Author’s final version

Periñán-Pascual, Carlos and Francisco Arcas-Túnez (2019) “Detecting environmentally-related problems on Twitter”. Biosystems Engineering 177, pp. 31-48.

Detecting environmentally-related problems on Twitter

Carlos PERIÑÁN-PASCUAL[a] and Francisco ARCAS-TÚNEZ[b]
[a]*Universitat Politècnica de València (Spain)*
[b]*Universidad Católica San Antonio de Murcia (Spain)*

Abstract. Social media networks such as Facebook and Twitter can be used as a valuable tool to report on environmentally related problems, e.g. landslides or wildfires, that are about to occur or have just occurred, so that response actions can be promptly executed. The goal of this article is to describe a knowledge-based system that is able to analyse tweets in Spanish to detect a variety of such problems. This research resulted in the implementation of CASPER, a proof-of-concept workbench where multi-domain problem detection has been devised as a two-fold task: topic categorisation and sentiment analysis.

Keywords. Twitter, social sensor, problem detection, topic categorisation, sentiment analysis.

1. Introduction

Sensors are event-driven devices for information pickup. Particularly, sensors are intended to detect changes that occur in the real-world environment, after which a signal is sent to a processor for its analysis and interpretation, thus enabling an event-oriented action to be executed. For example, when an optical gate sensor detects that something is approaching, it sends a signal to a micro-controller, and then the micro-controller opens the gate automatically. In other words, an event happens, a signal is sent and processed, and finally an action is executed. How does a gate sensor resemble a person tweeting a message about a traffic jam on the motorway? Much more than appears at first sight, because both of them are sensors. In the latter case, an event happens (e.g. an accident caused a traffic jam on the motorway), a signal is sent and processed (e.g. one of the drivers stuck in the traffic jam tweeted a message), and finally an action is executed (e.g. the tweet was shared with friends, followers, etc. to advise them to take an alternative route). Therefore, it can be concluded that there are two types of sensors: electronic sensors and social sensors. As explained by Crooks et al. (2013), social sensors operate in a manner comparable to electronic sensors: micro-bloggers play the role of both sensors and microcontrollers, since they collect the information that is important to communicate, whereas the micro-blogging service (e.g. Twitter or Facebook) is the transceiver, since it enables the dissemination of the information. Although social sensors are much noisier than electronic sensors, because "users sometimes misunderstand phenomena, sleep, and are not near a computer" (Sakaki & Matsuo, 2012: 314), social sensors stand out for their low operating cost, wide geographical dissemination and immediate information transfer.

In this context, social media contribute to situation awareness. For example, Twitter provides a real-time channel of communication to report natural disasters. According to Endsley (1995), the three primary components of situation awareness in a given environment are (a) the perception of the elements in the environment, (b) the comprehension of the current situation, and (c) the projection of future actions. The development of software systems that allow these tasks to be performed is actually very useful for citizens and emergency responders. Indeed, situation awareness is recognised as “a critical part of making successful and effective decisions for emergency response” (Yin et al., 2012: 52). Goswami (2016) stated that the objectives of this type of systems can be classified into three major categories: (a) prediction of an event that is about to occur, (b) detection of an event that has just occurred, and (c) management of the event. Our research focuses on the second category—more specifically, on the automatic analysis of Spanish micro-texts from Twitter to devise a protocol of action to manage environmentally related problems, such as overflowing rivers, waste discharge or wildfires, among many others. Therefore, we are interested in constructing a system that can detect messages that contribute to increasing our situation awareness with respect to an environmental hazard that is taking place at the time that the tweet is posted (example 1), ignoring messages that do not contribute to developing an effective notification system for first responders during a disaster (example 2).

(1) Derrame de melaza llegó al río Las Cañas en Apopa.
[Molasses spill reached the Las Cañas River in Apopa.]

(2) Que proyectos de prevencion de mantos acuiferos tiene el Salvador para los proximos 30 años?
[What aquifer protection plans does El Salvador have for the next 30 years?]

In this context, this research led to the design and development of CASPER (CAtegory- and Sentiment-based Problem FindER), which analyses tweets for the automatic detection of user-defined multi-domain problems by following a symbolic approach to topic categorisation and sentiment analysis.[1] The primary goal of this article is to provide detailed insight into the processing that takes place in CASPER. The remainder of this article is organised as follows. Sections 2 and 3 briefly describe some works related to social sensors and the challenge of our research, respectively. Sections 4 and 5 explore the Spanish knowledge resources used in this project and the database of our system, respectively. Section 6 provides an accurate account of our method of problem detection, and Section 7 evaluates the research. Finally, Section 8 presents some conclusions.

2. Related work

The use of social sensors for the development of emergency-response systems has become a relevant research topic over the last decade, where most of these studies have focused on the processing of English texts from a supervised machine-learning approach. In this section, we describe the most significant achievements in this field, whose primary goal has been to detect tweets related to natural disasters, mainly earthquakes, floods, hurricanes, tornados, and wildfires.

Vieweg et al. (2010) analysed the tweets generated during two emergency events (e.g. grassfires and floods) in North America. They concluded that these tweets enhanced situation awareness, so information extraction systems can be really useful both for citizens and for emergency responders. In other words, Twitter can be used as a valuable medium “for ‘harvesting’ information *during* a crisis event to determine what is happening on the ground” (Vieweg et al., 2010: 1079).

[1] CASPER, which has been developed in C# with ASP.NET 4.6 and MySQL Database, is freely accessible from the FunGramKB website (http://www.fungramkb.com/nlp.aspx).

Sakaki et al. (2010, 2013) presented one of the first applications to use Twitter as a medium for social sensors to detect real-time events. They devised a Support Vector Machine (SVM) classifier of tweets based on features such as the keywords in a tweet, the number of words and their context. Moreover, a probabilistic spatio-temporal model was used to find the event location. As a result, they developed a reporting system to promptly notify people of an earthquake in Japan. In the evaluation of the system, Sakaki et al. (2013) experimented with the tweets posted during nineteen months (2009-2011), in which 1,136 earthquakes occurred, and considered different values for the number of positive tweets that were analysed every ten minutes (N_{tweet}). They concluded that there was a trade-off between precision and recall. When N_{tweet} was 10, the system could detect 93% of the earthquakes that were 3 or higher on the JMA (Japan Meteorological Agency) seismic intensity scale, but the precision was very low (0.2), so many false-positive alarms were produced. However, when N_{tweet} was 100, the system could detect only 80% of the earthquakes of that intensity, but 75% of the alarms were correct.

Verma et al. (2011) used natural language processing (NLP) and machine learning techniques in combination with linguistic features such as subjectivity (i.e. subjective or objective), style (i.e. personal or impersonal) and register (i.e. formal or informal) to automatically detect messages that may contribute to situational awareness. They conducted the experiment with four datasets including tweets from different natural disasters, i.e. the Red River floods in 2009 (453 tweets) and 2010 (499 tweets), the Oklahoma grass fires in 2009 (527 tweets), and the Haiti earthquake in 2010 (486 tweets), from which the training and test data were generated. The training data were annotated with the linguistic features. They experimented with Naïve Bayes and Maximum Entropy, with the latter providing the best performance results. Their system was able to achieve accuracy scores ranging from 0.84 to 0.89 when categorizing tweets with a classifier trained on the same type of disaster. However, when the classification of the events was performed with a classifier trained on a different type of disaster (i.e. cross-event classification), accuracy decreased as low as 0.29. Moreover, the experiment could not provide conclusive evidence for the benefit of using the linguistic features in the classification.

