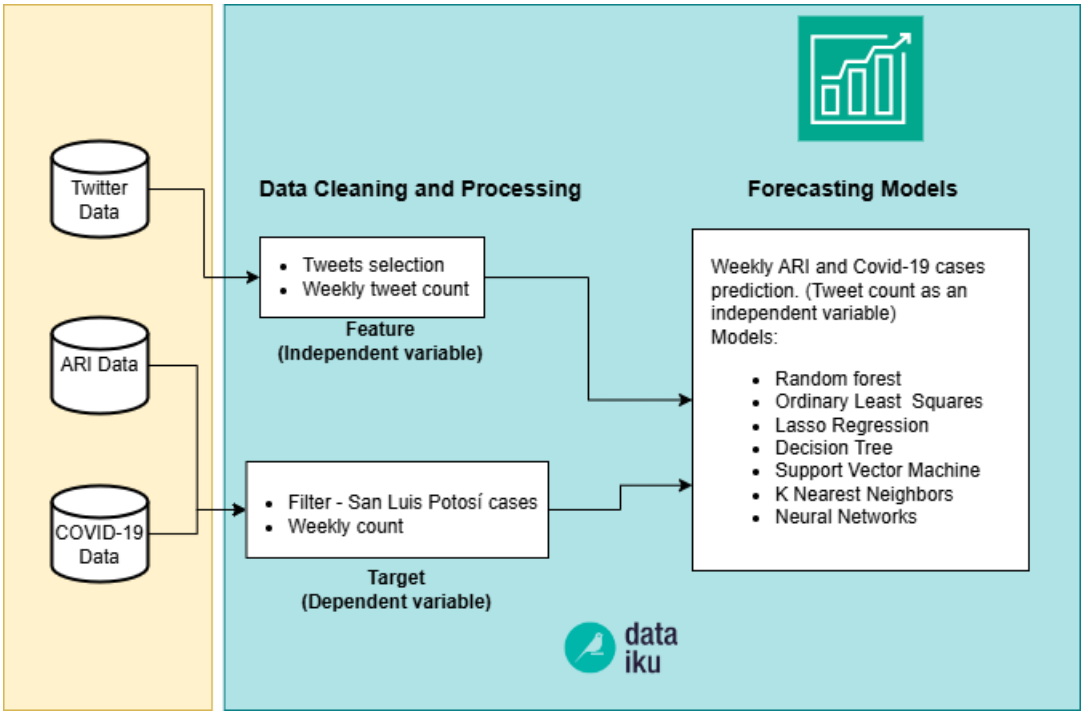

Figure 4.3 Pipeline of Phase 1 [43].

spite the increased complexity, this phase prioritizes maintaining agility and efficiency in executing the required experiments [43].

Three pipelines were developed to generate the datasets used in the prediction models (see Figure 4.4).

1. The DEP-IND-C pipeline incorporated Natural Language Processing (NLP) techniques, implemented in Python, to filter out tweets unrelated to COVID-19 or ARI. This process began with three main steps:

   (a) Stopword removal: The NLTK library was used to identify and eliminate Spanish stopwords.

   (b) Tokenization and vector transformation: Tweets were tokenized and transformed into word vectors using the NumPy library.

   (c) Conversion to numerical vectors: The Gensim library and the Word2Vec algorithm were employed to convert words into numeric vectors [109]. Word2Vec offers two training approaches—Continuous Bag of Words (CBOW) and Skip-gram—which produce 100-dimensional vector representations for each word. Each tweet was

then represented by the average of its constituent word vectors, enabling the use of textual data as numerical input for machine learning models [43].

In the second part of this pipeline, we applied the KMeans algorithm to cluster the tweets. Dataiku offers several clustering models, including KMeans, Gaussian Mixture, Mini-Batch KMeans, Agglomerative Clustering, Spectral Clustering, DBSCAN, Interactive Clustering, and Isolation Forest. KMeans was selected due to its simplicity, lower computational cost, and superior performance based on the Silhouette coefficient. This clustering step was essential for distinguishing tweets genuinely related to COVID-19 or ARI, as some posts—often from institutions or individuals—were unrelated to respiratory infectious diseases and were merely informational or general updates. The KMeans algorithm incorporates silhouette analysis to evaluate clustering quality. The silhouette coefficient measures how similar a data point is to its own cluster (cohesion) relative to other clusters (separation). This metric ranges from -1 to 1, with values closer to 1 indicating that data points are well matched to their own cluster and poorly matched to neighboring ones, suggesting compact and well-separated clusters [43].

2. DEP-IND. This pipeline used the raw tweet dataset without applying any additional filtering beyond the criteria set during data collection via the Twitter API. A basic data cleaning process was performed, followed by the computation of the weekly tweet count (see §3.2).

3. DEP. This pipeline was designed to assess the contribution of the “tweets" feature—used as an independent variable in the other two pipelines. In this case, only the historical weekly counts of confirmed COVID-19 and ARI cases were used for prediction, excluding any tweet-related data.

The final step in each of the three pipelines involved applying machine learning models for time series forecasting. Dataiku offers a range of forecasting models, including statistical approaches—such as Trivial Identity, Seasonal Naive, AutoARIMA, Seasonal Trend, and Non-Parametric Time Series—as well as deep learning models, including Feed Forward [110], DeepAR [34], and Transformer [35]. Statistical models were excluded due to their comparatively lower accuracy. Instead, deep learning models—specifically DeepAR, Feed Forward, and Transformer—were employed to predict the target variables (COVID-19 or ARI cases). Notably, these models have not been commonly used in prior research involving Twitter data. Model training and testing were conducted within the Dataiku platform. To evaluate and compare

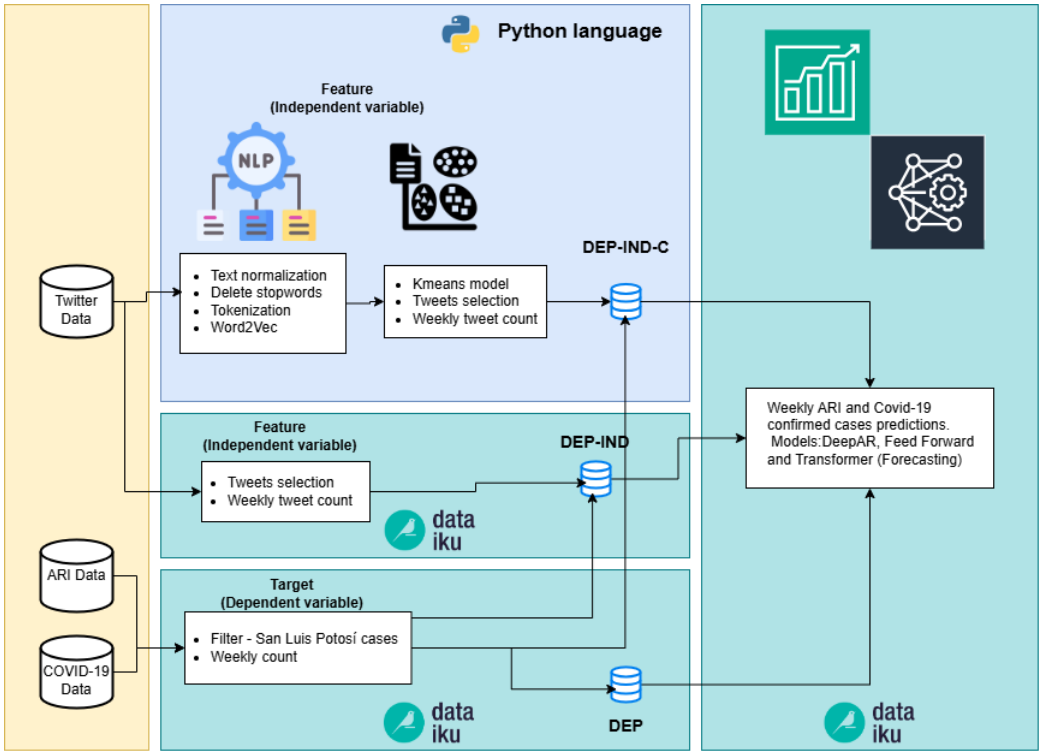

Figure 4.4 Pipelines of Phase two (DEP-IND-C, DEP-IND and DEP) [43].

performance across models, the MAPE was used as the primary metric (see Equation 5.1). The time series data spanned from January 2020 to December 2022 (formatted as yyyy-mm) [43].

4.3 DeepAR model (Phase three)

The third phase of the computational methodology was implemented using the Python programming language. The primary objective of this stage is to refine and enhance processes that could not be fully integrated within the Dataiku platform. In this phase, four distinct pipelines—labeled Category 1 through Category 4—were developed, each representing a unique combination of preprocessing and modeling procedures. These pipelines produced four separate datasets, which were subsequently used in the prediction models (see Figure 4.5) [43].

1. Category 1. This pipeline focuses on cleaning and processing the target variable (historical weekly counts of confirmed COVID-19 and ARI cases). It mirrors the DEP pipeline from Phase Two but is reimplemented entirely in Python. The predictions generated from this

pipeline serve as a baseline for evaluating the performance of the other pipelines that incorporate tweet counts as input features.

