Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Future Generation Computer Systems 112 (2020) 641–657

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America

José Antonio García-Díaz, Mar Cánovas-García, Rafael Valencia-García *

Departamento de Informática y Sistemas, Universidad de Murcia, 30100, Murcia, Spain

ARTICLE INFO

Article history:
Received 10 March 2020
Received in revised form 12 June 2020
Accepted 14 June 2020
Available online 18 June 2020

Keywords:
Aspect-based sentiment analysis
Infodemiology
Deep learning
Ontologies

ABSTRACT

Infodemiology is the process of mining unstructured and textual data so as to provide public health officials and policymakers with valuable information regarding public health. The appearance of this new data source, which was previously unimaginable, has opened up a new way in which to improve public health systems, resulting in better communication policies and better detection systems. However, the unstructured nature of the Internet, along with the complexity of the infectious disease domain, prevents the information extracted from being easily understood. Moreover, when dealing with languages other than English, for which some of the most common Natural Language Processing resources are not available, the correct exploitation of this data becomes even more difficult. We intend to fill these gaps proposing an ontology-driven aspect-based sentiment analysis with which to measure the general public's opinions as regards infectious diseases when expressed in Spanish by employing a case study of tweets concerning the Zika, Dengue and Chikungunya viruses in Latin America. Our proposal is based on two technologies. We first use ontologies in order to model the infectious disease domain with concepts such as risks, symptoms, transmission methods or drugs, among other concepts. We then measure the relationship between these concepts in order to determine the degree to which one concept influences other concepts. This new information is subsequently applied in order to build an aspect-based sentiment analysis model based on statistical and linguistic features. This is done by applying deep-learning models. Our proposal is available on a web platform, where users can see the sentiment for each concept at a glance and analyse how each concept influences the sentiment of the others.

1. Introduction

Outbreaks of infectious diseases are responsible for high mortality rates and the collapse of public health services, causing panic among citizens. Globalisation and the constant flow of people across political and geographic boundaries allow the rapid spread of infectious diseases around the world, thus increasing the contagion ratio and the consequent spread of these diseases. Proof of this is that in only the last two decades, humanity has had to confront the Severe Acute Respiratory Syndrome (SARS), the H1N1 outbreak, the Ebola virus, the Zika Virus, and the new Coronavirus (2019-nCoV).

During outbreaks of infectious diseases, public organisations should play a leading role as regards managing available resources, developing and carrying out prevention and action strategies, and coordinating with news media in order to provide citizens with proven and useful information. The analysis of past actions carried out by organisations has, however, proved to be inadequate, since it is insufficient or ineffective, and is often more part of the problem than the solution [1–3].

The social emergence derived from outbreaks has drawn the analysis of public opinions towards health-data to the attention of disciplines such as health communication, public relations, or medical informatics to a great significant [4]. Infodemiology, which is the process of mining public health data on the Internet, has consequently, opened up a new possibility as regards improving public health systems, thus resulting in better communication policies and better detection systems. For example, modern syndromic surveillance systems based on Infodemiology contribute to the early detection of outbreaks, thus complementing Sentinel Surveillance systems [5,6]. This new knowledge is highly strategic because it can detect outbreaks in early-stages and allow public

* Corresponding author.
E-mail addresses: joseantonio.garcia8@um.es (J.A. García-Díaz), mariamar.canovasg@um.es (M. Cánovas-García), valencia@um.es (R. Valencia-García).

https://doi.org/10.1016/j.future.2020.06.019
0167-739X/

authorities to reduce the devastating impact infectious diseases that cause in society [7].

Another interesting application of infodemiology consists of the analysis of social networks in order to measure the general public's interest, their behaviour, and how they respond to health policies during outbreaks [8]. In general, the opinions expressed in social networks regarding the health domain consist mostly of negative sentiments [9]. Negative feelings spread quickly owing to certain phenomena, such as the *echo-chamber effect*, which promotes existing beliefs and confirms biases [10]. These factors make social networks a breeding ground for misinformation and hoaxes, which can lead the general public to take fewer or contraindicated measures as regards preventing disease transmission [11]. Assessing the general public's sentiments in a quick and reliable manner, therefore, helps the public authorities to develop better communication strategies that will prevent misinformation and the spreading of false rumours. Moreover, this analysis can help public authorities to confirm whether the strategies adopted are perceived as useful by their citizens by tracking the dynamics of the sentiments (that is, how the polarity of the feelings change over time).

With regard to mining the Internet in search of health-related data, Natural Language Processing (NLP) encompasses a series of techniques that can retrieve and analyse information in texts written in natural language. Among other applications, NLP includes techniques with which to identify entities, sentiments and subjective polarities. However, the information on the Internet is commonly stored in an unstructured manner, which hinders the mining process. Moreover, certain linguistic phenomena related to figurative language, the inherent ambiguity of natural language and some forms of abbreviated language and slang popular in social media, may mask the author's true intention. Another added difficulty occurs when working with non-English languages for which the latest version of some NLP resources is not available.

Sentiment-Analysis is the NLP technique whose objective is to extract the sentiment polarity from a text, determining whether a piece of text is positive, negative or neutral [12]. Of the various alternatives available, one popular approach with which to perform Sentiment Analysis consists of using machine-learning models that find patterns in a set of pre-trained examples and employing this information to create a prediction model that can extract the sentiment of a document as a whole. Most modern approaches make it possible to do this in more fine-grained detail by measuring the sentiment for subtopics individually rather than dealing with the document as a whole. This approach is known as aspect-based sentiment analysis.

We argue that the inclusion of a formal representation of a specific domain, such as that of infectious diseases, in aspect-based sentiment analysis will improve the results attained by taking into account how the concepts are related. Moreover, the formal representation that ontologies provide may be used as regards representing and summarising the findings in a logical structure in which public authorities can see, at a glance, the degree of sentiment associated with different topics regarding infectious diseases such as social distance, confinement regulations, vaccination campaigns or the quality of services provided by a hospital.

This work consequently performs a fine-grained analysis of the general public's opinions of infectious diseases in texts written in Spanish. Our proposal is based on two technologies. We first use ontologies in order to model health-domain concepts and two metrics to measure the relationship among the concepts of the ontology: semantic similarity and semantic relatedness. We then apply aspect-based sentiment analysis models so as to determine the sentiment of tweets in a corpus of infectious diseases. Finally, the sentiments extracted are summarised by concepts and presented in such a way that the public authorities can see the degree of sentiment associated with each concept at a glance, and query new documents regarding infectious diseases for real-time monitoring.

During our research, we have evaluated the performance of various sentiment analysis models based on deep learning techniques, such as feed-forward neural networks, convolutional neural-networks and recurrent neural networks, by applying statistical and linguistic sentiment analysis features. We have also developed a domain-ontology with which to extract concepts from tweets and applied semantic annotation in order to identify concepts that appear explicitly in documents, along with other concepts that do not appear but are semantically related. Finally, we have shared a gold-standard corpus with the community, that has been manually annotated as regard infectious diseases caused by the Zika, Dengue, and Chikungunya viruses in Latin America. The corpus is composed of 10,843 positive tweets, 10,843 negative tweets and 7,659 neutral tweets, all written in Spanish. Each tweet was rated by several volunteers, signifying that there was a consensus when determining the sentiment of the tweets. This corpus is an extension of the work published in [13].

The remainder of this paper is organised as follows: Section 2 describes the state of the art regarding aspect-based sentiment analysis. Section 3 describes the materials and methods employed in our proposal. In Section 4, we analyse the results attained and present the lesson learned. Finally, Section 5 provides a summary of the conclusions of the paper and presents the directions for future research.

2. State of the art

In this paper, we show the process used to conduct an ontology-driven aspect-based sentiment analysis of texts concerning infectious diseases written in Spanish. We consequently: (1) present background information concerning aspect-based sentiment analysis (see Section 2.1), (2) describe the feature engineering approaches employed to represent text-documents (see Section 2.2), and (3) analyse modern approaches used to solve sub-tasks of aspect-based sentiment analysis, emphasising knowledge-based methods (see Section 2.3).

2.1. Aspect-based sentiment analysis classification

Sentiment Analysis (SA), also referred to as Opinion Mining (OM), is the field of Natural Language Processing (NLP) responsible for extracting users' subjective polarity concerning a specific topic [12]. Subjective polarity includes users' attitudes, appraisals and emotions as regards many aspects of society, such as products, organisations, events or services. SA has, therefore, attracted attention owning to its potential for application in marketing [14], customer service [15], infodemiology [16], hate-speech identification [17], spam-filters [18] or fake-news detection [19] among other domains.

SA makes it possible to assume that the whole document being analysed contains only a general opinion regarding that topic. A more detailed analysis can, however, be performed by employing Aspect-Based Sentiment Analysis (ABSA), in which the main topic is divided into subtopics and the sentiment is calculated individually for each subtopic. The two key-benefits of ABSA when compared to SA are, on the one hand, that it allows a detailed analysis of product reviews in which users present their conclusions after analysing and discussing different details of the products or services to which they have had access [20] and, on the other, it makes it possible to aggregate the sentiment score of a set of documents, which summarises what the feelings associated with each aspect are at a glance [21].

The following recurrent sub-tasks can be distinguished in an ABSA pipeline [21]: (1) sentiment and aspect identification, (2) sentiment classification, and (3) sentiment aggregation. There are, however, also solutions that vary or mix these sub-tasks.

The first sub-task, known as sentiment and aspect identification, consists of the identification of subjective opinions and their related items. Each sentiment should be formally defined as a tuple that contains: (1) the sentiment and its polarity; (2) the aspect of the topic to which it alludes; (3) the holder of that sentiment (which may not coincide with the person who expresses it); (4) and the time at which the opinion was expressed [12]. However, the bibliography also contains works that simplify this sub-task by identifying only the sentiments and the aspects, particularly when the main objective is to discover the overall polarity of each aspect rather than carry on an in-depth analysis of the user profiles that contains the sentiments. Some examples of this alternative approach can be found in [22] or at [23]. Further information on aspect extraction can be found in Section 2.3.

The second sub-task, which is known as sentiment classification, consists of assigning a sentiment to each aspect. As the most important indicators of sentiments are opinion words, such as *good*, *bad*, *poor*, or *terrible*, some researchers have applied sentiment lexicons in order to determine the subjective polarity of each word in context [24]. However, sentiment lexicons are not sufficient owing to the complex phenomena present in natural language, such as the usage of figurative language, which changes the meaning of an utterance from its literal meaning [25]. These drawbacks can be dealt with by using supervised machine-learning classifiers to complement sentiment-lexicons with linguistic, statistical, or contextual sentiment-analysis features in order to build a model that is capable of extracting the sentiments and adding semantic, lexical and contextual information. Further information regarding feature extraction can be found in Section 2.2.

The third and last sub-task, sentiment aggregation, consists of combining the sentiment tuples in a manner that will be comprehensive and useful for the final user. A great variety of solutions can be found in works in the bibliography, according to the problem domain. For example, in *ReviewMiner* [26], the authors developed a multi-modal user interface that supports three types of user interaction interfaces: (1) text-based opinion summary and comparison, which shows the reviews segmented into aspects and highlighted with different colours; (2) spatial-based opinion summary and comparison, which shows the opinions on heat-maps in order to attain rapid insights into where the places that the users have reviewed are according to each aspect, and (3) temporal-based opinion summary and comparison, which provides a timeline on which aspect sentiment ratings are displayed. On other occasions, however, this fine-grained detail is not necessary and other summarisation approaches simply calculate the sentiment per aspect by averaging all the sentiment scores by aspect.