Imran et al. (2013a, 2013b) designed a disaster-related information extraction system. First, informative messages were automatically detected, being categorised as caution and advice, casualty and damage, donation and offer, or information source. They experimented with several multi-label Naïve Bayes classifiers. Once the relevant tweets were selected, each message was analysed to decide the type of information to extract with respect to the category of the corresponding tweet, making use of the probabilistic model CRF (conditional random field) for the sequence-labelling task. Finally, the output consisted of "information nuggets", i.e. brief, self-contained information items relevant to disaster response. In the evaluation of the system, two datasets were employed: tweets posted during the tornado that struck Joplin in 2011, and tweets posted during the hurricane Sandy in 2012. The system achieved a precision between 0.8 and 0.9.

Karimi et al. (2013) experimented with multinomial Naïve Bayes and SVM to classify tweets for major types of disasters (e.g. earthquake, flood, fire and storm). From a random sample of 6,500 tweets posted in a range of two years, they created a dataset that covered a variety of disaster types from different locations in the world. In this regard, this work is different from previous research, since it does not rely on data from a specific geographical location or a specific type of event. SVM yielded better classification accuracy (0.73). However, this study also demonstrated that system performance can be affected not only by the type of incident described in the training data but also by the event dissimilarity between the training and test data, with the result that classification accuracy decreased up to 0.6.

Huang and Xiao (2015) developed an inventory of 47 message categories, such as casualty, clean-up, damage, evacuation, repair and rescue, which are commonly found during the different disaster phases (i.e. preparedness, response, impact and recovery). Several classifiers, including K-Nearest Neighbours (KNN), Naïve Bayes and logistic regression, were trained and tested with a dataset of 8,807 tweets posted during the hurricane Sandy. Logistic regression showed the best results, with an overall precision of 0.65 on average. The results also demonstrated the effects of unbalanced training datasets. In general, better precision is achieved for category classification when the number of tweets labelled with that category predominates in the training dataset. Indeed, if the prevalence of a given category is higher than 5%, the precision of the category classification can be higher than 0.8.

3. The challenge

In this research, problem detection is going to be addressed as an issue of classification, being comprised of two complementary tasks: topic categorisation and sentiment analysis. As contrasted by Pang et al. (2002), these two tasks cannot be addressed in the same way, since the latter requires special methods of classification, resulting in a more complex architecture of the system. Indeed, the ultimate goal of sentiment analysis (also known as opinion mining or opinion detection) makes this task play a key role in the process of problem detection. Since topic categorisation is intended to determine what the text is about and sentiment analysis is intended to determine the position that the holder of an opinion expresses about the topic, we developed an integrated environment where topic categorisation is guided by sentiment analysis. The design of such an environment was largely driven by two issues. First, we aim to detect a variety of environmentally related problems rather than processing tweets from a single domain, where Twitter-based detection systems of earthquakes, floods, storms or wildfires have been a recurring research topic in social sensors. Second, we aim to detect not only large-scale disasters, which affect a large number of people, but also problems related to small events, e.g. rubbish mounting up on a stream bank or traffic pollution in a town. Both issues determined our approach to problem detection.

The automatic classification of micro-texts can be conducted from two main approaches: a machine-learning approach (or corpus-based approach), which is usually implemented through a supervised method, and a symbolic approach (or lexicon-based approach), which is grounded on lexicons and rules. A supervised machine-learning method (e.g. KNN, Naïve Bayes or SVM) always requires a training dataset, that is, a collection of (micro-)documents where each instance has been manually annotated as positive or negative with respect to the target event (i.e. the problem). This training dataset should be carefully tagged as well as being sufficiently large and representative. For example, Sidorov et al. (2013) recommended having a training dataset containing at least 3,000 tweets. This requirement conflicts with not only the small volume of tweets that are posted during small events but also the development of a multi-domain system like ours, which was intended to classify new tweets on the ground of dynamically created categories of environmentally related problems. Thus, the effort to expand a given training dataset to fit new categories makes software applicability to new domains a non-trivial task. This fact actually becomes a great challenge for the performance of the system, since "successful results depend to a large extent on developing systems that have been specifically developed for a particular subject domain" (Moreno-Ortiz & Pérez Hernández, 2013: 93). In addition, we should bear in mind that the performance of supervised machine-learning classifiers is fairly good especially when processing takes places with a large document rather than with the few words found in a tweet:

> The success of these methods can be explained by the fact that larger texts contain redundant information, e.g. it does not matter whether a classifier cannot model a negation if the text to be classified contains twenty polar opinions and only one or two contain a negation. (Wiegand et al., 2010: 62)

For all of these reasons, our solution was aimed at dealing with problem detection from a lexicon-based approach.

Finally, there have been many studies about topic categorisation and sentiment analysis in English since the turn of the century (cf. Pang et al., 2002; Kim & Hovy, 2004; Popescu & Etzioni, 2005; Ding et al., 2008; Wiegand et al., 2010; Hogenboom et al., 2011; Young & Soroka, 2012, among many others). However, in comparison with English, there has been a smaller number of studies that have conducted research on the processing of Spanish tweets, most of them focused on sentiment analysis (cf. Aiala et al., 2010; Balbachan & Dell'Era, 2012; Gamallo et al., 2013; Moreno-Ortiz & Pérez Hernández, 2013; Sidorov et al., 2013; Araque et al., 2015; Vilares et al., 2015b; Gambino & Calvo, 2016; Plaza del Arco et al., 2016; Sixto et al., 2016; Jimenez-Zafra et al., 2017; Segura-Bedmar et al., 2017; Pla & Hurtado, 2018) and to a much lesser extent on topic categorisation (cf. Cordobés et al., 2014; Ayala et al., 2015; Vilares et al., 2015a; Khandelwal et al., 2017). Therefore, considering that Spanish is the second most common language used on Facebook and Twitter (Fernández Vítores, 2017), this article represents a priority area of interest for the research community.

4. Knowledge resources for Spanish

Without a doubt, the degree of success of knowledge-based approaches is closely dependent on the quality and coverage of the lexical resources involved in the system. This section describes the Spanish resources from which our lexical data stores were constructed (cf. Section 5).

4.1. *Multilingual Central Repository*

The Multilingual Central Repository (Atserias et al., 2004; Gonzalez-Agirre et al., 2012) integrates wordnets from six languages (i.e. Basque, Catalan, English, Galician, Portuguese and Spanish) following the model proposed by EuroWordNet (Vossen, 1998).[2] In each wordnet, every meaning of a word is linked to a synset (or set of synonyms), where in turn synsets are semantically interconnected. Moreover, the Inter-Lingual-Index allows the connection from words in one language to equivalent translations in any of the other languages by means of the synsets in English WordNet 3.0 (Fellbaum, 1998). In order to provide ontological coherence to the integrated wordnets, this knowledge base has also been enriched with a set of ontologies, such as Top Ontology (Àlvez et al., 2008), WordNet Domains (Magnini & Cavaglià,

[2] The Multilingual Central Repository was downloaded from http://adimen.si.ehu.es/web/MCR/

2000)**¡Error! No se encuentra el origen de la referencia.** and SUMO (Pease et al., 2002).

4.2. *SFU-Review-SP-Neg*

SFU-Review-SP-Neg (Martí et al., 2016) comprises 2,953 sentences that contain at least one negative structure extracted from user comments about a variety of topics: books, cars, computers, films, hotels, mobiles, music and washing machines.[3] This resource focuses on syntactic-level negation, where a negative expression can invert the truth value of a phrase or sentence; the cases of lexical negation (e.g. *dudar* [doubt], *descontento* [unhappy], etc.) are excluded from this resource. In this corpus, negation is annotated as (a) simple, expressed with a single particle in the form of an adverb (e.g. *Nunca han dado problemas* [They have never given rise to problems]), a pronoun (e.g. *Nadie quedará decepcionado* [Nobody will be disappointed]) or a preposition (e.g. *El teléfono está sin cobertura* [The cell phone is out of range]), or (b) complex, expressed with two or more particles— continuous (e.g. *Casi no llega a la reunión* [He almost didn't come to the meeting]) or discontinuous (e.g. *No vino nunca* [She never came]). This corpus has also proved to be useful for detecting expressions that, despite of containing negative particles, do not really convey negation (e.g. *a más no poder* [with might and main] or *ni que decir que* [it goes without saying that]). SFU-Review-SP-Neg has been inspired to a great extent by SFU Review Corpus (Konstantinova et al., 2012), a corpus of 400 product and service reviews annotated with negation and speculation tags.