2. Category 2. This pipeline utilizes the tweet dataset without applying additional filtering, aside from the criteria used during data collection via the Twitter API. A basic data cleaning process was performed—similar to the DEP-IND pipeline from Phase Two—and the weekly tweet count was calculated.

3. Category 3. This pipeline is the same one used in DEP-IND-C in the previous phase.

4. Category 4. This pipeline used a tweet classifier through BERT algorithm [36]; labeling tweets related to users who are related to COVID-19 or ARI. The advantage of BERT over other word embedding methods like Word2Vec is its ability to generate vector representations for words based on the context of the entire sentencee [40]. The BERT version used is “bert-base-multilingual-cased", since this can be used for the Spanish language. For example, other researchers used IndoBERT model for Indonesian language Kalanjati et al. [24]. Moreover, elimination of stopwords and tokenization were part of the cleaning process of this pipeline.

The final step of each pipeline involved training and testing the DeepAR model, a forecasting method based on autoregressive recurrent neural networks. This model is based on LSTM, which handle sequential data efficiently [34]. It was implemented using the Python GluonTS library [111] to forecast ARI and COVID-19 time series.

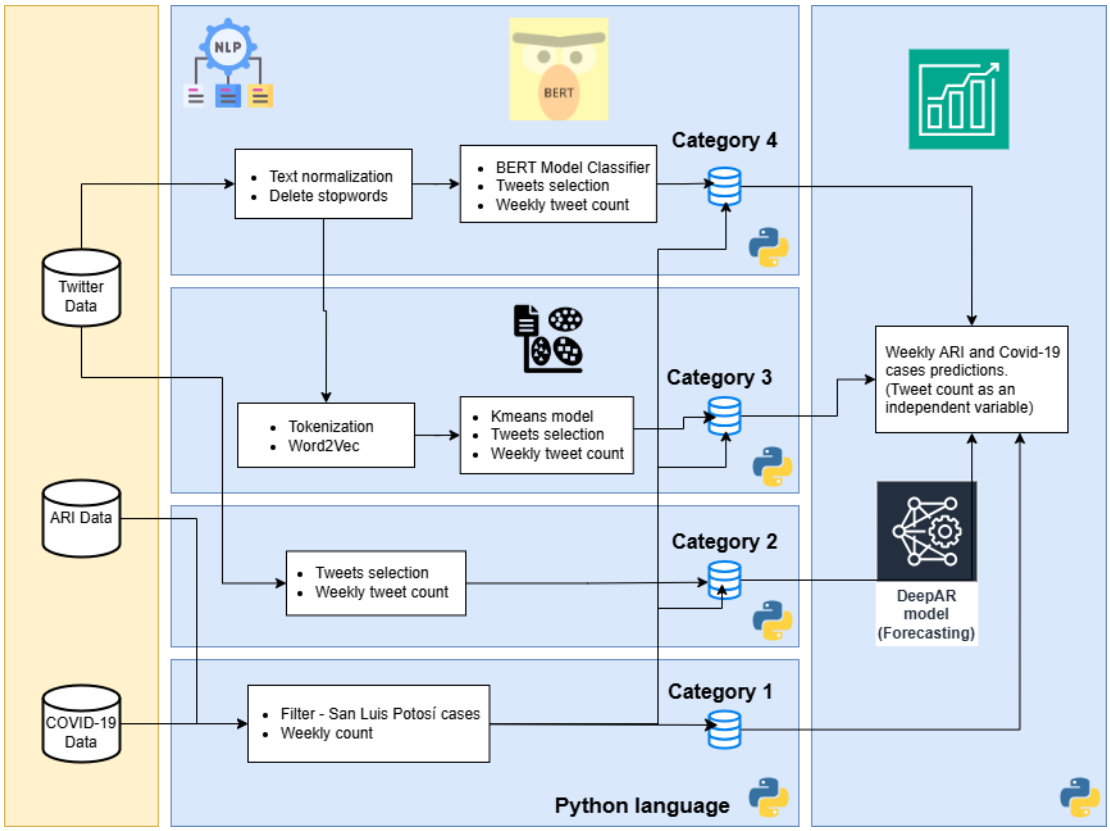

Figure 4.5 Pipelines of Phase three (Categories: 1, 2, 3 and 4) [43].

Chapter 5

Results

This section contains the different results obtained, each phase was developed taking into consideration the findings of the previous phase. The first phase helped us compare the use of models found in recent literature related to tweet analysis with machine learning. Popular models such as Support Vector Machine and Decision Tree were used. However, comparison with previous research is a complicated task; because the circumstances, parameters and data for each research project were too diverse, many of them in completely different contexts. The results of these tests were not as expected. In the second phase, the use of deep learning algorithms was further explored for predictions as well as the use of natural language processing algorithms to select tweets. In this phase, outstanding performance of the DeepAR (deep learning) model was found, and it was used in the last phase , through the Python GluonTS library. Additionally, in this phase three, the BERT algorithm was implemented for tweet classification. The results of each phase are explained below.

5.1 Maximum absolute percent error (MAPE) and Root mean squared error (RMSE)

To compare the results across the three phases and their respective forecasting models, two evaluation metrics were used: Maximum Absolute Percent Error (MAPE) 5.1 and Root Mean Squared Error (RMSE) 5.2 . MAPE indicates the largest percentage difference between predicted and observed values, while RMSE measures the average magnitude of these differences. In both cases, values closer to zero indicate a stronger agreement between predictions and actual

observations [32, 43].

$$MAPE = (\max_i \frac{|Y_i - \hat{Y}_i|}{Y_i}) * 100 \tag{5.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{5.2}$$

5.2 Dataiku machine learning models results (Phase one)

The results from Phase One correspond to predictions made on 20% of the total registered weeks for COVID-19 and ARI cases, using an 80/20 split for training and testing, respectively. These results are presented in Table 5.1. On average, predictions for ARI cases were more accurate than those for COVID-19. The models with the lowest error rates were Support Vector Machines, Neural Networks, and Ordinary Least Squares. However, the error margins remained relatively high (see Table 5.1), which prompted the development of a second phase focused on time series models within Dataiku. Unlike the models in Phase One, where predictions depended solely on the independent variable (i.e., tweet counts), time series models incorporate the historical behavior of the target variable itself. This integration of temporal patterns significantly improves prediction accuracy [43].

Table 5.1 Phase 1, models performance (MAPE and RMSE) [43].

Model	MAPE-ARI	RMSE-ARI	MAPE-COVID	RMSE-COVID
Random forest	0.774	5555	2.36	1909
Ordinary Least Squares	0.616	**4798**	9.09	1256
Lasso Regression	0.622	4852	8.08	1215
Decision Tree	0.91	6087	2.05	2107
Support Vector Machine	**0.475**	5076	3.23	**771**
K Nearest Neighbors	0.645	5027	2.73	2595
Neural Networks	1.0	8126	**0.993**	1065

* In bold fonts are highlighted the best results, the lower the better.

5.3 Deep learning models comparison (Phase two)

For this phase the results are detailed for: 1) tweet selection and 2) time series forecasting.

5.3.1 Tweet selection

In this phase, an additional step was introduced to select tweets based on their content. Beyond the initial filtering applied during data collection via the Twitter API, a word embedding technique—Word2Vec—was used to convert words into numerical vectors. These vectors were then used as input for a clustering model to group and classify semantically similar tweets. The KMeans algorithm was selected for clustering, and multiple tests were conducted by varying the number of clusters k from 3 to 7. The performance of each configuration was evaluated using the Silhouette metric. The results of these tests are presented in Table 5.2 [43].

Table 5.2 Results from KMeans algorithm for different number of clusters k, the values correspond to the silhouette metric (the higher, the better)[43].

Model	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**
KMeans	0.488	0.402	0.364	0.366	0.321

Although the silhouette score is a key metric for evaluating clustering performance, it is not the sole criterion. A high silhouette value can still occur even when the majority of data points are grouped into a single cluster, leading to poor distribution. To address this, the KMeans model with k = 4 was selected over k = 3, as it provided a better balance between silhouette score and data distribution. The inclusion of a clustering algorithm in the tweet selection process aimed to identify and exclude tweets that were not relevant. As an unsupervised method, clustering avoids the need for manual labeling by isolating specific clusters for inclusion or exclusion. Two clusters—Cluster 1 and Cluster Outliers—were identified as containing tweets that did not originate from individuals or were unrelated to COVID-19 or ARI. These included content from government institutions, unrelated topics, or non-interpretable tweets. These clusters were removed from the dataset used for prediction, resulting in a refined subset of 5,317 tweets. Examples of the excluded clusters are shown below [43].

1. Cluster 1, examples:

   (a) “status covid19 at san luis potosi. record infections, deaths. deaths: 143, infections, last [estatus covid19 san luis potosí. récord contagios fallecimientos. defunciones 143 contagios úl]".

   (b) “status covid19 at san luis potosi sum up 2,068 infections and 228 deaths. [estatus covid19 san luis potosí suman 2,068 contagios 128 decesos.] ".

(c) “status covid19 at san luis potosi 168 new infections deaths in last hours [estatus covid19 san luis potosi 168 nuevos contagios muertes últimas horas]".