2.2. Feature engineering

In order for a machine to work with texts written in natural language, the texts need to be represented as feature vectors. Feature engineering consists of selecting discerning features in order to classify tasks, and can include statistical, linguistic and contextual features. These types of features are described below:

- **Statistical features**. These encompass features that represent documents as vectors of words from a vocabulary. In their most simple form, the Bag of Words (BoW) model consists of a representation of each document of the corpus as a vector, with the frequencies of each word that is present in a certain vocabulary. Despite its simplicity, BoW has been widely adopted as a robust baseline with which to perform sentiment-analysis [27,28], or document classification tasks such as spam filtering [29]. However, the BoW model has two major drawbacks: (1) It is unaware of the context in which words are written because it handles words in isolation; and, (2) as the size of the corpus increases, the BoW model tends to produce sparse vectors that are time and memory consuming. Some authors have attempted to solve the absence of context, by proposing the incorporation of joint-words (bigrams, trigrams) or the usage of sequences of characters rather than words [30]. Other authors have, meanwhile, proposed preventing the high number of features by applying cut-off filters in order to discard words and joint-words that are not very representative. In this respect, rather than measuring the frequency of words in the corpus, some authors employ the term-frequency inverse document-frequency (TF–IDF) [31], which takes into account how often grams appear in the other documents in the corpus with the aim of dismissing popular and non-informative words.
  Word-embeddings are a more efficient solution than approaches based on BoW. In approaches based on word-embeddings, words are represented as dense vectors [32] in order to cluster similar concepts together and to capture certain semantic relationships among words based on the distance between word-vectors. The main idea of word-embeddings is that words with similar semantics are represented with similar vectors. Word-embeddings can be learned from scratch from the corpus, but it is possible to use pre-trained word embeddings and update them during the training of your models [33]. This approach is useful, because word-embeddings are initialised with general semantics rather than random values, and these vectors can still be adjusted during the training stage of a machine-learning pipeline.
  Sentence embeddings are similar to word-embeddings, and capture the document's meaning as a vector, typically by combining all words-embeddings of greater lexical units such as sentences or paragraphs. In their most simplest approach, sentence embeddings are calculated as the average vector of the word-embeddings in the text, although it is possible to find more sophisticated approaches, such as Sent2Vec [34] in which sentence embeddings are calculated as the average vector of source word embeddings of its constituent words, or by applying the Smooth Inverse Frequency (SIF) [35], which down-weights the relevance of common words. Sentence embeddings have been evaluated with several tasks regarding NLP in the medical domain. For example, the authors of [36] performed a comprehensive evaluation of different sentence embedding based models for different tasks, such as semantic similarity, question answering or text-classification. Although some of the models evaluated showed promising results, there was no clear winner that beat the other models for all the tasks.
- **Linguistic features**. These are features that measure the presence and frequency of certain linguistic phenomena. For example, the number of words that belong to a certain Part of Speech (PoS) category, the percentage of uppercase words, or stylistic features, such as expressive lengthening. Unlike statistical methods, such as BoW, linguistic features are less sparse and can better generalise a solution. However, these features are linguistic and context dependent, and translating them into other languages is not, therefore, a trivial task [37].

Some works in the bibliography also contain generic tools with which to capture linguistic features. For example, Linguistic Inquiry and Word Count (LIWC) [38] is a text analysis programme that counts words within pre-established psychological categories that capture content words which convey what people are saying (nouns, verbs or adjectives), and style words that convey how people are expressing something (prepositions, articles, conjunctions or auxiliary verbs). LIWC has been applied in several areas, such as suicide [39], cyber-bulling [40], and satire detection [41]. LIWC is a language-dependent tool that has been translated into several languages, including Spanish [42].

- **Contextual features**. These features are related to the context in which communication took place. Contextual features may include information concerning the author of the post, the date-time of the publication or other documents and conversational features. An in-depth analysis of contextual features can be found in [43], in which the authors classify contextual features as: (1) micro features, which exploit information about the author of the document, such as previous posts, or explore personal data, such as gender or age; (2) meso features, which reflect how users interact with each other, both small and medium scale; and (3) macro features, which incorporate information outside the scope of the social network. The usage of contextual features is popular in domains in which knowledge shared between the speaker and the audience assists in document classification task as regards, for example, sarcasm detection [44].

Once the features have been extracted, they can be used as input for machine-learning models in order to build a predictive model that is capable of inferring the polarity of new documents. The state of the art as regard Sentiment Analysis makes use of neural networks to build classifiers. The main techniques identified are listed below:

- **Multilayer perceptron**. This is the most basic form of feed-forward artificial neural networks (ANNs), composed of one or more layers of perceptrons. These kinds of networks have provided good results in SA tasks. For example, [45] used a deep-learning network based on multi-layer perceptrons to implement a classifier that outperformed other baseline methods based on genetic algorithms.
- **Convolutional Neural Networks (CNNs)**. These are artificial neural networks that exploit the spatial dimension in order to recognise similar word clusters, regardless of their position in the sentence. Although these kinds of neural networks were originally designed for the resolution of computer vision problems, they have outperformed several NLP tasks, such as text classification [46]. Their key aspect is that they handle word clusters, rather than individual words, thus allowing problems related to language ambiguity to be solved. For instance, the semantics of a polysemic word can be disambiguated by looking at the surrounding words. CNNs employ the concept of "convolution" to analyse and filter features in order to spot those that are relevant. CNNs have been applied to ABSA in recent years. For example, in [47], the authors achieved promising results with a dataset containing reviews of restaurants (SemEval 2014 Restaurant Dataset); or in [48], in which the authors applied a CNN for both aspect extraction and aspect-based sentiment analysis, and attained competitive results for different languages and domains.
- **Long Short Term Memory (LSTM)**. These are Recurrent Neural Networks (RNNs) that were designed to avoid the long-term dependency problem and which keep information for long periods of time. When compared with CNNs, which handle features in a space context, RNNs handle features in a time context, signifying that they can understand the meaning of a sentence when information concerning grammar is relevant or when they are long semantic dependencies rather than local keywords. Wang et al. [49] performed aspect-based sentiment classification by applying LSTM in order to connect different parts of the texts when different aspects are taken as input. In a similar vein, Yukun Ma [50] also employed an extension of the LSTM with a hierarchical attention mechanism. They specifically proposed the inclusion of commonsense knowledge of sentiment-related concepts into the training process of a deep neural network.

With regard to the Spanish language, both CNN and RNN have been regularly used to solve NLP tasks, as can observed in the case of evaluation of NLP workshops such as TASS 2018 [51]. In the work of Vilares et al. [52], the authors also focused on the Spanish language and employed LSTM to perform aspect-based sentiment analysis on tweets regarding different topics. They used unsupervised pre-training and sentiment-specific word embedding. LSTM provides some variants. For example, Bidirectional LSTMs (BiLSTM) can learn from past and future information states simultaneously.

The aforementioned deep-learning models can be combined to create more robust solutions. In this respect, Shad Akhtar et al. [53] used a Multi-Layer Perceptron for the sentiment analysis of the financial domain with features extracted from financial word embeddings and lexicon features. They combined three deep learning models based on CNN, LSTM and Gated Recurrent Unit (GRU). Another example of a hybrid approach is shown in [54], in which the authors proposed a hybrid approach based on contextual word embeddings and hierarchical attention models. They evaluated their approach with the datasets from SemEval 2015 and SemEval 2016 and showed that their proposal outperformed the testing accuracy of both datasets.

2.3. Aspect extraction

With regard to aspect extraction, in [21], Kim Schouten and Flavius Frasincar identified the following core approaches: (1) frequency-based, (2) syntax-based, (3) machine learning, and (4) hybrid approaches. Frequency-based approaches are based on the fact that the nouns or compounds names that appear most frequently in the texts are likely to be considered as aspects. The reliability of frequency-based approaches can be increased by making use of certain heuristics in order to prune some of the false positives of aspects [55] or by including implicit information [56]. Syntax-based methods rely on syntactical relations among the texts by, for example, searching for nouns that are modified by adjectives. Machine learning methods can, meanwhile, be classified as machine learning approaches and unsupervised machine learning approaches and it is also possible to create a hybrid approach based on the core approaches described previously. Regarding the approaches based on machine learning, those based on supervised methods rely on the Conditional Random Field (CRF) in order to assign aspects to documents. This can be based on multiple features, such as current and context words, part-of-speech tags, or sentiment scores, among other features. Unsupervised machine-learning approaches, meanwhile, make use of topic modelling to identify the aspect by clustering words. This is done by applying techniques such as latent Dirichlet allocation (LDA) or Latent Semantic Indexing (LMI).

Modern approaches benefit from knowledge-based methods in order to improve aspect-extraction in ABSA [57]. Ontologies have been shown to be an effective method for aspect extraction for a specific domain [23,58]. An ontology is *"a formal and*

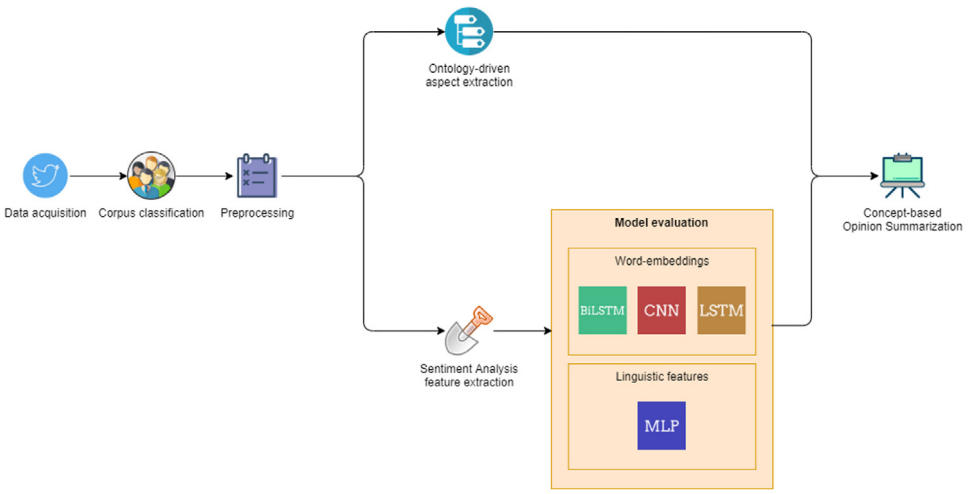

Fig. 1. System architecture of our ontology-driven aspect-based sentiment analysis proposal.

explicit specification of a shared conceptualisation" [59]. Ontologies facilitate knowledge acquisition and knowledge representation for complex domains, in addition to providing mechanisms with which to infer new knowledge from the formal semantics underlying ontology languages.

In order to determine what the subtopic of a text is, one basic approach consists of measuring the frequency of terms mapped onto ontology classes. This task can be extended by including mechanisms that take into account linguistic phenomena such as synonymy, polysemy, and the identification of tacit knowledge [60]. Semantic annotation is the process of tagging ontology class instances to a given text and mapping it onto ontology classes [61] in order to apply semantic reasoners. When performing Semantic annotation, it is first necessary to apply NLP techniques in order to identify sentences, part-of-speech and entities, among other structures. Once these linguistic units have been identified, the next step consists of identifying concepts and their relationships. Semantic annotation can be performed manually, automatically or using a mixed approach. The large amount of data on the Internet makes automatic semantic annotation crucial, since it allows existing data to be labelled [62,63].

Semantic annotation does not take into account the fact that a certain document may refer to subtopics, even if they do not appear explicitly. For example, a text containing a comprehensive list of symptoms is probably in some way related to the diseases of which these are symptoms. Two metrics that can be used to deal with this drawback are semantic similarity and semantic relatedness. On the one hand, semantic similarity is a metric with which to measure the similarity between two concepts that are semantically related by their hierarchy. On the other, semantic relatedness includes other semantic relationships in addition to hierarchy. Couto and Lamurias [64] discussed some metrics that could be used to measure the semantic similarity between two concepts in an ontology based on their hierarchy and shared properties.

3. Materials and methods

In this work, we present an ontology-driven aspect-based sentiment analysis regarding infectious diseases. The method employed to carry out our proposal can be summarised as follows: First, during the data-acquisition stage, a corpus of texts written in Spanish and concerning infectious diseases in Latin America, such as Dengue, Zika or Chikungunya was compiled from Twitter (see Section 3.1). A group of volunteers then manually labelled tweets as positive, neutral or negative (see Section 3.2). In the pre-processing stage, each tweet was subsequently tokenised, normalised and cleaned in order to remove noise (see Section 3.3). Statistical and linguistic features were then extracted for each tweet in the corpus (see Section 3.4), after which various deep-learning machine models and different combinations of the extracted features were evaluated by performing a multi-class classification in order to discover the most reliable model (see Section 3.5). We subsequently extracted the relevant subtopics from the tweets. This was done on the basis of semantic annotation, by employing an ontology of the infectious disease domain in order to match those subtopics that appeared either explicitly or implicitly in the texts (see Section 3.6). Finally, in the concept-based opinion summarisation stage, a sentiment was assigned to each concept in the ontology (see Section 3.7). The architecture of our proposal is depicted in Fig. 1.