4.3. *SentiWordNet*

SentiWordNet (Esuli & Sebastiani, 2006; Baccianella et al., 2010) is the result of automatically annotating all synsets in English WordNet 3.0 according to their degrees of positivity, negativity and objectivity.[4] Thus, different senses of the same term may have different opinion-related scores. Each of the three scores ranges from 0 to 1, where the sum of the three scores is 1 for each synset. This lexical resource was devised for supporting sentiment classification. To illustrate, Table 1 shows the scores assigned to the senses of *accident*.

Table 1. The word *accident* in SentiWordNet.

word	synset	positive	negative	objective	gloss
accident	07301336	0	0.75	0.25	an unfortunate mishap; especially one causing damage or injury
accident	07300960	0	0.125	0.875	anything that happens suddenly or by chance without an apparent cause

[3] SFU-Review-SP-Neg was downloaded from http://sinai.ujaen.es/sfu-review-sp-neg/

[4] SentiWordNet was downloaded from http://sentiwordnet.isti.cnr.it/

4.4. *Spanish Emotion Lexicon*

The Spanish Emotion Lexicon (Sidorov et al., 2013; Díaz Rangel et al., 2014) contains 2,036 words that are associated with a PFA (Probability Factor of Affective use) value with respect to at least one of the following emotions: anger, disgust, fear, joy, sadness and surprise.[5] To illustrate, Table 2 shows one example of each category.

Table 2. Some words in the Spanish Emotion Lexicon.

Word	PFA	emotion
accidente [accident]	0.696	fear
apestoso [stinking]	0.899	disgust
atontar [stun]	0.232	surprise
luto [mourning]	0.932	sadness
martirizar [torture]	0.397	anger
ovación [ovation]	0.796	joy

5. The database

From the previous knowledge resources, our database scheme can be partially characterised as follows:

$$KB = \left\{ \begin{array}{c} ABBREVIATIONS\text{: } [\{ABBREVIATION, NGRAM\}], \\ GLOSSES\text{: } [\{SYNSET, GLOSS\}], \\ MODIFIERS\text{: } [\{NGRAM, TYPE, SCOPE\}], \\ NEGATION\text{: } [\{NGRAM, POLARITY, SCOPE\}], \\ POS\text{: } [\{SYNSET, POS\}], \\ RELATIONS\text{: } [\{RELATION, SYNSET1, SYNSET2\}], \\ SENTIMENTS\text{: } [\{NGRAM, POLARITY, POS\}], \\ SYNSETS\text{: } [\{NGRAM, SYNSET\}] \end{array} \right\}$$

The complexity of the actual database design is underspecified in this scheme, which includes only those relations that are relevant for this article.

It is important to note that CASPER primarily deals with topic categorisation and sentiment analysis through two types of simple and complex lexical features—i.e. topic features and sentiment features, respectively. On the one hand, topic features take the form of ngrams (i.e. lexemes or stems) that serve to describe a given environmentally related problem (e.g. drought, flood, landslide, solid waste, wildfire, etc.). Unlike sentiment features, topic features are not pre-defined in the knowledge base but are introduced through a CSV file by the user. On the other hand, most sentiment features are stored in SENTIMENTS, in the form of an ngram (i.e. lexeme or stem), the polarity type (i.e. negative or positive) and the part of speech (POS, i.e. noun, verb, adjective or adverb).

[5] The Spanish Emotion Lexicon was downloaded from http://www.cic.ipn.mx/~sidorov/#SEL

$$SENTIMENTS\colon \begin{bmatrix} \{abandonar, n, v\}, \\ \{abastecer, p, v\}, \\ \{abuso, n, n\}, \\ \{accidente, n, n\}, \\ \dots \end{bmatrix}$$

In CASPER, stemming is performed with the SnowBall Analyser of Lucene.Net (Hatcher, Gospodnetic, & McCandless, 2010).[6]

Wiegand et al. (2010) described sentiment analysis as a task of "polarity classification" because the sentiment of opinions is usually categorised as positive or negative. Therefore, our sentiment features are polar expressions, i.e. sentiment-bearing words and phrases that convey a polarity type (also known as semantic orientation or valence). In terms of opinion mining, SENTIMENTS is the polarity lexicon or affective dictionary, which contains 1,934 non-neutral valenced lexical items. Polar expressions can be found in all open-word classes, e.g. nouns (e.g. *basura* [rubbish]), verbs (e.g. *odiar* [hate]), adjectives (e.g. *defectuoso* [defective]) and adverbs (e.g. *formidablemente* [tremendously]). What is noteworthy is the little agreement among researchers on the score to be assigned to the polarity type. For example, polar expressions have a discrete value of +1 or -1 in Ding et al. (2008) and Gamallo et al. (2013) but a continuous value between +1 and -1 in Rill et al. (2012), between +2 and -2 in Polanyi and Zaenen (2004) or between +5 and -5 in Taboada et al. (2011). In this regard, Polanyi and Zaenen (2004: 107) stated that "characterizing terms in binary terms as either positive or negative […] is too crude". However, Kim and Hovy (2004: 1373) concluded that "the mere presence of negative words is more important than sentiment strength". In CASPER, the polarity lexicon was constructed as follows. On the one hand, positively marked ngrams were extracted from those terms whose positive score is equal to or higher than 0.8 and the negative score is 0 in SentiWordNet. On the other hand, negatively marked ngrams were extracted from those terms that belong to the sentiment dimensions of anger, disgust, fear or sadness in the Spanish Emotion Lexicon. The dataset was also extended with swear words, as well as with complaint words found in a collection of 790 tweets. A manual validation of SENTIMENTS was finally required, since we found a large number of "context-dependent opinion words". As exemplified by Ding et al. (2008: 234), there are words whose polarity type depends on the context in which they appear:

(3) The battery of this camera lasts very long. [*long* is positive]
This program takes a long time to run. [*long* is negative]

Therefore, SENTIMENTS only holds words and phrases whose polarity type is clearly defined by looking at only the polar expression.

SYNSETS, GLOSSES, POS and RELATIONS were built from the Multilingual Central Repository. SYNSETS holds 118,591 synsets that are lexicalised in 91,332 simple and complex ngrams in Spanish, so each lexical unit (i.e. lexeme or stem) is linked to one or more synsets. GLOSSES stores the definitions used to describe the sense represented by each synset. POS holds the grammatical category (i.e. noun, verb, adjective or adverb) of the words linked to every synset. RELATIONS stores the semantic relations that can occur between two synsets; the only relations that are relevant for this research are *x-causes-y* (e.g. *kill-die*), *x-derived_from-y* (e.g. *markedly-*

[6] Lucene.net was downloaded from https://lucenenet.apache.org

marked), *x-has_hyponym-y* (e.g. *gene-sequence*), *x-has_subevent-y* (e.g. *fell-undercut*), *x-near_synonym-y* (e.g. *big-large*), *x-pertains_to-y* (e.g. *genetical-gene*) and *x-related_to-y* (e.g. *organism-organic*). These four datasets look like this:

$$SYNSETS\colon \begin{bmatrix} \{dar\ la\ tabarra, 1766638\}, \\ \{fastidiar, 1766638\}, \\ \{molestar, 1765908\}, \\ \{disgustar, 1765908\}, \\ \ldots \end{bmatrix}$$

$$GLOSSES\colon \begin{bmatrix} \{1766638, worry\ persistently\}, \\ \left\{\begin{matrix} 1765908, disturb\ the\ peace\ of\ mind\ of; afflict\ with\ mental \\ agitation\ or\ distress \end{matrix}\right\}, \\ \ldots \end{bmatrix}$$

$$POS\colon \begin{bmatrix} \{1766638, v\}, \\ \{1765908, v\}, \\ \ldots \end{bmatrix}$$

$$RELATIONS\colon \begin{bmatrix} \{12, 1765908, 1766638\}, \\ \ldots \end{bmatrix}$$

The above samples reveal that, for example, the verbs *dar la tabarra* and *fastidiar* pertain to the synset 1766638, whose meaning is *worry persistently (e.g. nag)*. In turn, these lexical units are hyponyms (i.e. relation 12) of the verbs *molestar* and *disgustar*, which are linked to the synset 1765908, whose meaning is *disturb the peace of mind of; afflict with mental agitation or distress (e.g. worry, vex)*.