2. Cluster Outliers, examples:

   (a) “@drmanuelortho @bulletgabo @salazartrejo @dimensionamural @rouge7475 @jhonlinx @random3e @lalizlizliz. . . ".

   (b) “@aaptmx inscripciones abiertas".

   (c) “@luisoncastaneda @gervivamexico @eddcampe @lopezobrador @claudiashein @mebrard aclaren medicinas".

5.3.2 Time series forecasting

Tables 5.3 and 5.4 present a summary of the results, with Table 5.3 corresponding to ARI data and Table 5.4 to COVID-19 data. Each table contains six columns showing the average MAPE values, as each prediction model was executed fifteen times to ensure consistency. The DEP column represents results from the pipeline using only the dependent variable. The DEP-IND column includes results from the pipeline that incorporates the weekly tweet count as an independent variable. The final column, labeled VAR (or VAR-C), shows the difference between the DEP and DEP-IND results (or DEP-C and DEP-IND-C for pipelines with tweets content filtering). A positive VAR value indicates that adding the tweet feature improved the model’s accuracy, while a negative value suggests a decline in performance. For the COVID-19 time series, the inclusion of tweets as a feature had minimal to no positive impact, as reflected by VAR values near zero across most models. In contrast, for the ARI time series, the DeepAR model demonstrated significant improvements: an 18.56% reduction in MAPE in the DEP-IND pipeline and a 16.74% improvement in the DEP-IND-C pipeline, indicating that the tweet feature contributed meaningfully to forecasting ARI cases [43].

Table 5.3 ARI Forecasting results [43].

MODEL	**DEP**	**DEP-IND**	**VAR**	**DEP-C**	**DEP-IND-C**	**VAR-C**
DeepAR	0.3673	0.1817	**0.1856**	0.3519	0.1845	**0.1674**
Feed Forward	0.1865	0.1851	0.0014	0.2037	0.2025	0.0013
Transformer	0.2562	0.1960	0.0602	0.2381	0.2256	0.0125

Table 5.4 COVID-19 Forecasting results [43].

MODEL	DEP	DEP-IND	VAR	DEP-C	DEP-IND-C	VAR-C
DeepAR	0.3183	0.4165	-0.0982	0.3085	0.4018	-0.0933
Feed Forward	0.5301	0.5282	**0.0019**	0.5288	0.5192	**0.0096**
Transformer	0.3412	0.4529	-0.1118	0.3199	0.4508	-0.1308

5.4 DeepAR model results (Phase three)

The results of the four categories into which this phase is divided are described below. Figures 5.1 and 5.2 depict the predictions for the four categories with respect to confirmed ARI cases. Figures 5.3 and 5.4 depict the confirmed cases of COVID-19. Category 1 refers to forecasts made using the historical data of the dependent variable (COVID-19 cases or ARI). On the other categories, the tweet count is added to the prediction calculations; varying the tweet filtering procedure depending on the category. Therefore, the independent variable “tweets" is expected to improve the prediction, otherwise it is not necessary. Figures 5.1–5.4 reveal a slight lag in the predictions, which is typical of time series prediction models. This occurs because the model learns the representative patterns of the time series and replicates them in its predictions, and occasionally with some delay [43].

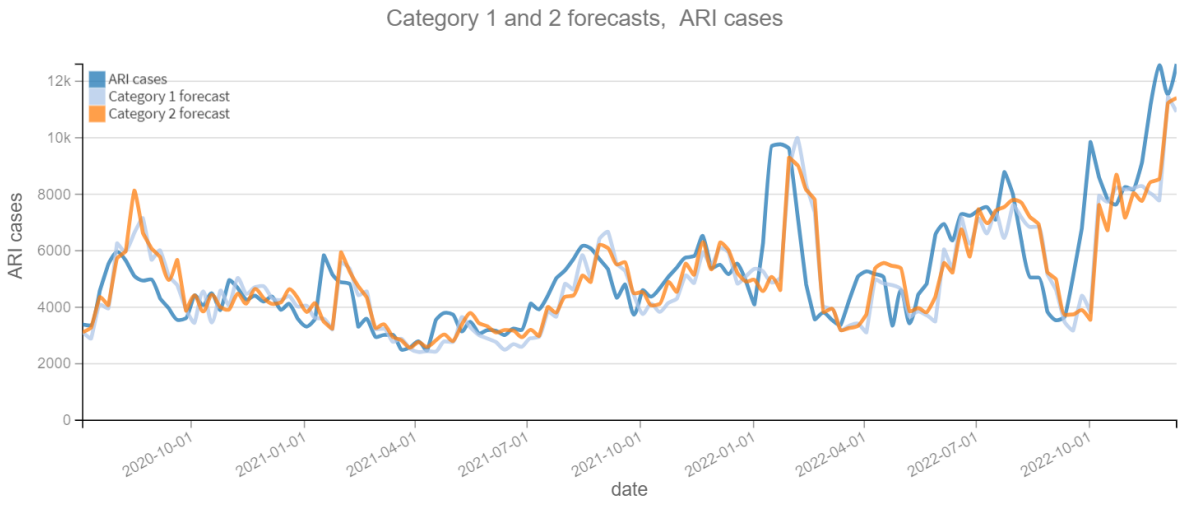

Figure 5.1 Positive ARI cases per week vs Forecast, Category 1 and 2 (2020 - 2022). Category 1: target variable without features. Category 2: target and tweets feature without filter [43].

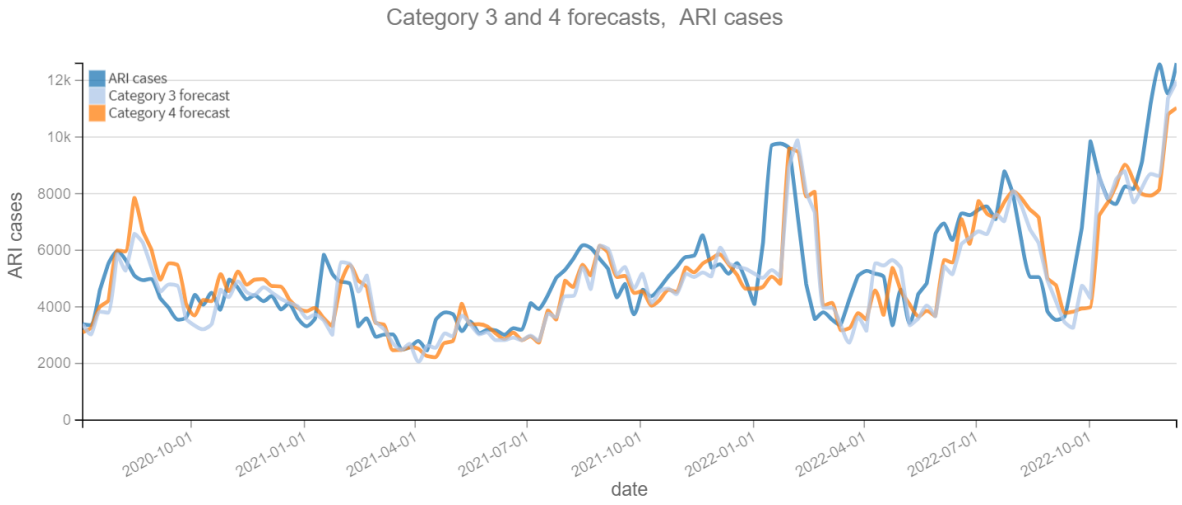

Figure 5.2 Positive ARI cases per week vs Forecast, Category 3 and 4 (2020 - 2022). Category 3: target and tweets feature with Kmeans filter. Category 4: target and tweets feature with BERT filter [43].

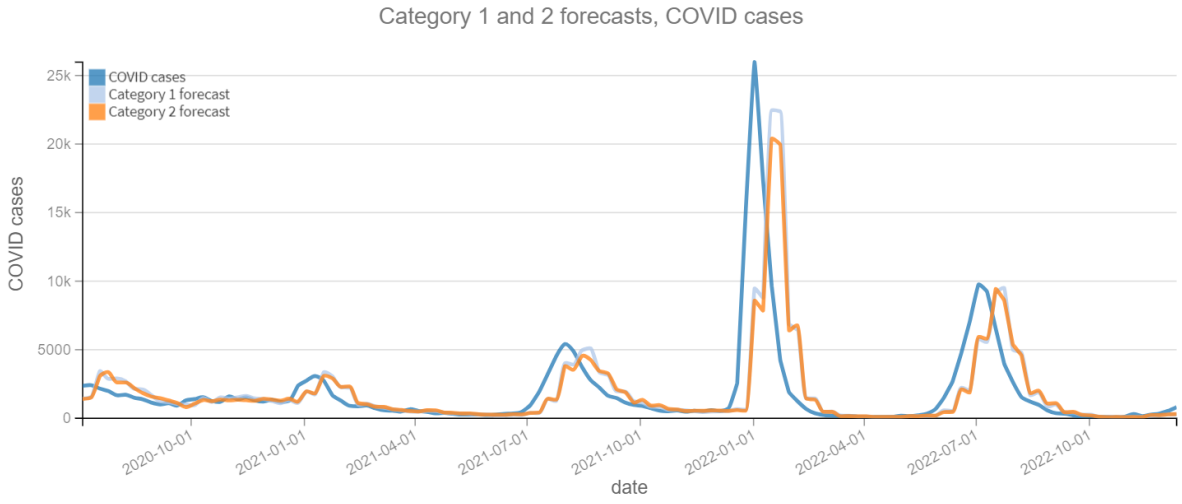

Figure 5.3 Positive COVID cases per week vs Forecast, Category 1 and 2 (2020 - 2022). Category 1: target variable without features. Category 2: target and tweets feature without filter [43].