3.1. Data acquisition

The case-study included in this paper concerns the study of the general public's social perceptions as regards infectious diseases such as Zika, Dengue, or Chikungunya in Latin America. We have consequently enlarged a corpus of this particular use case, available in [13], which Twitter, a popular micro-blogging platform, was used as a data-provider. Twitter was selected for the following reasons: (1) it is a popular means of spreading news and information [65] and is, therefore, suitable to capture health-related public information; and (2) it makes use of the hashtag, a mechanism that provides users with the ability to create and organise topics on the fly. Hashtags contribute to making this social network a Hub-and-Spoke network [66].

It is worth noting that tweets are restricted to a very short maximum length. After a manual analysis of the corpus, we assumed that, although a tweet can refer to more than one aspect (which is not, according to our analysis, common) it will contain only one sentiment. We have, therefore, assigned only one sentiment to each tweet, as detailed in Section 3.5. Fig. 2 contains an example of a compiled tweet.[1] More details concerning the compilation of the corpus are provided in [13].

One of the problems confronted was how to identify duplicated tweets. Twitter provides a mechanism called retweet that allows users to replicate a tweet from another user. Retweets

[1] In English: *A total of 49 children in Guatemala were born with microcephaly until October of this year in Guatemala as a result of the Zika virus, a figure that exceeds the 36 registered in the last year, according to official data.*

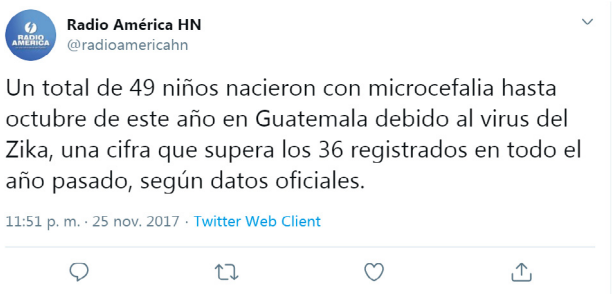

Fig. 2. Example of a tweet in the corpus.

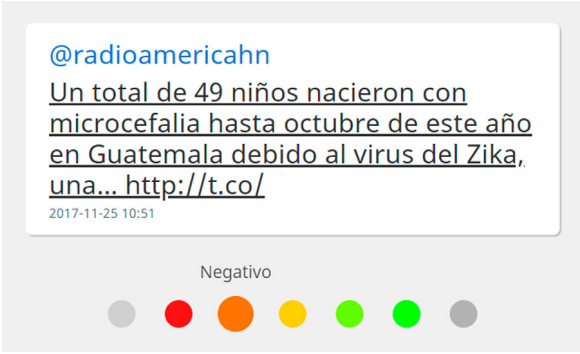

Fig. 3. Screen capture of the tool that the volunteers used to rate the tweets.

are easy to discard. However, not all users rely on the retweet mechanism and may manually copy–paste the original tweet. As Twitter uses a short-mechanism that generates random and reduced versions of URLs, it is difficult to identify duplicated content. We decided to solve this problem by replacing hyperlinks with the http://t.co token. However, we maintained the presence and position of hyperlinks in the tweet, but were able to discard most of the duplicated tweets automatically.

3.2. Corpus classification

The tweets were manually classified by a group of volunteers as *positive*, *neutral*, *negative* and *out-of-context*. The tweets were rated individually several times by different users. Each volunteer performed around 3,334 ratings, and each tweet was rated an average of 6.0216 times, thus achieving an inter-coder reliability of 0.6864 after applying Krippendorff's Alpha [67]. This resulted in the compilation of 10,843 *positive* tweets, 10,843 *negative* tweets, and 7,659 *neutral* tweets. Fig. 3 shows the tool used by the volunteers.

As mentioned in the Model evaluation section (see Section 3.5), we organised the corpus into three sets: (1) *multi-class*, (2) *neutral vs. non-neutral*, and (3) *positive vs. negative* tweets. Each division of the labelled corpus was balanced and shared with the community. However, accordingly to Twitter guidelines,[2] and because the users maintain their rights as regards the content of their tweets, only the IDs of the tweets are provided. As each corpus is balanced, the first half of each file corresponds to IDs of tweets labelled as *neutral* or *positive* for the *neutral vs-non-neutral*, and *positive vs. negative corpus*, respectively, whereas the last half of the list corresponds to IDs of tweets labelled as *non-neutral*, or *negative* for the *neutral-vs-non-neutral*, and *positive-vs-negative* corpus, respectively. This file can be downloaded at https://pln.inf.um.es/corpora/zika/zika-spanish-2020.rar.

[2] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.

3.3. Pre-processing

Prior to sentiment analysis feature extraction and the ontology-driven aspect extraction, we performed a pre-processing stage in order to remove noise from the data. Our pre-processing pipeline essentially consisted of: (1) transforming each letter into its lowercase form; (2) removing blank lines and HTML tags; (3) removing mentions; (4) removing the hashtag symbol while maintaining the rest of the word; and (5) fixing misspellings and removing continuously repeated symbols. For example, the tweet "Iniciará Semana Nacional contra el Dengue en jardín de niños http://bit.do/eNeUg #PiedrasNegras #Salud"[3] became "*iniciará semana nacional contra el dengue en jardín de niños. piedrasnegras salud*". We maintained the original version of each tweet in order to measure certain linguistic phenomena, such as the presence of linguistic errors.

3.4. Sentiment analysis feature extraction

The Sentiment Analysis feature identification process was divided into two major sets: (1) statistical features, and (2) linguistic features. We decided to avoid contextual features – features related to the context in which the tweet was written – because these kinds of features cannot be retrieved when analysing tweets from external news sources.

With regard to statistical features, we used pre-trained word-embeddings from fastText [68]. We specifically applied the pre-trained word vectors from Wikipedia and Common Crawl [69] and then configured the embeddings layer in order to allow the weights of the embeddings to be updated during training. In contrast with standard word vectors, fastText includes a bag of character n-gram vectors in the word vectors. The word vectors generated by fastText consequently take care of the internal structure of the words, which makes it more robust to unusual words that were not seen during training, because they can be broken down into character n-grams in order to obtain their embeddings. We decide to use these embeddings and this configuration because there are cultural and background differences amongst Spanish-speaking countries as regards the language spoken in each one [70,71].

In addition to the statistical features, we extracted a total of 253 different linguistic features, which we organised in the following categories:

- **Grammatical (GRA)**. These features measure the frequency of PoS, such as adjectives, pronouns, or verbs. We identified the main PoS categories by using the Stanford POS Tagger [72]. Spanish is a language that inflects for gender, nouns and adjectives. Unlike English, whose use of the inflection of verbs is limited, Spanish makes use of inflection to indicate the tense and mood of verbs, along with the person to whom they refer, which is useful as regards determining whether users are reporting events, talking about the future, or even talking about hypothetical events. However, the Stanford POS Tagger has some limitations in Spanish and is not able to capture some of these aspects from the verbs. We, therefore, extended the Stanford POS Tagger by compiling a list of a thousand Spanish popular verbs and their respective conjugations obtained from online resources.
- **Grammar and spelling mistakes (ERR)**. We used the original tweet before the pre-processing stage to detect stylistic and linguistic errors. For example, we identified words and expressions that capture informal speech language, such as colloquialisms, the usage of popular abbreviations in texting

[3] In English: National fight against Dengue Week will held in kindergarten.

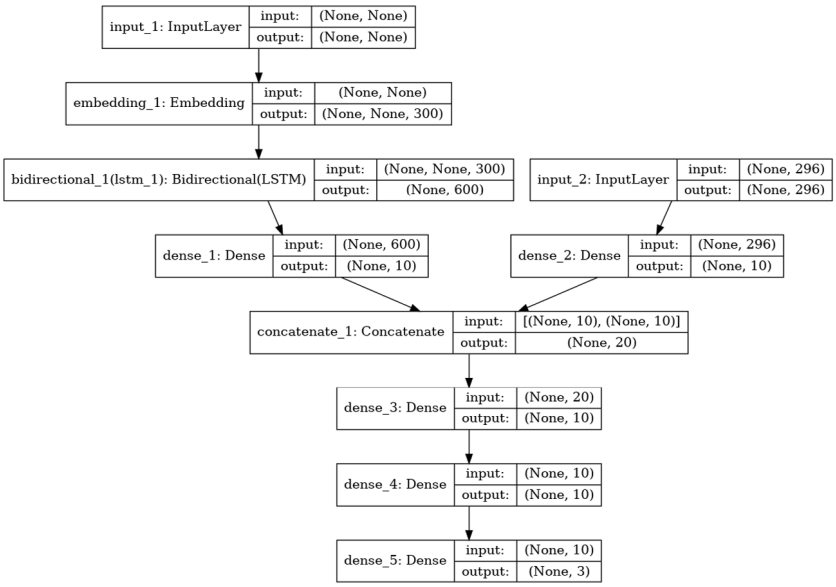

Fig. 4. Model architecture of the BiLSTM + LF model.

languages or non-fluent markers. These features are useful as regards capturing information about the writer's cultural level, and whether s/he has paid sufficient attention when writing the message. It is, for example, assumed that messages written by official media will contain less grammatical errors and less slang.

- **Figurative language (FIG)**. Figurative language is the usage of linguistic devices in order to change the meaning of an utterance from its literal meaning [25]. The identification of figurative language is challenging, because it is heavily dependent on the context and the writer's cultural background. However, we decided to compile regular expressions that capture some hyperboles, Spanish popular idiomatic expressions, rhetorical questions, verbal irony, understatements, metaphors and similes.
- **Pragmatics (PRA)**. These features capture semantics in order to measure emphasis and emotion. They include categories that measure the percentage of words written using uppercase letters, which indicates shouting, and expressive lengthening [73], which is the intentional repetition of letters in the text. We additionally captured the presence of discourse-markers in order to analyse the flow and structure of the utterances. We distinguish among structuring, connectors, reformers, conversational, and argumentative discourse markers.
- **Linguistic Processes (LPR)**. These features measure the length of the tweets, the number of words on average, how many sentences there are in a text and their types (declarative, exclamatory, interrogative, literal cites) and the percentage of long and short words. We also measured the readability of a text by applying different readability formulas based on the length of the tweets and the percentage of words and syllables.
- **Punctuation and symbols (SYM)**. These features capture typographical symbols. We distinguish between (1) sentence dividers, such as spaces, colons, and semicolons, in order to capture the rhythm of the text, and (2) general-purpose symbols, such as those used to express measures as units. We additionally captured some of the terms specifically employed in Twitter, such as hashtags, mentions or hyperlinks, because they have been proven to be discriminating features for Sentiment Analysis [41].
- **Socio-linguistics (SLI)**. These features include a collection of general topics regarding family, personal issues, health, food, animals or affiliations among other features.
- **Sentiment-Analysis (SA)**. These features capture positive and negative emotions with the usage of different lexicons. With regard to negative words, we made a fine-grained classification in the following subcategories: anger, despicable, anxiety and sad. We also compiled a list of positive and negative emoticons.

3.5. Model evaluation

Once the linguistic and statistical features had been extracted, we evaluated the following deep-learning models: (1) Multilayer perceptron, (2) Convolutional neural-networks (CNN), and (3) two variants of the Long Short Term Memory (LSTM): normal and bidirectional in order to create a classifier that would be capable of extracting the overall polarity of a tweet. The comparison between the sentiment-analysis models was performed using the following accuracy metric (see Eq. (1)).

$$Accuracy = TP + TN/(TP + TN + FP + FN) \quad (1)$$

Linguistic features are used as input for the statistical features, whereas word-embeddings are the inputs for CNN, LSTM and BILSTM. As mentioned in the State of the art section (see Section 2.2), these models were selected because CNN and RNN can respectively handle spatial and sequential data, which can assist in some of the challenges regarding NLP, such as word disambiguation. Statistical features are independent of time and space and were, therefore, trained using standard neural networks. However, linguistic features can identify stylistic features, pragmatics, idioms or other forms of figurative language that contain semantic information.

Classifiers were implemented using TensorFlow [74] and Keras [75]. For each experiment, we evaluated the accuracy of linguistic features (LF), the deep neural networks (CNN, LSTM and BiLSTM) and the combination of each deep neural model with linguistic features. We used Keras' functional API in order to combine the multiple inputs from linguistic features and word embeddings.