NEGATION and MODIFIERS compose the main source of knowledge for valence shifters (Polanyi & Zaenen, 2004), i.e. words and phrases that affect the values of the topic and sentiment attributes of some of the ngrams in the micro-text. NEGATION stores 38 negation cues (e.g. *brillar por su ausencia* [be non-existent], *carente de* [devoid of] or *sin* [without]), i.e. negative words and phrases that neutralise the positive or negative valence of other lexical items. Moreover, the scope of each negation cue must also be determined, i.e. *m* ngrams immediately to its left (*l*) or right (*r*), where *m* is a user-defined variable. This dataset also includes 31 false negation cues (e.g. *sin dejar de* [while continuing to]), i.e. negative expressions that do not really constitute a negation of the proposition. For this reason, negation cues are classified as negative (*n*) or non-negative (*x*). The latter case does not require information about the scope, for example:

$$NEGATION\colon \begin{bmatrix} \{brillar\ por\ su\ ausencia, n, l\}, \\ \{carente\ de, n, r\}, \\ \{sin, n, r\}, \\ \{sin\ dejar\ de, x, -\} \\ \ldots \end{bmatrix}$$

NEGATION resulted from the analysis of the negative particles in SFU-Review-SP-Neg, after which these particles were expanded with synonyms from the Spanish

WordNet. Another type of valence shifters is found in MODIFIERS, which includes 28 intensifiers (*i*) and 10 diminishers (*d*), which increase and decrease, respectively, the degree of polarity of the ngrams to which they modify (e.g. *levemente* [slightly]*, mucho* [many, much, a lot] or *muy* [very]). As in NEGATION, the scope of modifiers must also be determined. However, this time the direction of the scope of some modifiers is context-dependent, that is, the direction of the scope depends on the POS of the words to which they modify, e.g. *levemente* can modify a verb (e.g. *mejoró levemente* [it slightly improved]) or an adjective (e.g. *levemente desenfocado* [slightly blurred]), and *mucho* can modify a verb (e.g. *llueve mucho* [it rains a lot]) or a noun (e.g. *muchos accidentes* [many accidents]); the value *b* is used to describe these cases, like this:

$$MODIFIERS\colon \begin{bmatrix} \{levemente, d, b\}, \\ \{mucho, i, b\}, \\ \{muy, i, r\}, \\ \ldots \end{bmatrix}$$

Finally, ABBREVIATIONS holds 21 abbreviations (and their full forms) that are commonly used in social media (cf. Álvarez, 2011), like this:

$$ABBREVIATIONS\colon \begin{bmatrix} \{cdo, cuando\}, \\ \{pk, porque\}, \\ \{xfa, por\ favor\}, \\ \ldots \end{bmatrix}$$

The next section gives a detailed account of the process of problem detection that takes place in CASPER.

6. Method

6.1. Collecting data

To get a collection of micro-texts, we implemented CORSAIR (Getting miCrO-texts from Rss feedS And twItteR),[7] which scrapes the contents of RSS feeds and Twitter on the basis of user-defined settings, such as a list of RSS feed URLs and/or Twitter hashtags and references to be crawled and the maximum number of micro-texts to be retrieved. In this research, we are concerned only with Twitter accounts commonly used to deal with local environmental issues in El Salvador, including a citizen-based warning system (@alertux) and the official account of the Ministry of Environment and Natural Resources (@MARN_SV), among others. As regards the acquisition of tweets in real time, CORSAIR makes use of Tweetinvi 2.1, a library to access the Twitter API with RESTful web services from .NET applications (e.g. ASP.NET).[8] To overcome the rate limit imposed by Twitter, that is, 180 requests in 15-minute intervals, we chose to employ parallel threads when receiving data. Moreover, duplicate tweets are filtered out by checking the unique MD5 hash generated for each micro-text. Finally, the collection of tweets is downloaded as a CSV file, and those micro-texts that are not related to environmental issues (e.g. aquifers, drought, earthquakes, floods, landslides or waste

[7] CORSAIR, which has been developed in C# with ASP.NET 4.6, is also freely accessible from the FunGramKB website.

[8] Tweetinvi was downloaded from https://github.com/linvi/tweetinvi

discharge) are manually excluded from the dataset. Therefore, it is important to make it clear that, although all the tweets in our data collection focus on some environmental topic, not all of them are actually related to an environmental problem.

6.2 Setting parameters

CASPER allows researchers to model text processing through a variety of parameters connected with mainstream NLP techniques. In particular, the following parameters must be set before the micro-text is processed:

- whether or not the micro-text is spell checked,
- whether or not multi-words are taken into consideration (i.e. bigrams, trigrams and tetragrams), together with the choice of the form of ngrams (i.e. lexeme or stem)
- whether or not the POS of ngrams plays a relevant role, and
- whether or not valence shifters can affect the polarity of ngrams, consisting of (a) negation cues, (b) intensifiers and diminishers, and (c) irrealis markers; the scope of valence shifters is also determined.

Every possible permutation of the values of these parameters actually becomes a valuable chance for the researcher to make a sound decision on the most effective way to analyse the collection of tweets.

6.3. Describing topic categories

As CASPER was designed to detect environmentally related problems referred to in a multi-domain collection of micro-texts, one or several user-defined topic categories (e.g. drought, flood, etc.) must be previously recorded. A new category implies a semi-automatic process of selecting significant features, that is, relevant words and phrases that identify the target event. In fact, this approach is similar to that of Kim and Hovy (2004: 1368), i.e. “to assemble a small amount of seed words by hand, sorted by polarity into two lists—positive and negative—and then to grow this by adding words obtained from WordNet”. However, the difference lies in two core issues. First, Kim and Hovy’s seed terms are intended to contribute only to sentiment analysis; in CASPER, the seed terms are mainly used for topic categorisation. Second, Kim and Hovy expanded seed terms to words that were semantically connected by synonymy or antonymy; in CASPER, seven semantic relations from WordNet are exploited—i.e. *x-causes-y, x-derived_from-y, x-has_hyponym-y, x-has_subevent-y, x-near_synonym-y, x-pertains_to-y* and *x-related_to-y*.

Therefore, the first step in this stage consists in deciding a few seed terms that are representative of each category. Second, the user must also select the relevant synset(s) to which every seed term is linked, so the system explores GLOSSES in search of definitions related to the seed terms. As each seed term becomes a topic feature, this results in the vector $C_i = (f_{i1}, f_{i2}, \dots, f_{ik})$, where every f_{ij} identifies a user-defined feature in the form of a synset assigned to the category C_i. Third, the system recommends a set of additional words and phrases out of these seed terms, so that the researcher can select them to increase the original list of category descriptors. These additional terms are automatically discovered with the aid of SYNSETS and RELATIONS. In particular, the following four-level process of relation-driven expansion takes place in C_i:

- For each f_{ij} in C_i, expand to other synsets involved in the relations *x-has_hyponym-y* and *x-has_subevent-y*, where f_{ij} instantiates *x*.
- For each f_{ij} in C_i, expand to other synsets involved in the relations *x-near_synonym-y* and *x-related_to-y*, where f_{ij} instantiates *x*.
- For each f_{ij} in C_i, expand to other synsets involved in the relation *x-causes-y*, where f_{ij} instantiates *x*.
- For each f_{ij} in C_i, expand to other synsets involved in the relations *x-derived_from-y* and *x-pertains_to-y*, where f_{ij} can instantiate *x* or *y*.