Figure 5.6 shows the MAPE metric for test periods of 4 weeks. In the forecast of the variable confirmed COVID-19 cases, it can be noted that the four categories of models have the same behavior in this metric, which implies that the variable Independent “tweets" does not contribute enough to improve the model. Contrary, Figure 5.5 shows the MAPE on the prediction of the

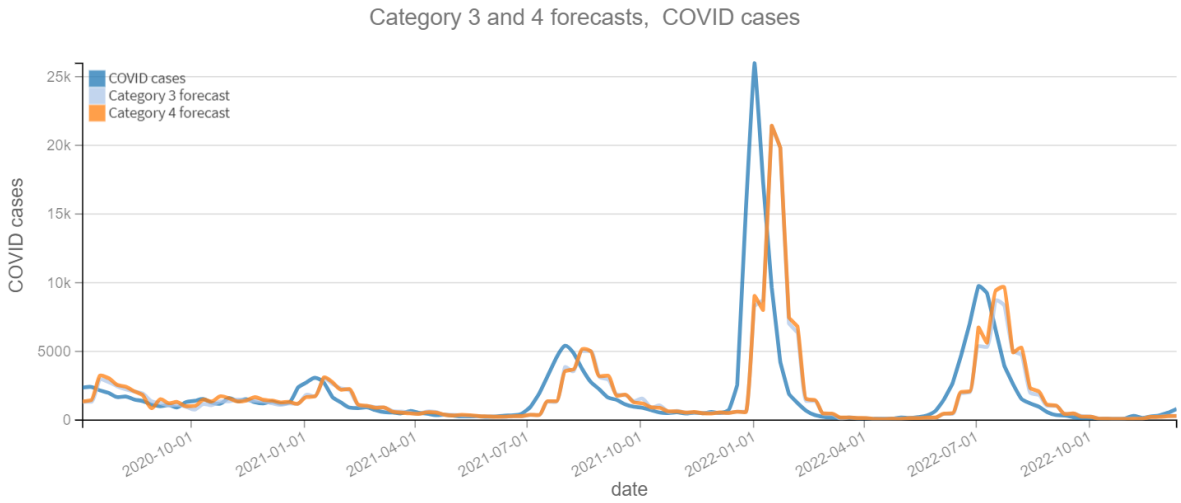

Figure 5.4 Positive COVID cases per week vs Forecast, Category 3 and 4 (2020 - 2022). Category 3: target and tweets feature with Kmeans filter. Category 4: target and tweets feature with BERT filter [43].

dependent variable confirmed cases of ARI, where a change is observed in the patterns of each model category. For example, for the period close to 2021-01-01 it is observed that the Category 2 model clearly improves the prediction by having a lower MAPE. In both plots, the dependent variable (confirmed cases of ARI and COVID-19) is also shown to assess whether there is a correlation between the slope changes of the dependent variable and the MAPE prediction error [43].

Table 5.5 contains the MAPE and RMSE values, over the entire prediction period (2020-06-28 to 2022-12-25). For the dependent variable ARI confirmed cases, the Category 3 shows the best performance, with the lowest MAPE (0.1746) and RMSE (1346.50). And for the variable COVID-19 confirmed cases, the best performance corresponds to the Category 2, with the lowest MAPE (0.5419) and RMSE (3048.52). It can be noted that the variation of the MAPE and RMSE between the different models is not substantial in the dependent variable COVID-19 confirmed cases, while for the variable ARI confirmed cases it is only $\pm 1\%$. This can be interpreted that there is a little contribution of the “tweets" variable to improve the prediction of the dependent variables [43].

Table 5.5 MAPE and RMSE of each forecast category [43].

Period	Variable	MAPE	RMSE	Test Category
2020-06-28 / 2022-12-25	ARI Cases	0.1847	1435.66	1
2020-06-28 / 2022-12-25	ARI Cases	0.1775	1431.53	2
2020-06-28 / 2022-12-25	ARI Cases	**0.1746**	**1346.50**	3
2020-06-28 / 2022-12-25	ARI Cases	0.1854	1438.21	4
2020-06-28 / 2022-12-25	COVID Cases	0.5437	3165.10	1
2020-06-28 / 2022-12-25	COVID Cases	**0.5419**	**3048.52**	2
2020-06-28 / 2022-12-25	COVID Cases	0.5442	3064.66	3
2020-06-28 / 2022-12-25	COVID Cases	0.5671	3077.47	4

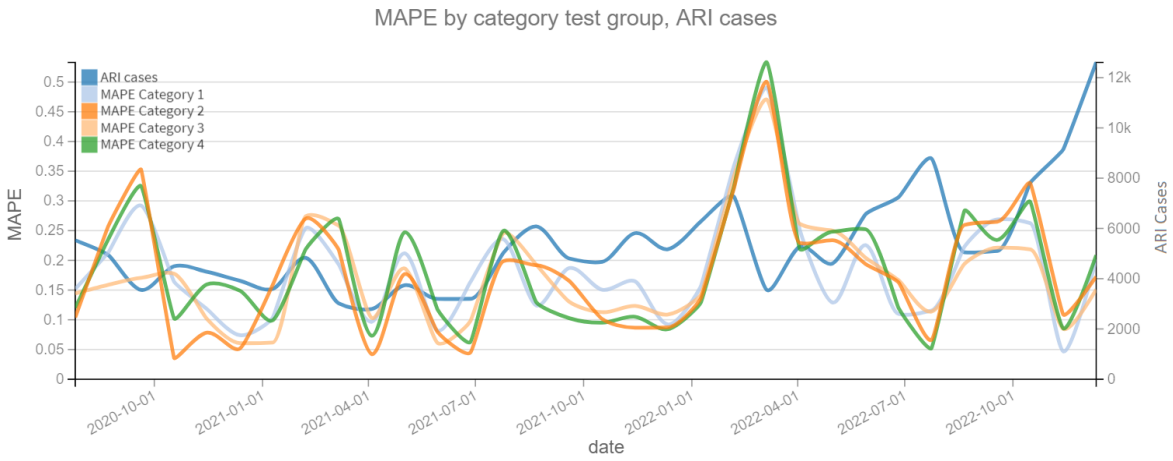

Figure 5.5 MAPE results, ARI cases (2020 - 2022). Category 1. Target variable without features. Category 2. Target and tweets feature without filter. Category 3. Target and tweets feature with Kmeans filter. Category 4. Target and tweets feature with BERT filter [43].

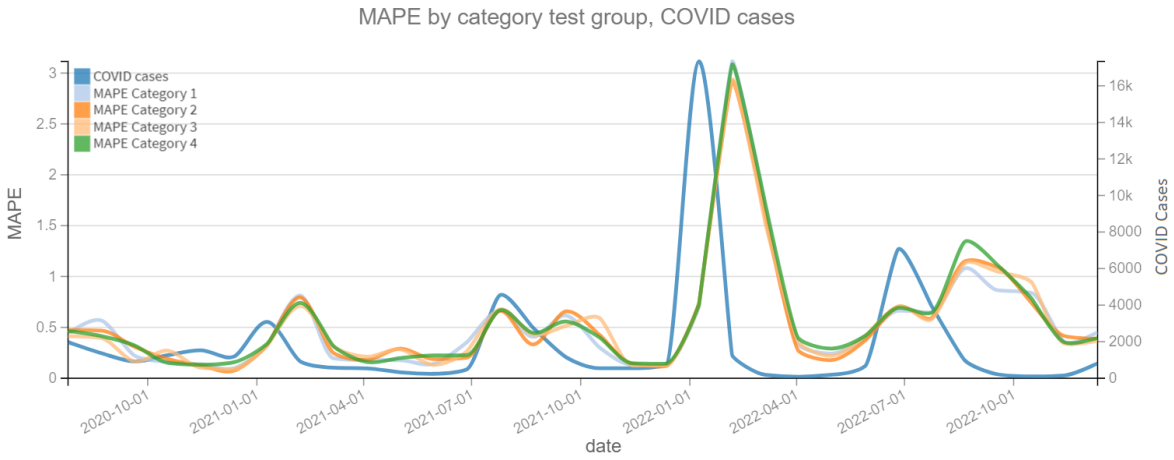

Figure 5.6 MAPE results, COVID cases (2020 - 2022). Category 1. Target variable without features. Category 2. Target and tweets feature without filter. Category 3. Target and tweets feature with Kmeans filter. Category 4. Target and tweets feature with BERT filter [43].

Table 5.6 MAPE difference between Category 1 and Categories 2, 3, 4 [43].