We include the description of the model architecture of BiLSTM + LF (see Fig. 4) and CNN + LF (see Fig. 5). In both cases, the linguistic features were trained with regular dense layers

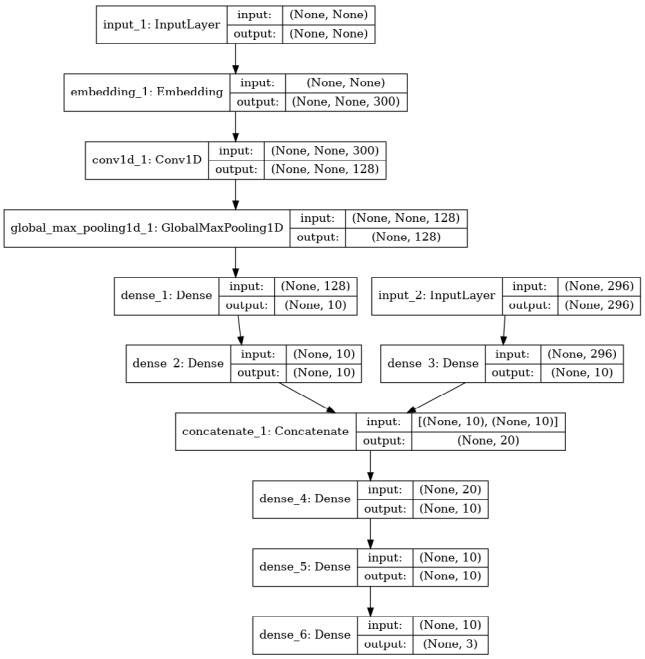

Fig. 5. Model architecture of the CNN + LF model.

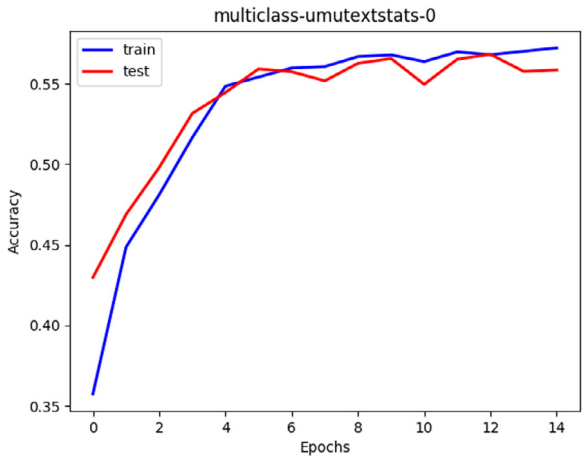

Fig. 6. Evolution of accuracy as regards training and testing with LF.

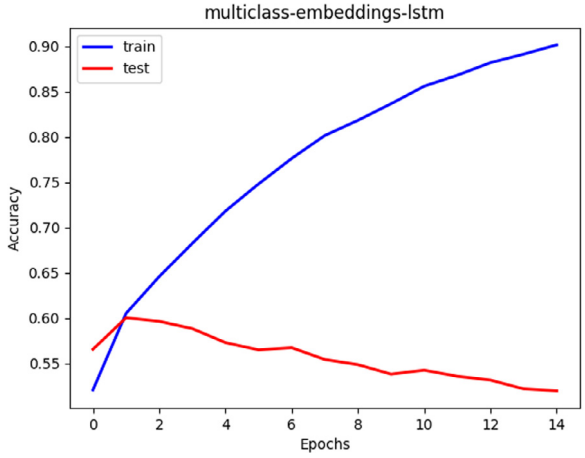

Fig. 7. Evolution of accuracy as regards training and testing with LSTM.

that reduced the 252 features to 10 features. In the case of word-embeddings, we used an Embedding Layer with pre-trained word-embeddings from fastText, in which the weights were updated during training. When applying CNN, the layer architecture consisted of a convolutional layer (Conv1D), a global max pooling (GlobalMaxPooling1D) and two regular dense layers. In the case of BiLSTM, we used a bidirectional layer with LSTM. Regardless of whether we use CNN or RNN, LF and word-embeddings are concatenated and the resulting layer passes through a two-layer neural network until the final prediction is obtained

All the experiments were evaluated using ten-cross validation to split the corpus into training and testing in order to estimate the performance of a model on unseen data in a robust manner. K-fold divides the training dataset into k subsets and uses all of them except one for training and the remaining one for evaluation. This process is repeated until all folds are used as a validation set. Finally, the performance measure calculates average accuracy for all the models. We trained our models using 15 epochs, and each model was tuned using hyper-parameter optimisation.

The first evaluation of the corpus was treated as a multi-class problem. As the corpus contains more positive and negative than neutral statements, we decided to randomly remove some instances in order to maintain the balance. In multi-class problems with *N* classes, neural networks have N output neurons, similar to the one-vs-all approach. We applied soft-max activation to assign the highest value in order to determine the class. The result of each fold and the average for each model is shown in Table 1, in which LF stands for linguistic features trained with a Multi-layer Perceptron, and LSTM, BiLSTM, and CNN respectively refer to Long Short Term Memory, BiLSTM for bi-directional Long Short Term Memory, and CNN for Convolutional neural-networks applied to word-embeddings.

As will be observed in Table 1, the best average accuracy is obtained by using linguistic features with Multilayer perceptron, with average accuracy of 55.3%, followed by BiLSTM + LF model, with an average accuracy of 54.2%. Note that the combination of linguistic features with word-embeddings when applying the LSTM neural networks does not improve the results, which suggests that some of the generalisations performed during training are contradictory in each feature set. With regard to word-embeddings with CNN, the results are similar regardless of whether or not the linguistic features are included.

The following figures plot the learning rate for training and testing sets for each model in isolation: linguistic features (see Fig. 6), and word-embeddings with LSTM (see Fig. 7), BiLSTM (see Fig. 8), and CNN (see Fig. 9) with a random fold. Note that LF and BiLSTM generalise well with the training data since the test data remain stable while CNN and regular LSTM continue to improve the training data, although the performance on the test data gradually decreases.

As mentioned previously in the description of the first experiment, because there were more positive and negative than neutral tweets, we randomly removed some tweets in order to keep the corpus balanced. However, in order to maintain as many tweets as possible, we tried an alternative approach with two binary classifications. We first classified all the tweets in order to form two sets, namely (1) *neutral vs. non-neutral*, and (2) *positive vs. negative*, thus evaluating the deep learning classifiers in order to predict whether or not the tweets were neutral. We then classified the non-neutral tweets through the use of a deep-learning model that was capable of distinguishing between positive and negative tweets. The results of the two binary classifiers are shown in Table 2, and Table 3 for the *neutral vs. non-neutral*, and *positive vs. negative* datasets respectively.

Table 1
Performance of Sentiment Analysis models for multi-class classification.

Model	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10	**AVG**
LF	57.1	55.4	55.6	55.1	51.9	56.1	52.3	56.7	57.0	55.9	**55.3**
LSTM	52.9	56.4	33.3	33.3	46.8	49.5	49.3	47.1	50.8	48.6	**46.8**
LSTM+LF	53.4	63.6	55.7	46.4	44.6	47.7	48.7	52.5	49.7	47.6	**51.0**
BiLSTM	33.2	51.7	33.4	52.2	52.9	33.2	33.2	51.8	53.6	33.5	**42.9**
BiLSTM+LF	52.3	52.1	56.5	52.4	53.9	57.0	56.6	52.3	51.9	56.7	**54.2**
CNN	51.2	56.0	49.9	45.6	45.6	48.7	50.6	47.7	49.1	48.1	**49.3**
CNN+LF	53.1	53.6	48.3	46.4	46.8	48.2	51.3	44.6	50.0	48.4	**49.1**

Table 2
Performance of Sentiment Analysis models for neutral vs. non-neutral.

Model	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10	**AVG**
LF	64.2	63.2	65.0	64.8	64.9	52.3	64.6	64.5	65.3	63.8	**63.26**
LSTM	58.6	50.0	57.8	58.4	50.0	54.6	51.1	53.6	51.8	50.0	**53.59**
LSTM+LF	59.7	70.0	64.0	60.7	58.4	55.4	52.1	58.0	53.9	53.4	**58.56**
BiLSTM	62.2	59.2	49.8	52.7	50.2	58.7	49.9	50.2	54.1	59.8	**54.70**
BiLSTM+LF	58.9	64.8	63.8	59.3	64.3	65.0	60.0	58.2	65.2	64.4	**62.37**
CNN	60.0	64.1	57.8	58.3	53.9	54.0	53.1	51.8	53.3	53.4	**55.96**
CNN+LF	61.2	61.6	58.7	59.0	56.1	54.1	53.9	51.0	54.3	52.5	**56.23**

Table 3
Performance of Sentiment Analysis models for positive vs. negative.

Model	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10	**AVG**
LF	68.8	69.9	69.8	68.3	67.5	69.8	69.9	69.3	67.9	69.2	**69.03**
LSTM	50.0	50.0	68.9	71.8	67.7	50.0	67.0	50.0	72.0	72.3	**61.96**
LSTM+LF	75.2	71.0	69.3	69.1	67.8	66.9	63.1	69.8	72.6	75.0	**69.98**
BiLSTM	71.8	71.1	49.2	49.2	71.0	70.1	50.8	71.8	71.5	60.2	**63.66**
BiLSTM+LF	69.5	69.3	68.9	71.2	69.4	69.8	69.7	70.2	70.0	69.8	**69.77**
CNN	73.8	73.0	69.8	69.5	68.6	66.8	65.7	71.6	72.0	74.3	**70.52**
CNN+LF	75.3	71.5	71.3	69.9	69.2	66.1	65.2	71.0	70.8	72.8	**70.31**

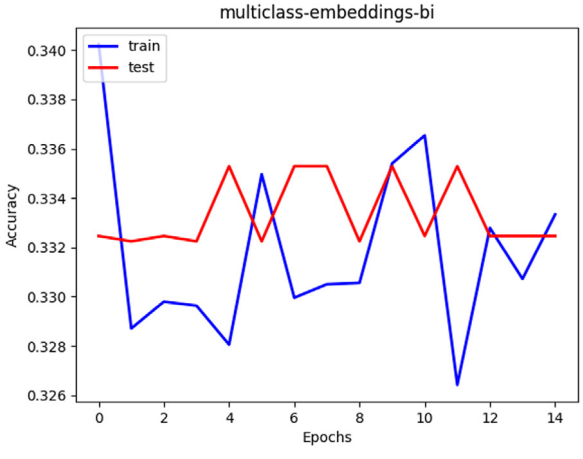

Fig. 8. Evolution of accuracy as regards training and testing with BiLSTM.

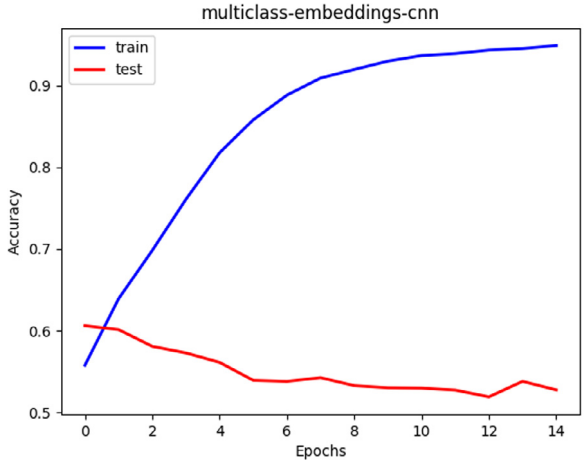

Fig. 9. Evolution of accuracy as regards training and testing with CNN.

The results of the binary classification of the neutral and non-neutral tweets (Table 2) show that linguistic features with Multilayer perceptron achieve the best accuracy with an average accuracy of 63.26%, followed by BiLSTM + LF, with an average accuracy of 62.37%. However, when comparing obtained for the positive and negative tweets (see Table 3), the best result is achieved with word-embeddings and applying CNN, with an average accuracy of 70.52%, followed by CNN + LF with an average accuracy of 70.31%. Needless to say, it is easier to distinguish among positive and negative statements, rather than neutral; and it is difficult to obtain neutral or objective information [76].

3.6. Ontology-driven aspect extraction

The aspect extraction was carried out using a domain-ontology to represent knowledge regarding the domain of infectious diseases. This ontology represents concepts such as infectious diseases, patients, phenotype, risks, symptoms or transmission methods, among other concepts. An ontology allows this knowledge to be represented formally and consistently for concepts and their relationships, along with instances of these concepts with axioms and constraints.

The development process employed to build this ontology followed the Methontology knowledge engineering methodology [77]. The knowledge described by this ontology was collected

Table 4
Top classes of the domain ontology.