In each sequential level, every synset returned by the expansion is added to C_i as f_{ik+1}, providing that the synset is not already included in C_i. Next, every f_{ij} in C_i is mapped into one or several lexemes, which are presented as a list of suggestions. In this lexical projection, proper names (e.g. people, places or organisations) are not considered. Thus, every topic feature that describes a given category is ultimately represented as an object that is defined with attributes such as the ngram (i.e. either a seed term or a word selected from the list of suggestions), the POS and the polarity (i.e. negative or non-negative).

6.4. Pre-processing data

The tweets are pre-processed to produce clean texts for NLP. First, the Spanish language used in social media is characterised by departing from the commonly accepted standard for written text. This is largely due not only to the need for fast and immediate communication but also to an environment of creative freedom that is not subject to genre norms. Consequently, the pre-processing stage can involve the spelling and typographical standardisation of micro-texts. In this regard, and following Álvarez (2011), three tasks can be performed:

- Reduction of duplicate vowels and consonants to a single one by means of regular expressions (e.g. *¡Qué calooor! -> ¡Qué calor!* [It's so hot!], *Besitossssssss!!! -> Besitos* [Kisses]); we exclude those cases in which the double consonant is valid in Spanish: *cc* (e.g. *accesorio* [accessory]), *ll* (e.g. *llave* [key]), *nn* (e.g. *innato* [innate]) and *rr* (e.g. *perro* [dog]).
- Spell checking with NHunspell,[9] a library that implements Hunspell (Nemeth et al. 2004) for the .NET platform, and the Open Office *Diccionario de corrección ortográfica, separación silábica y sinónimos para el idioma español (España y América Latina)*.[10]
- Transformation of abbreviations into their full-word equivalent with the aid of ABBREVIATIONS (e.g. cdo -> cuando [when], x -> por [by]).

Second, tweets typically show social-media elements in their texts, such as hashtags (i.e. any word starting with #), references (i.e. usernames headed by @) and URL links. In CASPER, these elements are automatically removed from the micro-text by means of regular expressions.

[9] NHunspell was downloaded from https://sourceforge.net/projects/nhunspell/

[10] This Spanish dictionary was downloaded from https://extensions.openoffice.org/en/project/diccionario-de-correccion-ortografica-separacion-silabica-y-sinonimos-para-el-idioma-espanol

It should be noted that neither emoticons (i.e. shorthands for facial expressions consisting of punctuation marks, letters and numbers) nor emojis (i.e. pictographs representing not only facial expressions but also concepts and ideas) were taken into consideration in this research, since we discovered that their presence in messages that are posted during an emergency is practically non-existent. The explanation probably lies in the fact that these visual cues are more commonly found in posts created in casual conversations, where they are used in lieu of the kinesic and prosodic markers that appear in face-to-face communication (Crystal, 2004). However, as the trend towards the use of these visual elements in text-based social media can be different during large-scale disasters and small incidents, we do not rule out the possibility of dealing with them in future research. In this case, a lexicon such as *The Emoji Sentiment Ranking* (Kralj Novak et al., 2015) would be very useful, where each of the 751 most frequently used emojis is mapped to a sentiment score between +1 and -1. As pointed by Miller et al. (2017), however, the emotional content of some emojis is not always clear-cut when considered in isolation, so the challenge will be to determine their sentiment interpretation by analysing the text that accompanies the emojis.

6.5. Processing natural language

Each micro-text is split into sentences, and then each sentence is tokenised and POS-tagged by using the Stanford Log-linear Part-Of-Speech Tagger.[11] At this point, a tweet is represented as the vector $T_m = (w_{m1}, w_{m2}, \ldots w_{mp})$, where w_{mn} represents an object for every word that occurs in the tweet and p is the total number of words. Each w_{mn} is defined with attributes such as the position in the micro-text, the word form, the unit of analysis (i.e. lexeme or stem), the POS, the topic and the sentiment, where the values of the latter two are discovered in the next stage. Word normalisation is essential, because the inflectional paradigm of Spanish is quite rich. In this context, we employed the LemmaGen library (Juršič et al., 2010)[12] for lemmatisation and the SnowBall Analyser for stemming. Moreover, the use of POS-tagged ngrams can serve as a crude form of word sense disambiguation. In fact, the POS plays a relevant role when the processing is based on stems; for example, the noun *paro* [unemployment], the verb *parir* [give birth], the preposition *para* [for] and the adjective *par* [even, with numbers], whose meanings are completely unrelated, happen to have the same stem, i.e. *par*.

6.6. Discovering topic-related ngrams

This stage consists in detecting significant ngrams with respect to the topic (i.e. the target event or category). The weight 1 is assigned to the attribute topic of every w_{mn} in T_m whose ngram is found as an f_{ij} in C_i. Upon the user's choice, the POS of the ngram can also be taken into consideration when the matching between T_m and C_i is involved.

6.7. Discovering sentiment-related ngrams

This stage consists in detecting significant ngrams with respect to the sentiment. Thus, the system attempts to assign the values +1 or -1 (for positively and negatively marked ngrams, respectively) to the attribute sentiment of every w_{mn} in T_m according to

[11] The Stanford POS Tagger was downloaded from https://sergey-tihon.github.io/Stanford.NLP.NET/StanfordPOSTagger.html

[12] LemmaGen was downloaded from http://lemmatise.ijs.si

the polarity of the ngram in SENTIMENTS or, failing that, to the attribute polarity of the ngram found as an f_{ij} in C_i.

6.8. Dealing with valence shifters

After having identified the values of the attributes topic and sentiment of w_{mn}, the next stage is to apply valence shifters to neighbouring words within the micro-text. We categorise valence shifters into (a) neutralisers, which take the form of negation cues and irrealis markers, and (b) modifiers, which take the form of intensifiers and diminishers. Neutralisers make all the ngrams involved in their scope be no longer significant for topic and sentiment, so the values of their attributes topic and sentiment are re-computed to 0. By contrast, modifiers change (i.e. increase or decrease) the degree of polarity of the ngrams involved rather than shifting the valence to 0. In other words, the polarity of the tweet can be strengthened or weakened by the presence of intensifiers and diminishers, respectively. For example, depending on the type of modifier, Polanyi and Zaenen (2004) suggested adding or subtracting 1, and Fernández Anta et al. (2013) preferred to multiply by 3 or 0.5. CASPER re-computes the attribute sentiment according to the coefficients of modification proposed by Fernández Anta et al. (2013).

The sequence of action of these valence shifters is as follows. First, all negation cues (e.g. *no, sin* [without]) in the whole micro-text are detected and applied by making use of NEGATION. Second, all intensifiers (e.g. *bastante* [enough]) and diminishers (e.g. *poco* [little]) in the whole micro-text are detected and applied by making use of MODIFIERS. Third, irrealis markers are taken into consideration. In this regard, problem detection is primarily concerned with realis contexts, which describe events that have happened or are happening as well as states that have been experienced or are being experienced. In this context, Spanish verbs either in subjunctive mood or in future and conditional tenses also play the role of neutralisers.