Year	Variable	Category 2	Category 3	Category 4
2020	ARI Cases	0.0222	**0.0333**	-0.0129
2020	COVID Cases	-0.0005	**0.0325**	-0.0019
2021	ARI Cases	**0.0182**	0.0078	0.0157
2021	COVID Cases	-0.0106	-0.0210	-0.0105
2022	ARI Cases	-0.0107	**0.0015**	-0.0114
2022	COVID Cases	**0.0152**	0.0046	-0.0461

Table 5.6 shows the difference between the MAPE of the Category 1 model predictions (dependent variable only) and the other categories. The objective is to observe the contribution of the “tweets" variable to the accuracy of the predictions. If the result is positive, it means that the model with the “tweets" variable has a lower MAPE and therefore the prediction is more accurate. Otherwise, the “tweets" variable worsens the prediction. In 2020, the Category 3 model has the better performance of 0.0333 for the prediction of the ARI variable and 0.0325 for the COVID-19 variable, as a percentage it means approximately 3% improvement over the Category 1 model. For subsequent years it is not observed a significant improvement, so it can

be concluded that in the first year of the pandemic, reading the tweets did represent an aid in the prediction. Something that did not happen in the following years [43].

Comparing Figure 5.2 and Figure 5.3, we can see that the scale in the number of cases is different for ARI and COVID-19. Therefore, the datasets were normalized and the experiments repeated to see the effect in the data scale. We use StandardScaler method of the Python Scikit-learn library[1], which removes the mean and then scales to unit variance. The results are in Table 5.7, and they must be compared with Table 5.5. For all tests there is no improvement except in Category 2 (COVID-19), which is not representative, only 0.0069 of difference in MAPE [43].

Table 5.7 MAPE and RMSE of each forecast category, standardized features [43].

Period	**Variable**	**MAPE**	**RMSE**	**Category**
2020-06-28 / 2022-12-25	ARI Cases	**0.2144**	1746.69	1
2020-06-28 / 2022-12-25	ARI Cases	0.2211	1784.18	2
2020-06-28 / 2022-12-25	ARI Cases	0.2163	**1661.63**	3
2020-06-28 / 2022-12-25	ARI Cases	0.2156	1752.79	4
2020-06-28 / 2022-12-25	COVID Cases	0.6538	3529.61	1
2020-06-28 / 2022-12-25	COVID Cases	**0.5350**	3424.18	2
2020-06-28 / 2022-12-25	COVID Cases	0.7721	3288.10	3
2020-06-28 / 2022-12-25	COVID Cases	0.6422	**3266.63**	4

[1]https://scikit-learn.org/

Chapter 6

Conclusions

We propose a computational methodology for predicting acute respiratory infections (ARI) using social media data from Twitter and ML algorithms, structured into three distinct phases. This phased approach enables a comparative analysis of various ML models and their effectiveness in processing tweet data as predictive input for respiratory disease trends. While the primary goal was to establish a foundational framework for tweet-based analysis that could be extended to other contexts and applications, the results also reveal key challenges and limitations in using tweets as predictive variables.Findings from Phase One indicate that tweet volume alone, used as an independent variable, is insufficient for accurately predicting respiratory disease cases. The best outcome achieved was a MAPE of 0.475 using Support Vector Machines for ARI cases. In contrast, Phases Two and Three, which employ time series models incorporating the historical behavior of the dependent variable, yield significantly better results. Key observations from Phase Two include: a) For COVID-19 predictions, in two out of three models, the inclusion of tweet data did not improve performance and in some cases worsened it. b) The only model that showed improvement with tweet data was Feed Forward (see Table 5.4). c) Tweet count had a noticeably stronger impact on ARI forecasts than on COVID-19. d) Among all models tested, DeepAR produced the best results for ARI time series predictions. e) The additional step of clustering and filtering tweets did not yield any measurable improvement in predictive accuracy during this phase [43].

Phase Three centers on analyzing the DeepAR model using different preprocessing approaches for Twitter data, specifically the pipelines labeled Categories 2 to 4 (see Figure 4.5). Over a three-year period, we found that the “tweets” variable was predictive only for confirmed ARI cases. The best-performing model was associated with the Category 3 pipeline, which incorporates Word Embedding and KMeans clustering, and relates to ARI case prediction (see

Tables 5.6 and 5.5). This highlights that applying NLP techniques and tweet selection improves prediction accuracy, though the improvement was modest—approximately 1.01% better than predictions without the tweets variable. Notably, during the second half of 2020, including the tweet variable enhanced prediction accuracy by around 3% for both ARI and COVID-19 confirmed cases. This suggests that Twitter data was more representative and informative at the pandemic's onset but diminished in relevance over time. Due to differences in regions, timeframes, and diseases studied, direct comparisons with other research are limited. Nonetheless, this study offers several valuable contributions. The comprehensive analysis of Twitter, ARI, and COVID-19 data helped identify the most effective preprocessing methods for Twitter data and confirmed DeepAR as the best predictive model. Given the challenges inherent in text mining of social media data, our findings provide useful insights for researchers working with similar datasets. Furthermore, the computational methodology presented here is adaptable to other data sources and domains. It offers a framework for identifying optimal machine learning models combined with Twitter data for time series prediction and can be extended to fields such as finance, energy, and beyond, wherever time series data plays a central role [43].

6.1 Contributions

The contributions of this thesis include the systematic literature review, the proposed computational methodology, and the creation and results of the computational model. The main contributions are listed below:

1. We developed an Systematic Review based on the PRISMA methodology to analyze the available studies published between January 2006 and December 2020. Unlike other Systematic Reviews, we implement a classification of the studies according to the algorithms used in mathematical models, either based on the objective they want to achieve (classification, regression or clustering) or on the complexity of the models according to the analysis of information from social networks. We analyzed the algorithms used based on the problem they solve and we also analyzed the metrics used, comparing them with other research. Additional, this review focuses on studies related to Twitter posts on ARI.

2. We introduce a computational methodology to predict acute respiratory infections with social media (Twitter) and ML algorithms, which is composed of three phases. The

different phases allow us to compare various ML models applied to the processing of tweets and their subsequent use as a predictors in respiratory diseases.

3. Application of Natural Language Processing (NLP) techniques as a process for analyzing and selecting tweets that correspond to users who have or had ARI symptoms. The algorithms used for this task are: Word Embedding Based Clustering Classification to create a cluster of tweets related to a particular topic, and a Transformer-based machine learning model as a second step to classify tweets in those that match the aforementioned requirements.

4. We create a computational model that process the two different data sources (tweets and publicly available surveillance databases, both related to respiratory infectious diseases) and predicts respiratory infectious disease outbreaks.

6.2 Limitations

Using social network data for research presents several limitations. One of the main challenges we encountered was the restricted access to tweets, as the Twitter API account employed had a download limit of 5,000 tweets per month. Moreover, the collected tweets may not accurately reflect the target population of San Luis Potosí, Mexico, in terms of geographic coverage and demographic representation. The API queries relied on seven keywords related to ARI and COVID-19, which limits the scope of captured content. This approach may overlook relevant tweets that use alternative terms or local expressions to refer to these illnesses. Furthermore, the vocabulary used on social media to describe the same condition can evolve over time, adding another layer of complexity to data collection [43]. Tweet collection was limited to Spanish-language posts, and appropriate libraries were configured for text cleaning and processing in this language. While the analysis of tweets in other languages is feasible by adjusting the parameters, it falls outside the scope of this study. Natural Language Processing techniques were applied in conjunction with BERT (for classification) and K-means (for clustering) to identify tweets from individuals who have experienced or are currently experiencing respiratory illnesses such as ARI or COVID-19. Nevertheless, the presence of false positives among the selected tweets remains a concern. Enhancing the accuracy of tweet classification would require a deeper contextual analysis using more advanced NLP methods. The selection of machine learning models in Phases 1 and 2 was constrained to those available within the Dataiku platform. Although Dataiku provides a variety of clustering and time series forecasting

tools, the inability to compare performance with external models represents a methodological limitation. Additionally, the preprocessing step that removed emoticons, non-alphanumeric symbols, and pictorial characters may have resulted in the loss of contextual cues, potentially affecting the precision of tweet classification [43].

6.3 Future research

The Twitter data used in this study is limited to the state of San Luis Potosí, Mexico. Expanding the analysis to include other cities in Mexico or additional countries could yield more generalizable insights into the potential of tweets as a predictive variable for respiratory diseases. Moreover, extending the time series beyond the 2020–2022 period may produce different results, particularly by capturing patterns unrelated to the recent pandemic outbreak. Analyzing tweets in multiple languages presents a promising avenue for comparative studies across regions or within multilingual populations. However, a major challenge for multilingual analysis lies in the availability and quality of text processing tools, such as NLTK, for the target language. In addition, disease-related vocabulary is broad and continuously evolving. Exploring patterns in word usage, expressions, and regional idioms could enhance classification accuracy, but such analysis would require access to a larger volume of tweets. It is also important to note that researchers working with Twitter (now X) must consider recent changes in the platform's data access policies, which impose stricter limitations on data downloads. As a result, alternative social media platforms should be considered as supplementary data sources, given the added predictive value of social media data. Tweet classification remains a complex task that often requires advanced NLP techniques. Models like Word Embeddings and BERT typically need large datasets to perform effectively, highlighting the ongoing challenge of developing models that work well with limited data. Finally, this study focused solely on textual content, despite the fact that modern tweets frequently include images, emojis, and other visual elements. Integrating these modalities through multimodal analysis represents a valuable opportunity to enhance tweet interpretation and improve classification accuracy [43].