Object property	Domain	Range	Inverse
Expresses	Person	Phenotype	isExpressedBy
hasPrevention	Disease	Vaccine	isPreventedBy
hasRisk	Region or Person	Risks	isRiskOf
hasSymptom	Disease or Syndrome	Symptom	isSymptomOf
hasTreatment	Disease	Drugs	isTreatmentOf
isActiveIn	Disease	Region	hasActiveDisease
isPreventedWith	Disease	Vaccine	prevents
Presents	Person	Symptom	isPresentedBy
transmitedBy	Disease	TransmissionProcess	isTransmitedBy
hasRisk	Person or Region	Risk	isRiskOf

from the Disease Ontology (DO) [78], and Infectious Diseases Ontology (IDO) [79]. On the one hand, DO is a open-source ontological description of human diseases that includes disease concepts from the Unified Medical Language System (UMLS) [80] and is mapped for specific terminology concepts (SNOMED CT) [81]. The DO ontology makes use of identifiers known as DOIDs for traceability. On the other hand, IDO is a set of inter-operable ontologies regarding infectious diseases. These are composed of the (1) IDO-Core, with relevant entities regarding shared biomedical and clinical aspects of most infectious diseases, and (2) domain specific extensions towards specific infectious diseases, such as Influenza, Dengue, Malaria and HIV, among others. In addition to the DO and IDO ontologies, we included concepts with which to represent: (1) direct and indirect transmission processes, (2) syndromes, and (3) proximate, intermediate and distal risks among other classes. The ontology was modelled using the Protégé tool [82] in the Web Ontology Language (OWL).

The main concepts defined are *Disease*, *Drugs*, *Person*, *Phenotype*, *Prevention*, *Region*, *Risk*, *Symptom*, *Syndrome*, *Transmission process* and *Vaccine*. The top classes are disjoint, signifying that an instance cannot belong to more than one concept. This ontology currently contains information principally concerning infectious diseases that affect Latin America, such as Dengue, Zika or Chikungunya; along with their transmission methods and their associated risks. We also developed object properties for the semantic relationship among concepts. Table 4 contains the main object properties of the domain ontology including their domain, range and inverse object-property. An excerpt from this ontology is illustrated in Fig. 10.

To the best our knowledge, there are no reliable translations into Spanish of either the DO or IDO ontologies. As we were dealing with tweets written in Spanish, we manually translated the concepts of the domain ontology. Each concept of the ontology includes annotations with regular expressions with the Spanish translation of the term. We additionally employed regular expressions in order to attain synonyms of each term. For example, the concept *fever* was labelled with the following regular expression, which captured the presence of synonyms and related terms. Eq. (2) contains the regular expression used to capture tweets related to fever with different Spanish synonyms.

As this paper tackles a very specific domain, the ambiguity of the terms is solved merely by including counterexamples with negative and positive look-ahead or negative and positive looking-behind patterns in order to avoid false positives. For example, the usage of the word "*tos*" (cough) is common in Spanish, but this is also an informal way in which to refer to everybody "*todos*" (all). This was one of the few problematic cases that we encountered and was solved by ensuring that the word "*tos*" was not preceded by the conjunction "*con*" (with).

$$fever = (fiebres?\ |decimas|calenturas?\ |hipertermia|temperatura) \tag{2}$$

3.7. Concept-based opinion summarisation

The last step in our pipeline consisted of assigning a sentiment based on the results of the sentiment analysis model to each concept of the ontology. We first looped over each tweet in order to measure the Term Frequency–Inverse Document Frequency (TF–IDF) of each concept. We specifically applied an approach similar to that of [83], in which the semantic distance between concepts is taken into account (see Eq. (3)). The main idea is to measure the frequency of those concepts that explicitly appear in the text and those concepts that are semantically related to them based on the distance between these concepts in the ontology. The way in which distance was calculated is explain is explained in greater detail further in this section. Note that the Term Frequency (see Eq. (5)) is calculated as the number of occurrences of the ontological concept in one specific document ($n_{i,d}$) divided by the sum of the occurrences of all the ontological entities identified in the same document ($\sum_k n_{kd}$). This version of the Term Frequency makes it possible to prioritise documents that refer to one specific subtopic. The Inverse Document Frequency (see Eq. (6)) is calculated as the logarithm between the size of the corpus ($|D|$) divided by the number of all documents annotated with this subtopic (N_i)

$$TF--IDF-e = \sum_{j=1}^{n} \frac{TF-IDF_{j,d}}{e^{distance(i,j)}} \tag{3}$$

$$TF-IDF_{j,d} = TF * IDF \tag{4}$$

$$TF = \frac{n_{id}}{\sum_{k} n_{kd}} \tag{5}$$

$$IDF = log\frac{|D|}{N_i} \tag{6}$$

The distance among the concepts in Eq. (3) is calculated by transforming the ontology into a graph, and applying Dijkstra's algorithm [84] to calculate the distance between each of the concepts and the others. It will be observed that the influence between two concepts is inversely proportional to their exponential distance in order to attenuate the influence of distant concepts. Moreover, we filtered out those nodes that were located at a greater distance than a threshold of 3. The consequence of including the semantic similarity and the semantic relatedness of distant terms was the consideration of terms that do not appear explicitly in the texts.

We assigned different weights according to the relationship between the node and adjacent nodes. The main idea was to represent the distance for object properties differently, as they represent properties such has *hasSymptom* or *hasTreatment* rather than hierarchical relationships, in which two different diseases have only their type in common. However, as the decision to

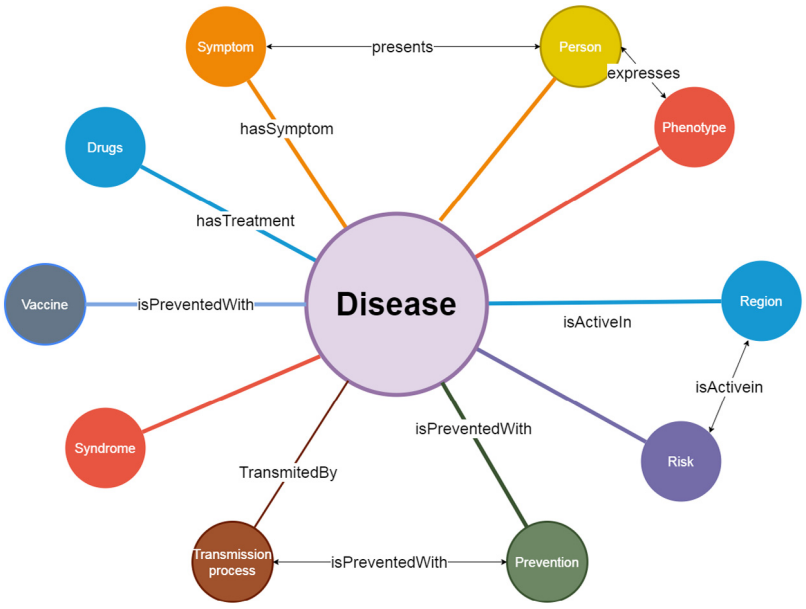

Fig. 10. Ontology.

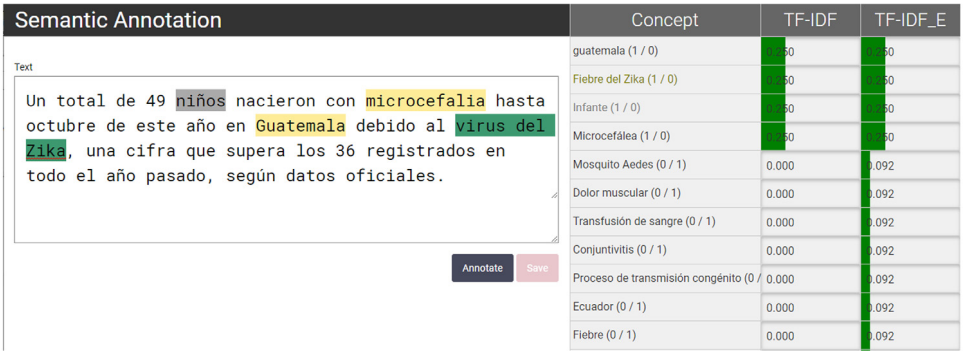

Fig. 11. Example of semantic annotation with a tweet from the corpus.

assign these weight factors would appear to be arbitrary, we decided to keep the distance between adjacent nodes constant, and allow users to adjust these weights in the configuration. Fig. 11 contains an example of the semantic annotation of one document in the corpus, in which the terms identified have been highlighted.

With regard to opinion summarisation, we performed a quantitative analysis by building a vector to represent the positive, neutral and negative scores for each of the aspects extracted. To do this, we performed an analysis of each aspect of the whole corpus after which we performed an iteration for each tweet in the corpus in order to obtain the TF–IDF and TF–IDF-e measures of each concept of the ontology.

The next step consisted of extracting the sentiment from the tweet by applying the multi-class deep-learning model based on BiLSTM and Linguistic features (see Section 3.5). It is worth noting that although LF provides an accuracy that is slightly better that the combination of LF + BiLSTM (55.3 vs. 54.2), we chose LF + BiLSTM as the final model because we considered that the combination of both feature sets would provide more consistent results in real environments, particularly when dealing with words that do not appear during the training process. This decision is commented on in greater depth in Section 4.1.

In the last iteration, we obtained a vector composed of TF–IDF positive, TF–IDF neutral and TF–IDF negative and another vector with TF–IDF–e positive, TF–IDF–e neutral and TF–IDF–e negative. We normalised the vectors in a range 0–1 in order to discover the degree to which the sentiment of each concept appeared in the corpus. Table 5 shows the degree of positive, neutral and negative sentiment for the concepts obtained with the TF–IDF, and the TF–IDF–e formula, ordered by the TF–IDE–e score.

Table 5 shows that the most popular subtopic is *fever*, followed by *aedesBorne*, which is a species of mosquito that transmits Dengue fever, yellow fever, the Zika virus, and Chikungunya. It also shows that the most popular terms are from the ontology concept *Symptom*, such as *fever*, *headhache*, and *conjunctivitis*.

The last step we consisted of sorting the normalised scores in order to see which subtopics were mostly considered as positive, neutral and negative. Figs. 12, 13, and 14 shows the twenty aspects that were classified as being mostly positive, neutral and negative respectively.

Finally, we developed a web interface in order to allow users to show the sentiment and subtopics of new tweets. This interface enables its users to see the last tweets compiled by a crawler that requests Twitter in intervals so as to obtain the last tweets regarding a list of predefined infectious diseases. The system extracts the statistical and linguistic features of each new tweet compiled, and uses them as input for the sentiment analysis model in order to guess its sentiment. The users can obtain information about the semantic annotated concepts for each tweet, along with how each concept influences the other concepts based on the TF–IDF–e. This interface is shown in Fig. 15

Table 5
TF–IDF and TF–IDF–e score for each concept in the ontology.

Class	TF	TF–IDF–e	vector TF–IDF	vector TF–IDF–e
Influenza	143	26,024	(0.23, 0.46, 0.31)	(0.35, 0.33, 0.32)
Chikungunya	1 426	24,801	(0.62, 0.23, 0.15)	(0.49, 0.28, 0.23)
Measles	33	24,543	(0.12, 0.27, 0.61)	(0.37, 0.31, 0.32)
YellowFever	80	24,468	(0.32, 0.31, 0.37)	(0.38, 0.31, 0.31)
Pain	149	24,225	(0.19, 0.54, 0.27)	(0.36, 0.35, 0.29)
dengueDisease	18,993	24,068	(0.32, 0.34, 0.35)	(0.33, 0.34, 0.34)
zikaFever	4 441	23,088	(0.39, 0.30, 0.31)	(0.38, 0.31, 0.30)
Headache	27	6 456	(0.27, 0.51, 0.22)	(0.43, 0.30, 0.27)
aedesBorne	142	6 417	(0.53, 0.26, 0.22)	(0.42, 0.32, 0.26)
mosquitoBorne	1 576	6 052	(0.46, 0.36, 0.18)	(0.45, 0.34, 0.21)
Symptom	401	5 642	(0.37, 0.29, 0.34)	(0.35, 0.31, 0.34)
Fever	270	5 428	(0.25, 0.40, 0.35)	(0.37, 0.32, 0.32)
Conjunctivitis	23	4 861	(0.42, 0.37, 0.21)	(0.39, 0.31, 0.30)
Person	494	3 092	(0.17, 0.25, 0.58)	(0.27, 0.28, 0.45)
Disease	1 619	1 769	(0.29, 0.31, 0.40)	(0.29, 0.31, 0.40)
insectBorne	38	1 706	(0.53, 0.40, 0.06)	(0.47, 0.36, 0.17)
Femenine	359	1 690	(0.24, 0.30, 0.46)	(0.24, 0.30, 0.46)
Masculine	861	1 677	(0.27, 0.34, 0.39)	(0.26, 0.33, 0.41)
Region	605	1 645	(0.24, 0.25, 0.51)	(0.23, 0.24, 0.52)
familyGroup	219	1 323	(0.58, 0.29, 0.13)	(0.41, 0.28, 0.32)
Pacient	361	1322	(0.28, 0.27, 0.44)	(0.28, 0.27, 0.45)
populationGroup	257	1 321	(0.48, 0.33, 0.19)	(0.38, 0.30, 0.32)
Infant	591	1 141	(0.15, 0.21, 0.64)	(0.16, 0.22, 0.62)
Colombia	100	1 128	(0.24, 0.17, 0.59)	(0.28, 0.25, 0.48)
Adult	38	1 121	(0.27, 0.25, 0.49)	(0.20, 0.24, 0.56)