A critical aspect in this stage of the processing is to determine the scope (or impact region) of valence shifters, i.e. those words in the neighbouring context that are affected by the presence of a given neutraliser or modifier. In this regard, we could have relied on a dependency parser for analysing the syntactic structure of each tweet. For example, FreeLing (Carreras et al., 2004; Padró & Stanilovsky, 2012), an open source library for a variety of NLP tasks, has two types of syntactic dependency parsers for Spanish: a rule-based parser (Atserias et al., 2005) and a probabilistic parser (Carreras, 2007). However, conventional parsers cannot successfully handle the high presence of ungrammatical phenomena that are commonly found during the analysis of tweets. In contrast with NLP tools for English, where Kong et al. (2014) developed a dependency parser designed explicitly for English tweets, there is currently no resource of this kind for Spanish. One option could have been to train a language-independent probabilistic parser, such as MaltParser (Nivre et al., 2007), with a large annotated corpus of Spanish tweets, but this task would have involved a costly manual annotation process. Therefore, the usual way of dealing with the scope of negation is by means of heuristics. Indeed, the common trend has been to deal with a fixed window length of words following and/or preceding a valence shifter, but researchers do not agree with the optimal window size. For example, Hogenboom et al. (2011) concluded that the best approach consists in considering two words following the negation cue. Grefenstette et al. (2004) and Hu and Liu (2004), however, proposed that the scope of negation should be the following five words. In CASPER, the user is given the opportunity to decide the window size of any type of valence shifter. However, the direction of the scope (i.e. following or preceding the valence shifter) is mostly determined by the information provided by the database for each negation cue and modifier. In the case of irrealis markers, the impact is only focused on the first preceding word and the following *n* words, where *n* is the window size. Moreover, whereas neutralisers are applied to all the words within the scope, modifiers act only on the first polar expression that is found in the scope.

6.9. Determining the topic score

As tweets and categories are represented as vectors, a similarity measure may be used to assess the degree of relatedness between both of them. In this context, we used cosine similarity (or normalised dot product) as a measure of semantic distance. In our case, since we deal with the binary values of the attribute topic and the number of distinct topic-related ngrams in the tweet T_m is equal to or less than the number of relevant features in the category C_i, the topic-relatedness function between T_m and C_i can be reduced to the equation (1).

$$rel(T_m, C_i) = \frac{\sum_{n=1}^{p} w_{mn}}{\sqrt{\sum_{n=1}^{p} w_{mn}} \times \sqrt{\sum_{j=1}^{k} f_{ij}}} \qquad (1)$$

Therefore, a tweet is linked to a given category if the similarity score is greater than 0.

6.10. Determining the sentiment score

A simple and commonly used approach to sentiment calculation could have been to sum up the sentiment values of the ngrams in the micro-text and, eventually, to determine the polarity of the text by the sign of the final score. However, we chose to assess the degree of sentiment in a given tweet through a metric originally used to assess political positions in texts. Particularly, Lowe et al. (2011) proposed the logit scale to locate party positions (i.e. left or right) on a continuous scale from the sentences of political texts that were previously coded into these two categories. Indeed, this scaling procedure allows the system to convert the counts of sentiment-coded ngrams in the tweet T_m into a point on the sentiment dimension *S*. In the analysis of political texts, these researchers demonstrated that logit scale is superior to other approaches used to estimate left-right positions, such as saliency (Budge, 1999) or relative proportional difference (Kim & Fording 2002). Consequently, we computed the sentiment score by means of the equation (2).

$$rel(T_m, S)'' = \log(P + 0.5) - \log(|N| + 0.5) \qquad (2)$$

In Lowe et al. (2011), P and N referred to the number of positively and negatively marked sentences in a given text, respectively. However, we interpret P and N as the total value of positively and negatively marked ngrams in T_m, respectively. Therefore, since the value of N is negative or zero in our case, we chose to take the absolute value of N, that is, |N|. The sentiment score is normalised with the equation (3).

$$rel(T_m, S) = \left\{ \begin{matrix} 1 - \frac{1}{\log(-rel(T_m, S)'' + 2)}, & \text{if}\, rel(T_m, S)'' < 0 \\ 0, & \text{otherwise} \end{matrix} \right\} \qquad (3)$$

It is important to note that, as we are only concerned with problem detection, the sentiment-relatedness function automatically assigns the value of zero to those tweets that do not express negative polarity, as stated in the second part of the equation (3).

6.11. Detecting the problem

As explained above, $rel(T_m, C_i)$ and $rel(T_m, S)$ can return zero—in case of irrelevance, or a positive value—when the tweet (T_m) has been tagged with a topic category (C_i) or classified with a negatively marked sentiment (S), respectively. In the end, the final score integrates the values for the topic- and sentiment-relatedness functions in the form of a problem-relatedness perception index (PPI), which results from the geometric mean of both values, as shown in the equation (4).

$$PPI(T_m, C_i, S) = \sqrt{rel(T_m, C_i) * rel(T_m, S)} \qquad (4)$$

In other words, the PPI serves to measure how reliable we can feel that a given tweet deals with a problem about an environmental topic. Therefore, any tweet is classified as a problem only if the above scoring function returns a positive value. In future research, the PPI will also be used to set alert thresholds from which the severity of the problem could be rated as, for example, minor, moderate and critical.

6.12. Visualizing results

As CASPER has been developed as a proof-of-concept system, researchers are provided with a variety of settings that can be configured to test critical functionalities (cf. Section 6.2). The application interface is split into two vertical panels: the input section on the left side and the output section on the right side of the screen. Thus, whereas the left-hand side shows the selected values of the parameters (i.e. spell checking, type of ngram, POS and valence shifters) as well as the category features, the right-hand side displays not only the topic score, sentiment score and PPI of each tweet but also the individual scores assigned to the topic and sentiment features detected in each micro-text. To conclude, Figure 1 illustrates the whole process of problem detection.

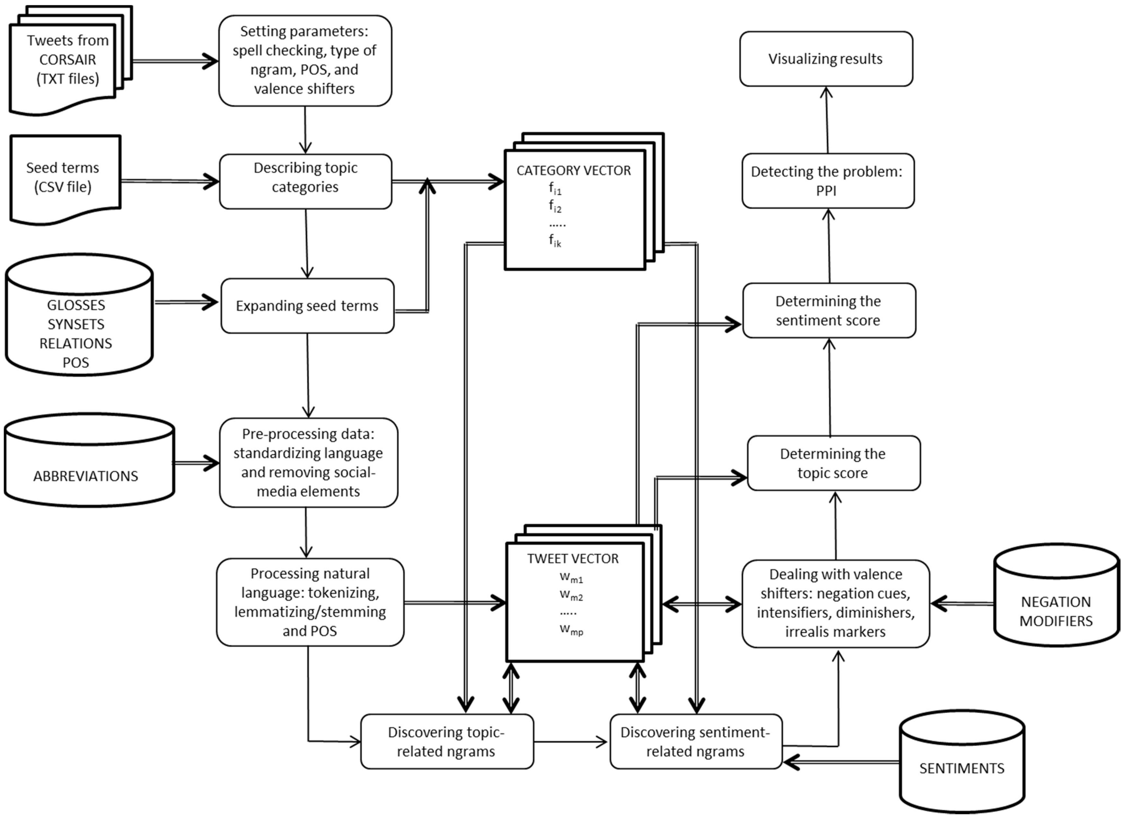
Tweets from CORSAIR (TXT files)
Setting parameters: spell checking, type of ngram, POS, and valence shifters
Visualizing results
Seed terms (CSV file)
Describing topic categories
CATEGORY VECTOR
f_i1
f_i2
.....
f_ik
Detecting the problem: PPI
GLOSSES SYNSETS RELATIONS POS
Expanding seed terms
Determining the sentiment score
ABBREVIATIONS
Pre-processing data: standardizing language and removing social-media elements
Determining the topic score
Processing natural language: tokenizing, lemmatizing/stemming and POS
TWEET VECTOR
w_m1
w_m2
.....
w_mp
Dealing with valence shifters: negation cues, intensifiers, diminishers, irrealis markers
NEGATION MODIFIERS
Discovering topic-related ngrams
Discovering sentiment-related ngrams
SENTIMENTS