Bibliography

[1] Patrick Pilipiec, Isak Samsten, and András Bóta. Surveillance of communicable diseases using social media: A systematic review. *PloS one*, 18:e0282101, 02 2023. doi: 10.1371/journal.pone.0282101.

[2] Marta Nunes, Edward Thommes, Holger Fröhlich, Antoine Flahault, Julien Arino, Marc Baguelin, Matthew Biggerstaff, Gaston Bizel-Bizellot, Rebecca Borchering, Giacomo Cacciapaglia, Simon Cauchemez, Alex Barbier—Chebbah, Carsten Claussen, Christine Choirat, Monica Cojocaru, Catherine Commaille-Chapus, Chitin Hon, Jude Kong, Nicolas Lambert, and Laurent Coudeville. Redefining pandemic preparedness: Multidisciplinary insights from the cerp modelling workshop in infectious diseases, workshop report. *Infectious Disease Modelling*, 9(2):501–518, 02 2024. doi: 10.1016/j.idm.2024.02.008.

[3] E. Velasco, T. Agheneza, K. Denecke, G. Kirchner, and T. Eckmanns. Social media and internet-based data in global systems for public health surveillance: a systematic review. *The Milbank quarterly*, 92(1):7–33, 2014.

[4] World Health Organization. Regional Office for the Western Pacific. *A guide to establishing event-based surveillance*. WHO Regional Office for the Western Pacific, 2008.

[5] Martin McKee, Rikard Rosenbacke, and David Stuckler. The power of artificial intelligence for managing pandemics: A primer for public health professionals. *The International Journal of Health Planning and Management*, 40(1):257–270, 10 2024. doi: 10.1002/hpm.3864.

[6] J. Liang, R. Liu, W. He, Z. Zeng, Y. Wang, B. Wang, L. Liang, T. Zhang, C. L. P. Chen, C. Chang, C. Hon, E. H. Y. Lau, Z. Yang, and K. Tong. Infection rates of 70*Journal of Infection*, 86:402–404, 2023. doi: 10.1016/j.jinf.2023.01.029.

[7] Deema Fallatah and Hafeez Adekola. Digital epidemiology: Harnessing big data for early detection and monitoring of viral outbreaks. *Infection Prevention in Practice*, 6(3): 1–7, 06 2024. doi: 10.1016/j.infpip.2024.100382.

[8] Lauren Charles, Tera Reynolds, Mark Cameron, Mike Conway, Eric Lau, Jennifer Olsen, Julie Pavlin, Mika Shigematsu, Laura Streichert, Katie Suda, and Courtney Corley. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PLoS ONE*, 10, 10 2015. doi: 10.1371/journal.pone.0139701.

[9] Aakansha Gupta and Rahul Katarya. Social media based surveillance systems for healthcare using machine learning: A systematic review. *Journal of Biomedical Informatics*, 108:103500, 07 2020. doi: 10.1016/j.jbi.2020.103500.

[10] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, 240:112552, 09 2019. doi: 10.1016/j.socscimed.2019.112552.

[11] Monica Giancotti, Milena Lopreite, Marianna Mauro, and Michelangelo Puliga. Innovating health prevention models in detecting infectious disease outbreaks through social media data: an umbrella review of the evidence. *Frontiers in Public Health*, 12(1):1–20, 11 2024. doi: 10.3389/fpubh.2024.1435724.

[12] Xiangfeng Dai, M. Bikdash, and B. Meyer. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *SoutheastCon 2017*, pages 1–7, 03 2017. doi: 10.1109/SECON.2017.7925400.

[13] Karolos Talvis, Kostantinos Chorianopoulos, and Katia Kermanidis. Real-time monitoring of flu epidemics through linguistic and statistical analysis of twitter. In *Proceedings - 9th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2014*, pages 83–87, 11 2014. doi: 10.1109/SMAP.2014.38.

[14] Hideo Hirose and Liangliang Wang. Prediction of infectious disease spread using twitter: A case of influenza. *2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming*, pages 100–105, 2012.

[15] Ruchit Nagar, Qingyu Yuan, Clark Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and John Brownstein. A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *Journal of medical Internet research*, 16:e236, 10 2014. doi: 10.2196/jmir.3416.

[16] Guido Zuccon, Sankalp Khanna, Anthony Nguyen, Justin Boyle, Matthew Hamlet, and Mark Cameron. Automatic detection of tweets reporting cases of influenza like illnesses in australia. *Health information science and systems*, 3:S4, 04 2015. doi: 10.1186/2047-2501-3-S1-S4.

[17] José Carlos Santos and Sérgio Matos. Analysing twitter and web queries for flu trend prediction. *Theoretical biology and medical modelling*, 11:S6, 05 2014. doi: 10.1186/1742-4682-11-S1-S6.

[18] Víctor Prieto, Sérgio Matos, Manuel Alvarez, Fidel Cacheda, and José Oliveira. Twitter: A good place to detect health conditions. *PloS one*, 9:e86191, 01 2014. doi: 10.1371/journal.pone.0086191.

[19] Jianhong Jiang, Chenyan Yao, and Xinyi Song. A multidimensional comparative study of help-seeking messages on weibo under different stages of covid-19 pandemic in china. *Frontiers in Public Health*, 12(1):1–17, 02 2024. doi: 10.3389/fpubh.2024.1320146.

[20] Shweta Agrawal, Sanjiv Jain, Shruti Sharma, and Ajay Khatri. Covid-19 public opinion: A twitter healthcare data processing using machine learning methodologies. *International Journal of Environmental Research and Public Health*, 20:432, 12 2022. doi: 10.3390/ijerph20010432.

[21] Ryuichiro Ueda, Feng Han, Hongjian Zhang, Tomohiro Aoki, and Katsuhiko Ogasawara. Correction: Verification in the early stages of the covid-19 pandemic: Sentiment analysis of japanese twitter users. *JMIR infodemiology*, 4:e57880, 03 2024. doi: 10.2196/57880.

[22] Aisha Aldosery, Robert Carruthers, Karandeep Kay, Christian Cave, Paul Reynolds, and Patty Kostkova. Enhancing public health response: a framework for topics and sentiment analysis of covid-19 in the uk using twitter and the embedded topic model. *Frontiers in Public Health*, 12:1105383, 02 2024. doi: 10.3389/fpubh.2024.1105383.

[23] Yawen Guo, Jun Zhu, Yicong Huang, Lu He, Changyang He, Chen Li, and Kai Zheng. Public opinions toward covid-19 vaccine mandates: A machine learning-based analysis of u.s. tweets. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2022:502–511, 04 2023.

[24] Viskasari Kalanjati, Nurina Hasanatuludhhiyah, Annette d'Arqom, Danial Habri Arsyi, Ancah Marchianti, Azlin Muhammad, and Diana Purwitasari. Sentiment analysis of indonesian tweets on covid-19 and covid-19 vaccinations. *F1000Research*, 12:1007, 11 2023. doi: 10.12688/f1000research.130610.2.

[25] Moez Farokhnia Hamedani, Mostafa Esmaeili, Yao Sun, Ehsan Sheybani, and Giti Javidi. Paving the way for covid survivors' psychosocial rehabilitation: Mining topics, sentiments, and their trajectories over time from reddit. *Health Informatics Journal*, 30 (2):1–18, 2024. doi: 10.1177/14604582241240680. PMID: 38739488.

[26] Ireneus Kagashe, Zhijun Yan, and Imran Suheryani. Enhancing seasonal influenza surveillance: Topic analysis of widely used medicinal drugs using twitter data. *Journal of Medical Internet Research*, 19:e315, 09 2017. doi: 10.2196/jmir.7393.

[27] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[28] Sicong Kuang and Brian Davison. Learning word embeddings with chi-square weights for healthcare tweet classification. *Applied Sciences*, 7:846, 08 2017. doi: 10.3390/app7080846.

[29] Erdenebileg Batbaatar and Keun Ryu. Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. *International Journal of Environmental Research and Public Health*, 16:3628, 09 2019. doi: 10.3390/ijerph16193628.

[30] Chen Chen, Lina Zhou, Yunya Song, Qian Xu, Ping Wang, Kanlun Wang, Yaorong Ge, and Daniel Janies. Comparison of viral covid-19 sina weibo and twitter contents: a novel feature extraction and analytical workflow. *Journal of Medical Internet Research*, 23(1): e24889, 10 2021. doi: 10.2196/24889.

[31] Yosra Didi, Ahlem Walha, Mohamed Ben Halima, and Ali Wali. Covid-19 outbreak forecasting based on vaccine rates and tweets classification. *Computational Intelligence and Neuroscience*, 2022, 10 2022. doi: 10.1155/2022/4535541.

[32] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney Corley. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PLOS ONE*, 12:e0188941, 12 2017. doi: 10.1371/journal.pone.0188941.

[33] Soheila Molaei, Mohammad Khansari, Hadi Veisi, and Mostafa Salehi. Predicting the spread of influenza epidemics by analyzing twitter messages. *Health and Technology*, 9, 03 2019. doi: 10.1007/s12553-019-00309-4.