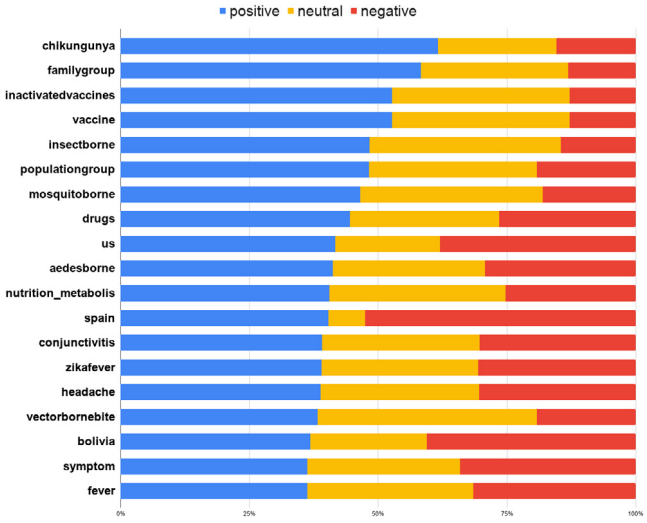

Fig. 12. Aspects most frequently classified as positive with TF–IDF–e.

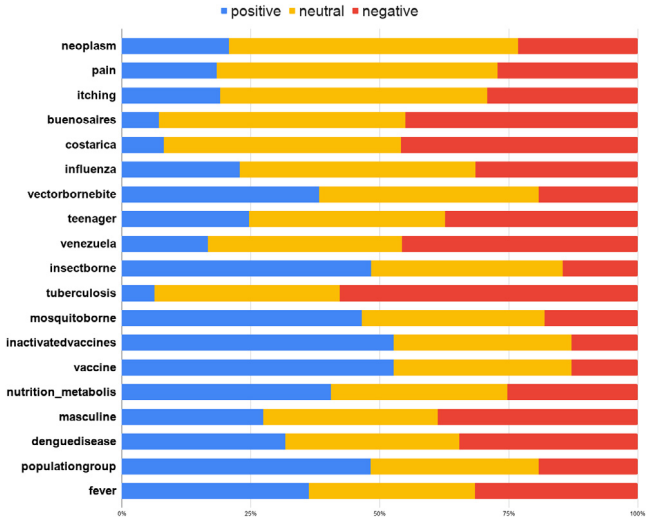

Fig. 13. Aspects most frequently classified as neutral more times with TF–IDF–e.

4. Analysis

Having described our proposal, we shall now provide some insights into the evaluation of the results. This section analyses the results attained after carrying out the Sentiment-Analysis model evaluation (see Section 4.1), and the Linguistic feature analysis (see Section 4.2).

4.1. Sentiment-analysis model evaluation

We evaluated the Sentiment-Analysis model by testing several deep-learning models, such as feed-forward neural networks, convolutional neural-networks, and recurrent neural networks. We also tested different combinations for text representation based on linguistic features, statistical features with word-embeddings, and a combination of both. We additionally evaluated these models from a multi-class perspective, and employed two binary classifiers to discern between neutral and non-neutral tweets, after which we classified the non-neutral tweets as positive and negative.

With regard to the model evaluation (see Section 3.5), Table 1 shows that the combination of LF significantly improves the results attained by the deep-learning models based on RNN (LSTM and BiLSTM) whereas there is no difference when combining LF with CNN. The best improvement occurs in the BiLSTM + LF model, which increases the accuracy by 11.3% with respect BiLSTM. This finding can also be found in Table 2, which LF also increases the accuracy of the RNN when comparing neutral and non-neutral tweets. Moreover, the CNN also benefits from the combination of LF by increasing its accuracy. These findings suggest that LF makes it easier to distinguish between neutral and non-neutral tweets. In Table 3, however, it will be noted that deep-learning models work fine in isolation, as is the case of CNN, which achieves the best result. Moreover, the accuracy is less significant when comparing LSTM with LSTM + LF (4.97%)

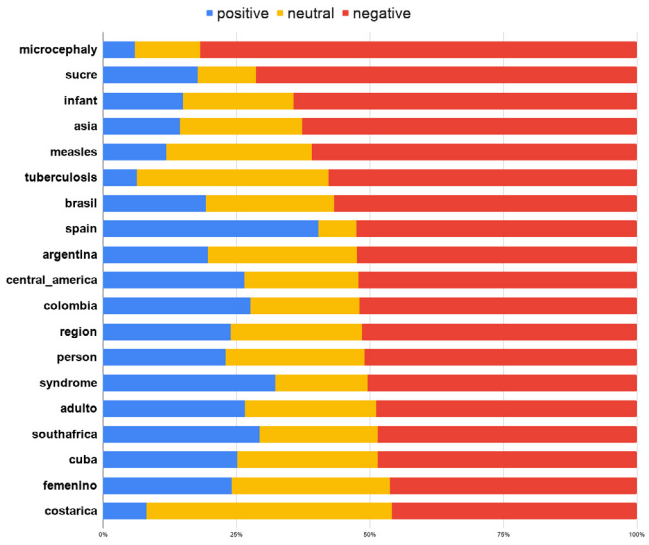

Fig. 14. Aspects most frequently classified as negative more times with TF–IDF–e.

and BiLSTM + LF with BiLSTM (7.67%) than the multi-class classification problem, owing to the high reliability of deep learning methods in sentiment analysis.

As mentioned earlier in Table 1 the classification with linguistic features (LF) in isolation obtained the best result, followed by the combination of linguistic features with BiLSTM (BiLSTM + LF), which attained almost the same accuracy as LF. It is clear that the models that do not use linguistic characteristics obtain much worse results. When observing Table 2 as regards the classification of neutral and non-neutral tweets, the difference between LF and CNN + LF is quite small and the difference between BiLSTM + LF with BiLSTM and between LSTM + LF with LSTM are far superior, thus enabling to assume that the use of LF allows an effective identification of neutral tweets when compared to positive or negative tweets. Furthermore, when comparing positive and negative tweets (see Table 3, it will be observed that RNN and CNN achieved good results on their own, but that LF improves the results only for the RNN models.

As stated previously when describing the Concept-based Opinion summarisation process (see Section 3.7), we decided to use the multiclass classification based on BiLSTM and Linguistic features, despite the fact that LF provided a slightly better accuracy. The justification for this decision is that we consider that the combination of both feature sets will provide more consistent results in real environments, particularly when dealing with words that do not appear during the training process.

When the neural networks were evaluated, our attention was drawn to the fact that regular LSTM attained much worse results than BiLSTM. The main improvement as regards BiLSTM over LSTM is that BiLSTM preserves information from the past and the future. The difference between the best (LF) and the worst (LSTM) deep-learning models is 8%. During our experimentation, we tested classical machine-learning models (not showed) such as Random-Forest [85], and Sequential Minimal Optimisations (SMO) [86], and the results were similar. Since different models have been tested, it is possible that there are contradictory elements in the corpus that make it difficult to exceed this upper bound. Mozetič et al. [87] argue that the quality of a manually labelled corpus is conditioned by the subjective judgement of the annotators. The same authors measured the degree of self-agreement of annotators along with the inter-annotator agreements among all the annotators, for a set of corpora in different languages. They discovered that the ratings for some cases (including Spanish) were low. They proposed that annotators should be monitored throughout the annotation process. In our case, the manual annotation was monitored weekly to ensure that all volunteers performed a similar number of ratings to avoid bias. However, it is possible that the self-agreement of the annotators decreases over time because they observe and learn from the documents they have already labelled, which may influence their own criteria.

With regard to the aspects that have been most frequently classified as positive with TF–IDF–e (see Fig. 12), it will be noted that the subtopic that received the most direct and indirect positive comments was *chikungunya* disease. Another infectious disease, Zika fever, also appears in this list, but with less positive comments. Other types of aspects that were labelled as positive were those referring to population groups, such as family groups. In the case of neutral sentiments (see Fig. 13), the subtopic that received most ratings was neoplasm, which is an excessive growth of tissue. Symptoms, such as *pain*, and *itching* appear in the second and third position respectively, along with the regions of Argentina and Venezuela. Finally, the aspect with most negative sentiments (see Fig. 14) was microcephaly, which is an abnormal development of the brain and is related with maternal Zika virus infection [88].

4.2. Linguistic feature analysis

In order to determine the most discriminating linguistic features of the linguistic features, we calculated the Information Gain (IG). IG is used as metric for feature selection, by evaluating the mutual information gain between the linguistic feature in the context of the class [89]. It is also used in ensemble methods, such as decision trees, in order to determine when a new branch must be created. Fig. 16 shows the 20 best features with major-information for the multi-class version of the corpus.

As Fig. 16 shows, the most discriminating feature is *numerals*, which refers to words and symbols that denote to a numerical quantity. This feature includes only cardinal numbers, excluding the ordinals. In this corpus, cardinals are used for media sharing in order to report new cases of infectious diseases, and these tweets are usually classified with negative sentiment.

The length of the words, represented by the number of words longer than six characters *words-longer-6tr*, and the average number of words *words-length-avg* are the second and third most discriminating features, respectively. Next is the Flesch-Szigriszt (INFLESZ) readability formula [90], which assigns a score to the texts based on the number of syllables per word (*syllabes-per-word*) followed by the average number of words per sentence (*word-length-average*).

With regard to the PoS, a high number of linguistic features appear in the top linguistic features, including: (1) the percentage of prepositions (*prepositions*); (2) the percentage of nous (*grammatical-pos-nouns-common*); (3) the percentage of conjunctions (*grammatical-pos-conjunctions*), and (4) different verb categories (*verbs-nonfinite-infinitive*, *verbs-transitive*, and *verbs-inflection-irregular*).

Apart from pure linguistic features, hyperlinks *twitter-urls* are also a strong indicator of the polarity of the tweet, as tweets are commonly used for the purpose of sharing news [65]. Another discriminating feature is casual communication *colloquialisms*, which is a functional style of speech that is characterised by the usage of interjections and other expressive devices. Colloquial language is not very common in the media, and this linguistic feature, therefore, identifies many of the tweets written by normal users and not by news media.

Relativity-space is a set of words and expressions related to the position and direction of objects or events. It can say that, for

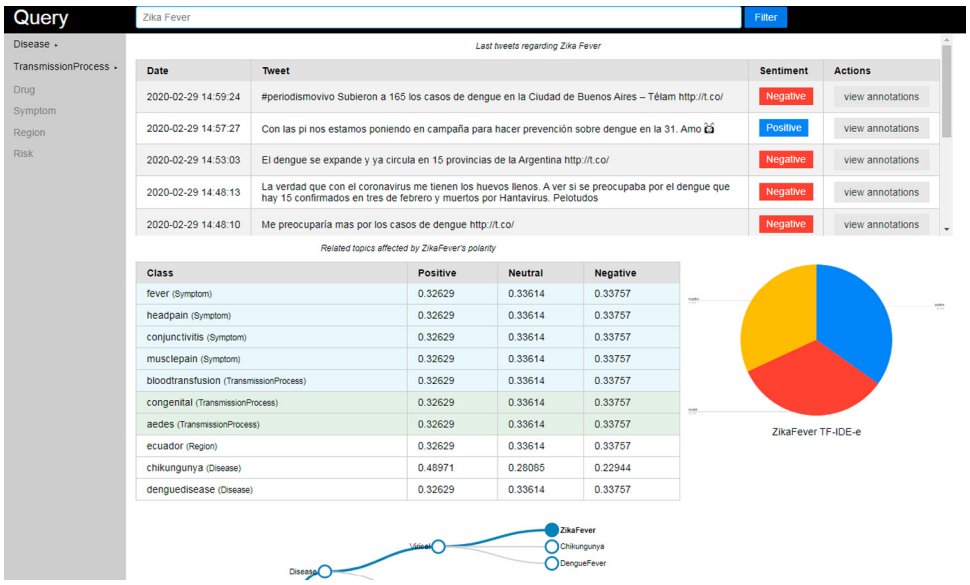

Fig. 15. User interface.