Figure 1. Problem detection in CASPER.

7. Evaluation

7.1. Experiments

We evaluated this research by means of two experiments. The goal of the first experiment was to discover the optimal settings for the parameters described in Section 6.2. For this purpose, a corpus of 383 tweets was compiled, where most of their micro-texts were related to a given environmental problem: 108 tweets were categorised as *flood* (FLO), 84 tweets as *drought* (DRO), 74 tweets as *landslide* (LAN) and 66 tweets as *aquifer* (AQU). It should be noted that 19 of these tweets were associated with two categories, namely with both FLO and LAN, as illustrated in the example (4).

(4) 4 personas fallecidas, 102 deslizamientos, 98 viviendas anegadas, 7 casas totalmente destruidas por lluvias en El Salvador
[4 people dead, 102 landslides, 98 houses under water, 7 houses completely destroyed by rain in El Salvador]

Two researchers, who did not take part in the construction of the corpus, were in charge of identifying the seed terms and compiling the final list of topic features for each category, for which the procedure described in Section 6.3 was applied. Table 3 shows the features that were selected for each topic category.

Table 3. Features of the topic categories AQU, DRO, FLO and LAN.

Category	**Features**
AQU	acuífero, agua del subsuelo, agua subterránea
DRO	aridez, árido, déficit hídrico, desértico, desertificación, desertización,

	desierto, escasez de agua, estrés hídrico, falta de agua, quedar sin agua, reseco, secar, seco, sequía
FLO	agua hasta la rodilla, anegar, inundación, inundar
LAN	alud, corrimiento, deslizamiento, desprendimiento

At the first stage of this experiment, we evaluated CASPER through 40 tests, where we could tweak the parameters *Spell checking*, *Ngram form*, *POS* and the *Scope* of the valence shifters to determine the relevance of each parameter during text processing and thus decide the most effective way to analyse micro-texts from our knowledge-based approach. In other words, this experiment is aimed at revealing the extent to which some linguistic factors, such as (a) whether or not the micro-text is spell checked, (b) the choice of the form of ngrams (i.e. lexeme or stem), (c) whether or not the ngrams are labelled with POS tags, and (d) the number of words that determine the impact region of valence shifters, can affect the performance of our problem detection system. In this case, we chose to ignore parameters such as *Multi-word* (i.e. whether or not ngrams can take the form of multi-words) and *Valence shifters* (i.e. whether or not the different types of valence shifters can affect the polarity of ngrams). On the one hand, we acknowledged simple and complex ngrams because several topic features took the form of multi-words (e.g. *agua subterránea* [groundwater] in AQU, *déficit hídrico* [water deficit] in DRO or *agua hasta la rodilla* [up to one's knees in water] in FLO, among others). On the other hand, we started from the premise that negation cues, modifiers and irrealis markers are crucial in determining the polarity of ngrams, thereby contributing to increasing the precision of the system. Indeed, at the second stage of this experiment, we demonstrated that our initial assumption on valence shifters was correct.

The goal of the second experiment was to compare the performance of our knowledge-based approach with that of machine learning. To this end, we compiled a corpus of 683 tweets, which resulted from adding 300 tweets to the previous corpus. On this occasion, the new tweets did not pertain to any of the target categories, although they took the form of comments describing issues about the environmental domain (e.g. waste discharge, earthquakes, etc.). By using a small dataset, we also evaluate the adequacy of machine-learning methods for problem detection with small events, which do not generate a large number of tweets. For data mining research, datasets of this size are also acceptable, since a random sample of roughly 500 tweets is "a manageable set for human annotation and provides sufficient training data for machine-learning classification" (Verma et al., 2011: 387). In this experiment, we used multinomial Naïve Bayes and SVM as the supervised machine-learning algorithms. It should be noted that the pre-processing stage involved the automatic standardisation of the spelling and typography of the micro-texts, as described in Section 6.4. Moreover, during the construction of the doc-ngram matrix from the collection of tweets, lexical features took the form of stems filtered by functional stopwords.

The evaluation of each experiment was actually conducted as four separate tasks of binary classification. In other words, CASPER managed to determine whether or not a given micro-text could be classified as AQU, whether or not it could be classified as DRO, whether or not it could be classified as FLO, and whether or not it could be classified as LAN. Therefore, 1,532 classifications (i.e. 383 tweets multiplied by four categories) took place for each test in the first experiment, and 2,732 classifications (i.e. 683 tweets multiplied by four categories) in the second experiment. Moreover, k-fold cross validations (k = 4) were performed for machine-learning classifications in the second experiment. In other words, after dividing the collection of tweets into four samples of equal size, the validation process was repeated four times. At each iteration, one sample was selected as the test dataset and the remaining three samples were used as the training dataset, where every sample became the test dataset only once. In the end, the results generated by each iteration were averaged.

Finally, with respect to evaluation metrics for binary classification, it should be recalled that most of them are built over a 2x2 contingency matrix—as shown in Table 4, where TP, FP, FN and TN denote the number of true positives, false positives, false negatives and true negatives, respectively.

Table 4. Contingency matrix for binary classification.

			Expected	
			Does the tweet really pertain to class *x*?	
			yes	no
Predicted	Was the tweet classified as *x*?	yes	TP	FP
		no	FN	TN

In this regard, we employed two typical evaluation metrics that come from information retrieval, i.e. Precision and Recall, which are shown in the equations (5) and (6).

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

Moreover, we used one of the most popular measures that combines Precision and Recall, i.e. F1, which is presented in the equation (7).

$$F1 = \frac{2*Precision*Recall}{Precision+Recal} \tag{7}$$

7.2 Results

With respect to the first experiment, Tables 5-12 show the average scores for Precision, Recall and F1 in the detection of problems related to the topic categories AQU, DRO, FLO and LAN. In this experiment, we conducted 40 tests, where parameter tweaking was done for *Spell checking* (i.e. yes/no), *Ngram form* (i.e. stem/lemma), *POS* (i.e. yes/no) and the *Scope* of the valence shifters (i.e. 1-5).

Table 5. Evaluation of problem detection with CASPER.

	#1	#2	#3	#4	#5
Parameters					
Spell checking	yes	yes	yes	yes	yes
Ngram form	stem	stem	stem	stem	stem
POS	no	no	no	no	no
Scope	1	2	3	4	5
Scores					
Precision	0.99096	0.98979	0.98973	0.98958	0.98958
Recall	0.81841	0.72388	0.71890	0.70895	0.70895

F1	0.89646	0.83621	0.83285	0.82609	0.82609

Table 6. Evaluation of problem detection with CASPER [continued].