[34] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

[35] A Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[37] Tom M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[38] Oguzhan Gencoglu. Large-scale, language-agnostic discourse classification of tweets during covid-19. *Machine Learning and Knowledge Extraction*, 2:603–616, 11 2020. doi: 10.3390/make2040032.

[39] Janhavi Lande, Arti Pillay, and Rohitash Chandra. Deep learning for covid-19 topic modelling via twitter: Alpha, delta and omicron, 02 2023.

[40] Quyen To, Kien To, Van-Anh Huynh, Nhung Nguyen, Diep Ngo, Stephanie Alley, Anh Tran, Anh Tran, Ngan Thi Thanh Pham, Thanh Bui, and Corneel Vandelanotte. Applying machine learning to identify anti-vaccination tweets during the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, 18:4069, 04 2021. doi: 10.3390/ijerph18084069.

[41] Yuan-Chi Yang, Angel Xie, Sangmi Kim, Jessica Hair, Mohammed Al-Garadi, and Abeed Sarker. Automatic detection of twitter users who express chronic stress experiences via supervised machine learning and natural language processing. *CIN: Computers, Informatics, Nursing*, Publish Ahead of Print, 11 2022. doi: 10.1097/CIN.0000000000000985.

[42] Peipei Chen, Yuwei Jin, Xinfang Ma, and Yan Lin. Public perception on active aging after covid-19: an unsupervised machine learning analysis of 44,343 posts. *Frontiers in Public Health*, 12(1):1–8, 03 2024. doi: 10.3389/fpubh.2024.1329704.

[43] Jose Manuel Ramos-Varela, Juan C. Cuevas-Tello, and Daniel E. Noyola. A machine learning-based computational methodology for predicting acute respiratory infections using social media data. *Computation*, 13(4), 2025. ISSN 2079-3197. doi: 10.3390/computation13040086. URL https://www.mdpi.com/2079-3197/13/4/86.

[44] Dylan George, Wendy Taylor, Jeffrey Shaman, Caitlin Rivers, Brooke Paul, Tara O'Toole, Michael Johansson, Lynette Hirschman, Matthew Biggerstaff, Jason Asher, and Nicholas Reich. Technology to advance infectious disease forecasting for outbreak management. *Nature Communications*, 10, 12 2019. doi: 10.1038/s41467-019-11901-7.

[45] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas Altman. Moher d, liberati a, tetzlaff j, altman dg, group ppreferred reporting items for systematic reviews and meta-analyses: the prisma statement. plos med 6: e1000097. *Open medicine : a peer-reviewed, independent, open-access journal*, 3:e123–30, 07 2009. doi: 10.1016/j.jclinepi.2009.06.005.

[46] Margarida Pereira, Cosima Prahm, Jonas Kolbenschlag, Eva Oliveira, and Nuno Rodrigues. Application of ar and vr in hand rehabilitation: A systematic review. *Journal of Biomedical Informatics*, 111, 11 2020. doi: 10.1016/j.jbi.2020.103584.

[47] Moisés Pacheco Lorenzo, Sonia Valladares, Luis Anido-Rifón, and Manuel José Fernández Iglesias. Smart conversational agents for the detection of neuropsychiatric disorders: A systematic review. *Journal of Biomedical Informatics*, 113, 12 2020. doi: 10.1016/j.jbi.2020.103632.

[48] Alessio Signorini, Alberto Segre, and P.M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. *during the influenza A H1N1 pandemic. PLoS One*, 6, 01 2011.

[49] Aron Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Language Resources and Evaluation*, 47, 03 2012. doi: 10.1007/s10579-012-9185-0.

[50] Liang Zhao, Jiangzhuo Chen, Virginia Tech, Feng Chen, Wei Wang, and Chang-Tien Lu. Simnest: Social media nested epidemic simulation via online semi-supervised deep learning. In *Proceedings / IEEE International Conference on Data Mining. IEEE International Conference on Data Mining*, volume 2015, pages 639–648, 11 2015. doi: 10.1109/ICDM.2015.39.

[51] Mauricio Santillana, Andre Nguyen, Mark Dredze, Michael Paul, and John Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11, 08 2015. doi: 10.1371/journal.pcbi.1004513.

[52] Xiangfeng Dai and M. Bikdash. Hybrid classification for tweets related to infection with influenza. *Conference Proceedings - IEEE SOUTHEASTCON*, 2015, 06 2015. doi: 10.1109/SECON.2015.7133015.

[53] Donal Simmie, Nicholas Thapen, Chris Hankin, and Joe Gillard. Defender: Detecting and forecasting epidemics using novel data-analytics for enhanced response. *PLOS ONE*, 11, 04 2015. doi: 10.1371/journal.pone.0155417.

[54] David Hartley, Courtney Giannini, Stephanie Wilson, Ophir Frieder, Peter Margolis, Uma Kotagal, Denise White, Beverly Connelly, Derek Wheeler, Dawit Tadesse, and Maurizio Macaluso. Coughing, sneezing, and aching online: Twitter and the volume of influenza-like illness in a pediatric hospital. *PLOS ONE*, 12:e0182008, 07 2017. doi: 10.1371/journal.pone.0182008.

[55] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Feature constrained multi-task learning models for spatiotemporal event forecasting. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 01 2017. doi: 10.1109/TKDE.2017.2657624.

[56] V.K. Jain and S. Kumar. Rough set based intelligent approach for identification of h1n1 suspect using social media. *Kuwait Journal of Science*, 45:8–14, 01 2018.

[57] Koustav Rudra, Ashish Sharma, Niloy Ganguly, and Muhammad Imran. Classifying and summarizing information from microblogs during epidemics. *Information Systems Frontiers*, 20, 10 2018. doi: 10.1007/s10796-018-9844-9.

[58] Hongping Hu, Haiyan Wang, Feng Wang, Daniel Langley, Adrian Avram, and Maoxing Liu. Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network. *Scientific Reports*, 8, 03 2018. doi: 10.1038/s41598-018-23075-1.

[59] Shoko Wakamiya, Yukiko Kawai, and Eiji Aramaki. Twitter-based influenza detection after flu peak via tweets with indirect information: Text mining study. *JMIR Public Health Surveill*, 4(3):e65, Sep 2018. ISSN 2369-2960. doi: 10.2196/publichealth.8627.

[60] Ali Alessa and Miad Faezipour. Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study. *JMIR Public Health Surveill*, 5(2):e12383, Jun 2019. ISSN 2369-2960. doi: 10.2196/12383.

[61] Hongxin Xue, Yanping Bai, Hongping Hu, and Hdsajkkd Ldfs. Regional level influenza study based on twitter and machine learning method. *PLOS ONE*, 14:e0215600, 04 2019. doi: 10.1371/journal.pone.0215600.

[62] Xiaoyi Fu, Xu Jiang, Yunfei Qi, Meng Xu, Yuhang Song, Jie Zhang, and Xindong Wu. An event-centric prediction system for covid-19. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 195–202, 2020. doi: 10.1109/ICBK50248.2020.00037.

[63] Mark Abraham, Peter Nabende, and Ernest Mwebaze. Ontology driven machine learning approach for disease name extraction from twitter messages. In *2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*, pages 68–73, 09 2017. doi: 10.1109/CIAPP.2017.8167182.

[64] Paola Velardi, Giovanni Stilo, Alberto Tozzi, and Francesco Gesualdo. Twitter mining for fine-grained syndromic surveillance. *Artificial Intelligence in Medicine*, 61, 07 2014. doi: 10.1016/j.artmed.2014.01.002.

[65] Liangzhe Chen, K. S. M. Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B. Aditya Prakash. Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. In *2014 IEEE International Conference on Data Mining*, pages 755–760, 2014. doi: 10.1109/ICDM.2014.137.

[66] Michael J. Paul and Mark Dredze. Discovering health topics in social media using topic models. *PLOS ONE*, 9:1–11, 08 2014. doi: 10.1371/journal.pone.0103408.

[67] Liangzhe Chen, K. S. M. Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B. Prakash. Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. *Data Mining and Knowledge Discovery*, 30, 08 2015. doi: 10.1007/s10618-015-0434-x.

[68] Tim Mackey, Vidya Lakshmi Purushothaman, Jiawei Li, Neal Shah, Matthew Nali, Cortni Bardier, Bryan Liang, Mingxiang Cai, and Raphael Cuomo. Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with covid-19 on twitter: Retrospective big data infoveillance study. *JMIR Public Health and Surveillance*, 6, 04 2020. doi: 10.2196/19509.

[69] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *2011 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2011*, pages 702–707, 05 2011. doi: 10.1109/INFCOMW.2011.5928903.

[70] Eui-Ki Kim, Jong Seok, Jang Oh, Hyong Lee, and Kyung Hyun Kim. Use of hangeul twitter to track and predict human influenza infection. *PloS one*, 8:e69305, 07 2013. doi: 10.1371/journal.pone.0069305.

[71] Anoshé Aslam, Ming-Hsiang Tsou, Brian Spitzberg, Li An, Jean Gawron, Dipak Gupta, K. Peddecord, Anna Nagel, Christopher Allen, Jiue-An Yang, and Suzanne Lindsay. The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of medical Internet research*, 16:e250, 11 2014. doi: 10.2196/jmir.3532.