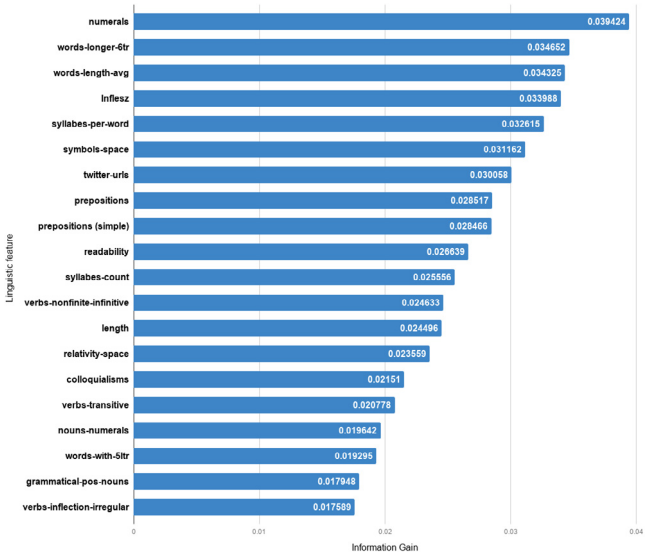

Fig. 16. Information gain of the twenty most discriminating linguistic features.

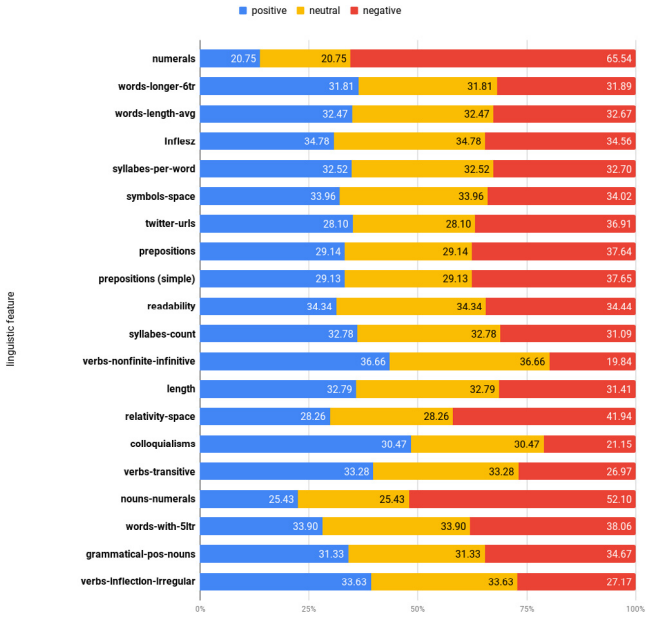

Fig. 17. Mean average of the twenty most discriminating linguistic features according to their sentiment.

example, one object is smaller than other. In this respect, there is a relationship between the microcephaly phenotype, caused by the Zika virus, with terms such as small, smaller, or diminutive.

In order to determine the degree to which each linguistic feature is related to each sentiment, we obtained the frequency of the best 20 linguistic features, which is shown in Fig. 17. Note that the values of this table have been normalised, signifying that each bar represents the extent to which each feature appears in positive, neutral or negative documents, as a percentage. It is possible to verify that the percentage of cardinal numerals (*numerals*) appears more frequently in tweets classified as negative (on 65.54% occasions), while cardinals appear in only the 13.71% of positive tweets and 20.75% of neutral tweets. With regard the use of *colloquialisms*, they were mostly identified in positive (48.38%) and neutral tweets (30.47%).

5. Conclusions and further work

In this paper, we have performed an aspect-based sentiment analysis based on ontologies for infodemiology. We have specifically described a use-case regarding infectious diseases in Latin America in texts written in Spanish whose aspects were extracted from a domain ontology, and whose sentiments were extracted by applying deep-learning models with word-embeddings and linguistic features.

Several means could be employed to improve our proposal. During the corpus compilation stage, we confronted the issue of identifying very similar tweets that may vary as regards only a few words that do not change the meaning of the purpose of the tweet. Although we discarded retweets and exact tweets,

we shall explore the reliability of applying lexical and semantic similarity measures in order to discard very similar tweets. Furthermore, and with regard to the corpus, it is not very large and we shall, therefore, continue to compile tweets concerning Zika and Dengue, along with other infectious diseases, such as Influenza or Covid-19.

We also intend to improve the aspect extraction by including techniques for aspect disambiguation rather than relying on lists of terms and regular expressions. With regard to sentiment classification, we shall include attention mechanisms [91] and shall explore the reliability of training our classifier by including documents from other Spanish sentiment datasets from other domains such as TASS dataset [51,92].

The development of the ontology, and the corresponding translation of some concepts into Spanish, was a very time-consuming task. In this respect, we are exploring the reliability of using automatic tools and thesaurus for the automatic translation of the concepts of the ontology so as to automatically build the regular expression in order to match the terms with those in Spanish. This feature will facilitate the internationalisation of our proposal for other languages.

Finally, the semantic distance and semantic relatedness used during the experimentation stage do not distinguish between the type of relationship between two concepts. For example, the distance between *Zika* with *Infectious Diseases* through the *hasParent* relationship, and the distance between *Zika* and *Fever* through the *hasSymptom* relationship, is the same. During our research, we considered the possibility of applying a weight factor to each object property in order to strengthen or decrease certain relationships. However, as we did not find a clear criterion with which to specify each weight factor, we left them parameterisable, thus enabling them to be adjusted according to the needs of each user. This suggests a further research line for the automatic calculation of weights based on counting the number of incoming and outgoing relationships of each concept in the ontology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER/ERDF) through projects KBS4FIA (TIN2016-76323-R) and LaTe4PSP (PID2019-107652RB-I00). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the Doctorado industrial programme.

References

[1] A. Wilkinson, M. Leach, Briefing: Ebola–myths, realities, and structural violence, Afr. Aff. 114 (454) (2015) 136–148.
[2] A. Gesser-Edelsburg, Y. Shir-Raz, S. Hayek, O.S.-B. Lev, What does the public know about ebola? the public's risk perceptions regarding the current ebola outbreak in an as-yet unaffected country, Amer. J. Infect. Control 43 (7) (2015) 669–675.
[3] P. Vinck, P.N. Pham, K.K. Bindu, J. Bedford, E.J. Nilles, Institutional trust and misinformation in the response to the 2018–19 ebola outbreak in north kivu, dr congo: a population-based survey, Lancet Infect. Dis. 19 (5) (2019) 529–536.
[4] L. Tang, B. Bie, S.-E. Park, D. Zhi, Social media and outbreaks of emerging infectious diseases: A systematic review of literature, Amer. J. Infect. Control 46 (9) (2018) 962–972.
[5] O. Serban, N. Thapen, B. Maginnis, C. Hankin, V. Foot, Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification, Inf. Process. Manage. 56 (3) (2019) 1166–1184, http://dx.doi.org/10.1016/j.ipm.2018.04.011.
[6] Y. Zhang, L. Yakob, M.B. Bonsall, W. Hu, Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data, Sci. Rep. 9 (1) (2019) 1–7.
[7] G. Eysenbach, Infodemiology and infoveillance: tracking online health information and cyberbehavior for public health, Amer. J. Prev. Med. 40 (5) (2011) S154–S158.
[8] G. Eysenbach, Infodemiology and nfoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet, J. Med. Internet Res. 11 (1) (2009) e11.
[9] M. Salathé, D.Q. Vu, S. Khandelwal, D.R. Hunter, The dynamics of health behavior sentiments on a large online social network, EPJ Data Sci. 2 (1) (2013) 4.
[10] P. Törnberg, Echo chambers and viral misinformation: Modeling fake news as complex contagion, PLoS One 13 (9) (2018).
[11] J. Brainard, P.R. Hunter, Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus, SIMULATION 96 (4) (2020) 365–374, http://dx.doi.org/10.1177/0037549719885021.
[12] B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Hum. Lang. Technol. 5 (1) (2012) 1–167.
[13] O. Apolinardo-Arzube, J.A. García-Díaz, J. Medina-Moreira, H. Luna-Aveiga, R. Valencia-García, Evaluating information-retrieval models and machine-learning classifiers for measuring the social perception towards infectious diseases, Appl. Sci. 9 (14) (2019) 2858.
[14] K. Howells, A. Ertugan, Applying fuzzy logic for sentiment analysis of social media network data in marketing, Procedia Comput. Sci. 120 (2017) 664–670.
[15] M. Geetha, P. Singha, S. Sinha, Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis, Tour. Manag. 61 (2017) 43–54.
[16] M. Roccetti, G. Marfia, P. Salomoni, C. Prandi, R.M. Zagari, F.L.G. Kengni, F. Bazzoli, M. Montagnani, Attitudes of crohn's disease patients: Infodemiology case study and sentiment analysis of facebook and twitter posts, JMIR Public Health Surveill. 3 (3) (2017) e51.
[17] D. Robinson, Z. Zhang, J. Tepper, Hate speech detection on twitter: feature engineering vs feature selection, in: European Semantic Web Conference, Springer, 2018, pp. 46–49.
[18] W. Herzallah, H. Faris, O. Adwan, Feature engineering for detecting spammers on twitter: Modelling and analysis, J. Inf. Sci. 44 (2) (2018) 230–247.
[19] G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, A. Mittal, Combining neural, statistical and external features for fake news stance identification, in: The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018, 2018, pp. 1353–1357, http://dx.doi.org/10.1145/3184558.3191577.
[20] P. Ray, A. Chakrabarti, A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis, Appl. Comput. Inform. (2019).
[21] K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Trans. Knowl. Data Eng. 28 (3) (2015) 813–830.
[22] M. del Pilar Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M.Á. Rodríguez-García, R. Valencia-García, Sentiment analysis on tweets about diabetes: An aspect-level approach, Comp. Math. Methods Med. 2017 (2017) 5140631:1–5140631:9, http://dx.doi.org/10.1155/2017/5140631.
[23] M. del Pilar Salas-Zárate, R. Valencia-García, A. Ruiz-Martínez, R.C. Palacios, Feature-based opinion mining in financial news: An ontology-driven approach, J. Inf. Sci. 43 (4) (2017) 458–479, http://dx.doi.org/10.1177/0165551516645528.
[24] A. Konjengbam, N. Dewangan, N. Kumar, M. Singh, Aspect ontology based review exploration, Electron. Commer. Res. Appl. 30 (2018) 62–71, http://dx.doi.org/10.1016/j.elerap.2018.05.006.
[25] M. del Pilar Salas-Zárate, G. Alor-Hernández, J.L. Sánchez-Cervantes, M.A. Paredes-Valverde, J.L. García-Alcaraz, R. Valencia-García, Review of english literature on figurative language applied to social networks, Knowl. Inf. Syst. (2019) 1–33.
[26] D. Wu, H. Wang, Reviewminer: An aspect-based review analytics system, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1285–1288, http://dx.doi.org/10.1145/3077136.3084148.
[27] B. Agarwal, N. Mittal, Machine learning approach for sentiment analysis, in: Prominent Feature Extraction for Sentiment Analysis, Springer, 2016, pp. 21–45.
[28] S. Wang, C.D. Manning, Baselines and bigrams: Simple, good sentiment and topic classification, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 90–94.
[29] N.A.M. Ariff, A. Abdullah, M.F. Nasrudin, Experimental approach based on ensemble and frequent itemset mining for image spam filtering, J. Telecommun. Electron. Comput. Eng. 10 (1–5) (2018) 121–126.