[72] Michael Paul, Mark Dredze, and David Broniatowski. Twitter improves influenza forecasting. *PLoS currents*, 6, 10 2014. doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

[73] Daniel Janies, Zachary Witter, Christian Gibson, Thomas Kraft, Izzet Senturk, and Umit Catalyurek. Syndromic surveillance of infectious diseases meets molecular epidemiology in a workflow and phylogeographic application. *Studies in health technology and informatics*, 216:766–70, 08 2015.

[74] David Andre Broniatowski, Mark Dredze, Michael J Paul, and Andrea Dugas. Using social media to perform local influenza surveillance in an inner-city hospital: A retrospective observational study. *JMIR Public Health Surveill*, 1(1):e5, May 2015.

[75] Fred Lu, Suqin Hou, Kristin Baltrusaitis, Manan Shah, Jure Leskovec, Rok Sosic, Jared Hawkins, John Brownstein, Giuseppe Conidi, Julia Gunn, Josh Gray, Anna Zink, and Mauricio Santillana. Accurate influenza monitoring and forecasting using novel internet data streams: A case study in the boston metropolis. *JMIR Public Health and Surveillance*, 4:e4, 01 2018. doi: 10.2196/publichealth.8950.

[76] Balsam Alkouz and Zaher Al Aghbari. Analysis and prediction of influenza in the uae based on arabic tweets. In *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, pages 61–66, 03 2018. doi: 10.1109/ICBDA.2018.8367652.

[77] Yasmin Khan, Garvin Leung, Paul Bélanger, Effie Gournis, David Buckeridge, Li Liu, Ye Li, and Ian Johnson. Comparing twitter data to routine data sources in public health surveillance for the 2015 pan/parapan american games: an ecological study. *Canadian Journal of Public Health*, 109, 04 2018. doi: 10.17269/s41997-018-0059-0.

[78] Carmela Comito, Agostino Forestiero, and Clara Pizzuti. Improving influenza forecasting with web-based social data. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 963–970, 08 2018. doi: 10.1109/ASONAM.2018.8508563.

[79] Moritz Wagner, Vasileios Lampos, Ingemar Cox, and Richard Pebody. The added value of online user-generated content in traditional methods for influenza surveillance. *Scientific Reports*, 8, 09 2018. doi: 10.1038/s41598-018-32029-6.

[80] Daniel O'Leary and Veda Storey. A google–wikipedia–twitter model as a leading indicator of the numbers of coronavirus deaths. *Intelligent Systems in Accounting, Finance and Management*, 27:151–158, 07 2020. doi: 10.1002/isaf.1482.

[81] Lauren Sinnenberg, Alison Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina Merchant. Twitter as a tool for health research: A systematic review. *American Journal of Public Health*, 107:143–143, 01 2017. doi: 10.2105/AJPH.2016.303512a.

[82] Davide Golinelli, Erik Boetto, Gherardo Carullo, Andrea Nuzzolese, Maria Landini, and Fantini Mariapia. Adoption of digital technologies in health care during the covid-19 pandemic: Systematic review of early scientific literature. *Journal of Medical Internet Research*, 22, 11 2020. doi: 10.2196/22280.

[83] Alana Corsi, Fabiane Souza, Regina Pagani, and Joao Kovaleski. Big data analytics as a tool for fighting pandemics: a systematic review of literature. *Journal of Ambient Intelligence and Humanized Computing*, 10 2020. doi: 10.1007/s12652-020-02617-4.

[84] A. Alamoodi, O.s Albahri, A.s Albahri, Bilal Bahaa, Rami Malik, A. Zaidan, Musaab Alaa, K. Mohammed, Hamsa Hammed, Z Al-Qays, and M. Alqaysi. Sentiment analysis and its applications in fighting covid-19 and infectious diseases: A systematic review. *Expert Systems with Applications*, 167, 10 2020. doi: 10.1016/j.eswa.2020.114155.

[85] Yi Guo, Yahan Zhang, Tianchen Lyu, Mattia Prosperi, Fei Wang, Wang Qi, and Jiang Bian. The application of artificial intelligence and data integration in covid-19 studies: A

scoping review. *Journal of the American Medical Informatics Association*, 28, 06 2021. doi: 10.1093/jamia/ocab098.

[86] A. R. Sanaullah, Anupam Das, Anik Das, Ashad Kabir, and Kai Shu. Applications of machine learning for covid-19 misinformation: A systematic review. *Social Network Analysis and Mining*, 12(1):1–34, December 2022. ISSN 1869-5469. doi: 10.1007/s13278-022-00921-9. Publisher Copyright: © 2022, The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature.

[87] Barbara Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. 2, 01 2007.

[88] Bhavani Devi Ravichandran and Pantea Keikhosrokiani. Classification of covid-19 misinformation on social media based on neuro-fuzzy and neural network: A systematic review. *Neural Computing and Applications*, 35:1–19, 09 2022. doi: 10.1007/s00521-022-07797-y.

[89] Dinesh Gunasekeran, Alton Chew, Eeshwar Chandrasekar, Priyanka Rajendram, Vasundhara Kandarpa, Mallika Rajendram, Audrey Chia, Helen Smith, and Choon Leong. The impact and applications of social media platforms for public health responses before and during the covid-19 pandemic: Systematic literature review (preprint). 09 2021. doi: 10.2196/preprints.33680.

[90] Christopher Mckinley and Yam Limbu. Promoter or barrier? assessing how social media predicts covid-19 vaccine acceptance and hesitancy: A systematic review of primary series and booster vaccine investigations. *Social Science & Medicine*, 340:116378, 01 2024. doi: 10.1016/j.socscimed.2023.116378.

[91] Mahmud Omar, Dana Brin, Benjamin Glicksberg, and Eyal Klang. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. 01 2024. doi: 10.1101/2024.01.14.24301289.

[92] Thayer Alshaabi, David Rushing Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *EPJ data science*, 10(1):15, 2021.

[93] Ching Nam Hang, Pei-Duo Yu, Siya Chen, Chee Wei Tan, and Guanrong Chen. Mega: Machine learning-enhanced graph analytics for infodemic risk management. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[94] Tyler Mccormick, Hedwig Lee, Nina Cesare, Ali Shojaie, and Emma Spiro. Using twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods amp Research*, 09 2015. doi: 10.1177/0049124115605339.

[95] tweepy.org. *Tweepy Documentation*, accessed June 22, 2021. URL https://docs.tweepy.org/en/stable/.

[96] Jie Yin, Sarvnaz Karimi, Bella Robinson, and Mark Cameron. Esa: emergency situation awareness via microbloggers. 10 2012. doi: 10.1145/2396761.2398732.

[97] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Twitcident: Fighting fire with information from social web streams. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion*, 04 2012. doi: 10.1145/2187980.2188035.

[98] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and Christian Bizer. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6, 01 2014. doi: 10.3233/SW-140134.

[99] Daniel A. Gónzalez-Bandala, Juan C. Cuevas-Tello, Daniel E. Noyola, Andreu Comas-García, and Christian A. García-Sepúlveda. Computational forecasting methodology for acute respiratory infectious disease dynamics. *Int. J. Environ. Res. Public Health*, 17 (12):4540–, 2020.

[100] Dirección General de Epidemiología Secretaría de Salud. *Boletín Epidemiológico. Sistema Nacional de Vigilancia Epidemiológica. Sistema Único De Información.*, accessed June, 2023. URL https://www.gob.mx/salud/documentos/boletinepidemiologico-sistema-nacional-de-vigilancia-epidemiologica-sistema-unico-de-informacion.

[101] Dirección General de Epidemiología Secretaría de Salud. *Datos Abiertos Dirección General de Epidemiología*, accessed December, 2022. URL https://datos.covid-19.conacyt.mx/#DownZCSV.

[102] Atish Pawar and Vijay Mago. Calculating the similarity between words and sentences using a lexical database and corpus statistics. 02 2018.

[103] Anthony N. Nguyen, M. Lawley, David P. Hansen, and S. Colquist. A simple pipeline application for identifying and negating snomed clinical terminology in free text. 2009.

[104] Jon Patrick, Yefeng Wang, and Peter Budd. An automated system for conversion of clinical notes into snomed clinical terminology. volume 68, pages 219–226, 01 2007.

[105] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. 01 2009. ISBN 978-0-596-51649-9.

[106] Sean Bleier. *NLTK's list of english stopwords*, 2010 (accessed June 18, 2021). URL https://gist.github.com/sebleier/554280.

[107] Kruti Nargund and S. Natarajan. Public health allergy surveillance using micro-blogs. pages 1429–1433, 09 2016. doi: 10.1109/ICACCI.2016.7732248.

[108] Gunjit Bedi. *A guide to Text Classification(NLP) using SVM and Naive Bayes with Python*, 2018. URL https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34.

[109] Long Ma and Yanqing Zhang. Using word2vec to process big text data, 10 2015.

[110] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[111] Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. GluonTS: Probabilistic and Neural Time Series Modeling in Python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL http://jmlr.org/papers/v21/19-820.html.