[30] S. Aiyar, N.P. Shetty, N-gram assisted youtube spam comment detection, Procedia Comput. Sci. 132 (2018) 174–182.
[31] Z. Yun-tao, G. Ling, W. Yong-cheng, An improved TF-IDF approach for text classification, J. Zhejiang Univ.-Sci. A 6 (1) (2005) 49–55.
[32] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.
[33] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[34] M. Pagliardini, P. Gupta, M. Jaggi, Unsupervised learning of sentence embeddings using compositional n-gram features, in: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, Vol. 1, 2018, pp. 528–540.
[35] S. Arora, Y. Liang, T. Ma, A simple but tough-to-beat baseline for sentence embeddings, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017, URL https://openreview.net/forum?id=SyK00v5xx.
[36] N.S. Tawfik, M.R. Spruit, Evaluating sentence representations for biomedical text: Methods and experimental results, J. Biomed. Inform. 104 (2020) 103396, http://dx.doi.org/10.1016/j.jbi.2020.103396, URL http://www.sciencedirect.com/science/article/pii/S1532046420300253.
[37] J. Sylak-Glassman, C. Kirov, D. Yarowsky, R. Que, A language-independent feature schema for inflectional morphology, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 674–680.
[38] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (1) (2010) 24–54.
[39] B. O'dea, M.E. Larsen, P.J. Batterham, A.L. Calear, H. Christensen, A linguistic analysis of suicide-related twitter posts., Crisis: J. Crisis Interv. Suicide Prev. 38 (5) (2017) 319.
[40] V.K. Singh, S. Ghosh, C. Jose, Toward multimodal cyberbullying detection, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2017, pp. 2090–2099.
[41] M. del Pilar Salas-Zárate, M.A. Paredes-Valverde, M.Á. Rodríguez-García, R. Valencia-García, G. Alor-Hernández, Automatic detection of satire in twitter: A psycholinguistic-based approach, Knowl.-Based Syst. 128 (2017) 20–33, http://dx.doi.org/10.1016/j.knosys.2017.04.009.
[42] N. Ramirez-Esparza, C.K. Chung, E. Kacewicz, J.W. Pennebaker, The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches, in: Proceedings of the Second International Conference on Weblogs and Social Media, ICWSM 2008, Seattle, Washington, USA, March 30 - April 2, 2008, The AAAI Press, 2008, URL http://www.aaai.org/Library/ICWSM/2008/icwsm08-020.php.
[43] J.F. Sánchez-Rada, C.A. Iglesias, Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison, Inf. Fusion 52 (2019) 344–356, http://dx.doi.org/10.1016/j.inffus.2019.05.003, URL http://www.sciencedirect.com/science/article/pii/S1566253518308704.
[44] D. Bamman, N.A. Smith, Contextualized sarcasm detection on twitter, in: Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015, 2015, pp. 574–577, URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10538.
[45] D.A. Alboaneen, H. Tianfield, Y. Zhang, Sentiment analysis via multi-layer perceptron trained by meta-heuristic optimisation, in: 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 4630–4635.
[46] Y. Kim, Convolutional neural networks for sentence classification, 2014, arXiv preprint arXiv:1408.5882.
[47] D. Zeng, Y. Dai, F. Li, J. Wang, A.K. Sangaiah, Aspect based sentiment analysis by a linguistically regularized cnn with gated mechanism, J. Intell. Fuzzy Systems 36 (5) (2019) 3971–3980.
[48] S. Ruder, P. Ghaffari, J.G. Breslin, INSIGHT-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis, 2016, CoRR, abs/1609.02748. arXiv:1609.02748. URL http://arxiv.org/abs/1609.02748.
[49] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 606–615.
[50] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 5876–5883, URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16541.
[51] E. Martínez Cámara, Y. Almeida Cruz, M.C. Díaz Galiano, S. Estévez-Velarde, M.Á. García Cumbreras, M. García Vega, Y. Gutiérrez, A. Montejo Ráez, A. Montoyo, R. Muñoz, et al., Overview of TASS 2018: Opinions, Health and Emotions, Sun SITE Central Europe, 2018.
[52] D. Vilares, Y. Doval, M.A. Alonso, C. Gómez-Rodríguez, Lys at tass 2015: Deep learning experiments for sentiment analysis on spanish tweets, in: TASS@ SEPLN, 2015, pp. 47–52.
[53] M.S. Akhtar, A. Kumar, D. Ghosal, A. Ekbal, P. Bhattacharyya, A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 540–546.
[54] M.M. Trusca, D. Wassenberg, F. Frasincar, R. Dekker, A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention, 2020, arXiv preprint arXiv:2004.08673.
[55] T.A. Rana, Y.-N. Cheah, A two-fold rule-based model for aspect extraction, Expert Syst. Appl. 89 (2017) 273–285, http://dx.doi.org/10.1016/j.eswa.2017.07.047, URL http://www.sciencedirect.com/science/article/pii/S0957417417305249.
[56] T.A. Rana, Y.-N. Cheah, Improving aspect extraction using aspect frequency and semantic similarity-based approach for aspect-based sentiment analysis, in: International Conference on Computing and Information Technology, Springer, 2017, pp. 317–326.
[57] G. Zhu, C.A. Iglesias, Computing semantic similarity of concepts in knowledge graphs, IEEE Trans. Knowl. Data Eng. 29 (1) (2017) 72–85.
[58] M. Dragoni, S. Poria, E. Cambria, Ontosenticnet: A commonsense ontology for sentiment analysis, IEEE Intell. Syst. 33 (3) (2018) 77–85.
[59] R. Studer, V.R. Benjamins, D. Fensel, Knowledge engineering: principles and methods, Data Knowl. Eng. 25 (1–2) (1998) 161–197.
[60] L. Derczynski, D. Maynard, G. Rizzo, M. Van Erp, G. Gorrell, R. Troncy, J. Petrak, K. Bontcheva, Analysis of named entity recognition and linking for tweets, Inf. Process. Manage. 51 (2) (2015) 32–49.
[61] L. Reeve, H. Han, Survey of semantic annotation platforms, in: Proceedings of the ACM Symposium on Applied Computing, Vol. 2, 2005, pp. 1634–1638, http://dx.doi.org/10.1145/1066677.1067049.
[62] X. Liao, Z. Zhao, Unsupervised approaches for textual semantic annotation, a survey, ACM Comput. Surv. 52 (4) (2019) 1–45.
[63] X.H. Wang, D.Q. Zhang, T. Gu, H.K. Pung, Ontology based context modeling and reasoning using OWL, in: IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second, Ieee, 2004, pp. 18–22.
[64] F. Couto, A. Lamurias, Semantic similarity definition, in: Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, Vol. 1–3, 2018, pp. 870–876, http://dx.doi.org/10.1016/B978-0-12-809633-8.20401-9.
[65] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, Vol. 19, 2010, pp. 591–600, http://dx.doi.org/10.1145/1772690.1772751.
[66] S. Goel, A. Anderson, J. Hofman, D.J. Watts, The structural virality of online diffusion, Manage. Sci. 62 (1) (2016) 180–196.
[67] K. Krippendorff, Reliability in content analysis: Some common misconceptions and recommendations, Hum. Commun. Res. 30 (3) (2004) 411–433.
[68] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: LREC 2018 - 11th International Conference on Language Resources and Evaluation, 2019, pp. 52–55.
[69] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, 2018, arXiv preprint arXiv:1802.06893.
[70] A.I. Noskova, L.G. Gazizova, D.O. Estrella, The problem of national and cultural semantics of lexical units in spanish (on material of venezuelan and nicaraguan words reflecting forms of work), Rev. Publ. 4 (13 (2)) (2017) 215–224.
[71] P.M. Carter, A. Lynch, Multilingual miami: Current trends in sociolinguistic research, Lang. Linguist. Compass 9 (9) (2015) 369–385.
[72] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The stanford coreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 55–60, http://dx.doi.org/10.3115/v1/P14-5010, URL https://www.aclweb.org/anthology/P14-5010.
[73] E. Fersini, E. Messina, F.A. Pozzi, Expressive signals in social media languages to improve polarity detection, Inf. Process. Manage. 52 (1) (2016) 20–35.
[74] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, Tensorflow: Large-scale machine learning on heterogeneous systems, 2015, Software available from tensorflow.org. https://www.tensorflow.org/.
[75] F. Chollet, et al., Keras, 2015, https://keras.io.
[76] M. Koppel, J. Schler, The importance of neutral examples for learning sentiment, Comput. Intell. 22 (2) (2006) 100–109.

[77] M. Fernández-López, A. Gómez-Pérez, N. Juristo, Methontology: from ontological art towards ontological engineering, in: Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering, American Asociation for Artificial Intelligence, Stanford, USA, 1997, pp. 33–40.
[78] L.M. Schriml, C. Arze, S. Nadendla, Y.-W.W. Chang, M. Mazaitis, V. Felix, G. Feng, W.A. Kibbe, Disease ontology: a backbone for disease semantic integration, Nucl. Acids Res. 40 (D1) (2012) D940–D946.
[79] L.G. Cowell, B. Smith, Infectious disease ontology, in: Infectious Disease Informatics, Springer, 2010, pp. 373–395.
[80] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucl. Acids Res. 32 (suppl_1) (2004) D267–D270.
[81] W.A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C.J. Mungall, J.X. Binder, J. Malone, D. Vasant, H. Parkinson, L.M. Schriml, Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data, Nucleic Acids Res. 43 (D1) (2014) D1071–D1078, http://dx.doi.org/10.1093/nar/gku1011, arXiv:https://academic.oup.com/nar/article-pdf/43/D1/D1071/17435884/gku1011.pdf.
[82] H. Knublauch, R.W. Fergerson, N.F. Noy, M.A. Musen, The protégé OWL plugin: An open development environment for semantic web applications, in: International Semantic Web Conference, Springer, 2004, pp. 229–243.
[83] M.Á. Rodríguez-García, R. Valencia-García, F. García-Sánchez, J.J.S. Zapater, Ontology-based annotation and retrieval of services in the cloud, Knowl.-Based Syst. 56 (2014) 15–25, http://dx.doi.org/10.1016/j.knosys.2013.10.006.
[84] E.W. Dijkstra, et al., A note on two problems in connexion with graphs, Numer. Math. 1 (1) (1959) 269–271.
[85] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1, IEEE, 1995, pp. 278–282.
[86] J. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Tech. rep., MSR-TR-98-14, 1998.
[87] I. Mozetič, M. Grčar, J. Smailović, Multilingual twitter sentiment classification: The role of human annotators, PLoS One 11 (5) (2016).
[88] L. Schuler-Faccini, E.M. Ribeiro, I.M. Feitosa, D.D. Horovitz, D.P. Cavalcanti, A. Pessoa, M.J.R. Doriqui, J.I. Neri, J.M. de Pina Neto, H.Y. Wanderley, et al., Possible association between zika virus infection and microcephaly—Brazil, 2015, Morb. Mortal. Weekly Rep. 65 (3) (2016) 59–62.
[89] J. Rogers, S. Gunn, Identifying feature relevance using a random forest, in: International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection", Springer, 2005, pp. 173–184.
[90] S.J. Nassif, K. Wong, J.R. Levi, The índice flesch-szigriszt and spanish lexile analyzer to evaluate spanish patient education materials in otolaryngology, Laryngoscope 128 (1) (2018) E21–E26.
[91] G. Liu, J. Guo, Bidirectional LSTM with attention mechanism and convolutional layer for text classification, Neurocomputing 337 (2019) 325–338.
[92] M.C. Dıaz-Galiano, M. Garcıa-Vega, E. Casasola, L. Chiruzzo, M.Á. Garcıa-Cumbreras, E.M. Cámara, D. Moctezuma, A.M. Ráez, M.A.S. Cabezudo, E. Tellez, et al., Overview of TASS 2019: One more further for the global spanish sentiment analysis corpus, 2019.

José Antonio García-Díaz received the B.Sc. and M.Sc. degrees in computer science from the University of Murcia, Espinardo, Spain. He is currently pursuing the Ph.D. degree in computer science with the University of Murcia where he is a member of the TECNOMOD (Knowledge Modelling, Processing and Management Technologies) Research Group. His research interests include Natural Language Processing and infodemiology.

Mar Cánovas-García received the B.Sc. degree in computer science from the University of Murcia, Espinardo, Spain. He is currently pursuing the M.Sc. entitled New Technologies in Computer Science in the University of Murcia, specialised in Intelligent and knowledge technologies with applications in medicine. Her research interests include Natural Language Processing and Big Data.

Rafael Valencia-García received the B.E., M.Sc., and Ph.D. degrees in Computer Science from the University of Murcia, Espinardo, Spain. He is currently a Full Professor with the Department of Informatics and Systems, University of Murcia. His main research interests are natural language processing, Semantic Web and recommender systems. He has participated in more than 35 research projects. He has published over 150 articles in journals, conferences, and book chapters, 50 of them in JCR-indexed journals. He is the author or coauthor of several books. He has been guest editor of five JCR-indexed journals (CSI, IJSEKE, JRPIT, JUCS, SCP).