	#6	#7	#8	#9	#10
Parameters					
Spell checking	yes	yes	yes	yes	yes
Ngram form	lemma	lemma	lemma	lemma	lemma
POS	no	no	no	no	no
Scope	1	2	3	4	5
Scores					
Precision	0.99074	0.99071	0.99062	0.99048	0.99041
Recall	0.79851	0.79602	0.78856	0.77612	0.77114
F1	0.88430	0.88276	0.87812	0.87029	0.86713

Table 7. Evaluation of problem detection with CASPER [continued].

	#11	#12	#13	#14	#15
Parameters					
Spell checking	yes	yes	yes	yes	yes
Ngram form	stem	stem	stem	stem	stem
POS	yes	yes	yes	yes	yes
Scope	1	2	3	4	5
Scores					
Precision	0.99365	0.99361	0.99355	0.99283	0.99283
Recall	0.77861	0.77363	0.76617	0.68905	0.68905
F1	0.87308	0.86993	0.86517	0.81351	0.81351

Table 8. Evaluation of problem detection with CASPER [continued].

	#16	#17	#18	#19	#20
Parameters					
Spell checking	no	no	no	no	no
Ngram form	lemma	lemma	lemma	lemma	lemma
POS	no	no	no	no	no
Scope	1	2	3	4	5
Scores					
Precision	0.99010	0.99010	0.98997	0.98983	0.98976
Recall	0.74627	0.74627	0.73631	0.72637	0.72139
F1	0.85106	0.85106	0.84451	0.83788	0.83453

Table 9. Evaluation of problem detection with CASPER [continued].

	#21	#22	#23	#24	#25
Parameters					
Spell checking	no	no	no	no	no
Ngram form	stem	stem	stem	stem	stem
POS	no	no	no	no	no
Scope	1	2	3	4	5
Scores					
Precision	0.98917	0.98913	0.98905	0.98893	0.98893
Recall	0.68159	0.67910	0.67413	0.66667	0.66667
F1	0.80707	0.80531	0.80177	0.79643	0.79643

Table 10. Evaluation of problem detection with CASPER [continued].

	#26	#27	#28	#29	#30
Parameters					
Spell checking	no	no	no	no	no
Ngram form	stem	stem	stem	stem	stem
POS	yes	yes	yes	yes	yes
Scope	1	2	3	4	5
Scores					
Precision	0.99251	0.99248	0. 99239	0.99234	0.99234
Recall	0.65920	0.65672	0. 64736	0.64428	0.64428
F1	0.79223	0.79042	0.78217	0.78130	0.78130

Table 11. Evaluation of problem detection with CASPER [continued].

	#31	#32	#33	#34	#35
Parameters					
Spell checking	yes	yes	yes	yes	yes
Ngram form	lemma	lemma	lemma	lemma	lemma
POS	yes	yes	yes	yes	yes
Scope	1	2	3	4	5
Scores					
Precision	0.99149	0.99145	0.99134	0.99127	0.99119
Recall	0.57960	0.57711	0.56965	0.56468	0.55970
F1	0.73155	0.72956	0.72354	0.71949	0.71542

Table 12. Evaluation of problem detection with CASPER [continued].

	#36	#37	#38	#39	#40
Parameters					
Spell checking	no	no	no	no	no
Ngram form	lemma	lemma	lemma	lemma	lemma
POS	yes	yes	yes	yes	yes
Scope	1	2	3	4	5
Scores					
Precision	0.99057	0.99057	0.99038	0.99034	0.99024
Recall	0.52239	0.52239	0.51244	0.50995	0.50497
F1	0.68404	0.68404	0.67541	0.67323	0.66886

As Test #1 provided the highest F1 score, we also performed a series of tests (#1.1-#1.8) with the same settings in Test #1 but tweaking the *Valence shifters* parameters, i.e. *Negation* (i.e. yes/no), *Modifiers* (i.e. yes/no) and *Irrealis markers* (i.e. yes/no). Table 13 shows the results of this second stage of the experiment, where Test #1.1 fully corresponds to Test #1.

Table 13. Tweaking the *Valence shifters* parameters.

	#1.1	#1.2	#1.3	#1.4	#1.5	#1.6	#1.7	#1.8
Valence shifters								
Negation	yes	yes	yes	yes	no	no	no	no
Modifiers	yes	yes	no	no	no	no	yes	yes
Irrealis	yes	no	yes	no	yes	no	yes	no

markers								
Scores								
Precision	0.99096	0.99096	0.99000	0.99000	0.99000	0.99000	0.98983	0.98983
Recall	0.81841	0.81841	0.73880	0.73880	0.73880	0.73880	0.72637	0.72637
F1	0.89646	0.89646	0.84615	0.84615	0.84615	0.84615	0.83788	0.83788

With respect to the second experiment, Tables 14, 15 and 16 show the scores of each topic category for Precision, Recall and F1, which were calculated with the settings in Test #1 and with the supervised algorithms of multinomial Naïve Bayes and SVM, respectively.

Table 14. Evaluation of problem detection by topic categories with CASPER.

Category	AQU	DRO	FLO	LAN	average
Precision	1.000	0.964	0.990	1.000	0.989
Recall	0.702	0.643	0.954	0.905	0.801
F1	0.825	0.771	0.972	0.950	0.880

Table 15. Evaluation of problem detection by topic categories with multinomial Naïve Bayes.

Category	AQU	DRO	FLO	LAN	average
Precision	0.895	0.932	0.852	0.937	0.904
Recall	0.929	0.810	0.979	0.953	0.917
F1	0.909	0.866	0.910	0.943	0.907

Table 16. Evaluation of problem detection by topic categories with SVM.

Category	AQU	DRO	FLO	LAN	average
Precision	1.000	0.989	0.856	0.832	0.919
Recall	0.825	0.830	0.836	0.784	0.819
F1	0.903	0.902	0.845	0.806	0.864

7.3 Discussion

We can draw a number of conclusions from analysing the data in Tables 5-16. First, keeping the values of *Spell checking*, *Ngram form* and *POS* constant, both Precision and Recall decreased when *Scope* increased. Therefore, we conclude that the best value for *Scope* was 1. Second, with a view to analysing the relative contribution of *Spell checking*, *Ngram form* and *POS* to the performance of the system, we built a multiple linear regression model with the values shown in Table 17, which correspond to those in Tables 5-12 where *Scope* was 1.[13] In this model, the independent variables were *Spell checking, Ngram form* and *POS*, and the dependent variable was *Precision* or *Recall*.

Table 17. Evaluation of problem detection with CASPER (Scope = 1).

Test	**Spell checking**	**Ngram form**	**POS**	**Precision**	**Recall**
#1	1	1	0	0.99096	0.81841
#6	1	0	0	0.99074	0.79851

[13] In the case of *Ngram form*, lemma = 0 and stem = 1. In the remaining parameters, yes = 1 and no = 0.

#11	1	1	1	0.99365	0.77861
#16	0	0	0	0.99010	0.74627
#21	0	1	0	0.98917	0.68159
#26	0	1	1	0.99251	0.65920
#31	1	0	1	0.99149	0.57960
#36	0	0	1	0.99057	0.52239

On the one hand, the regression analysis for predicting Precision returned an R^2 score of 0.76399; in other words, *Spell checking, Ngram form* and *POS* explained 76.40% of the variability of Precision. Indeed, as shown in Table 18, there was a statistically significant relationship between Precision and *POS*, since p-value is less than 0.05. The magnitude of the t-value also served to confirm that *POS* was judged as the most decisive of the three parameters, which implies that Precision is more likely to increase if *POS* is taken into consideration.

Table 18. Regression analysis for predicting Precision.

Variable	**Coefficient**	**T-value**	**P-value**
Spell	0.00112	1.76062	0.15310
Ngram	0.00084	1.32929	0.25450
POS	0.00181	2.84287	0.04673

On the other hand, the regression analysis for predicting Recall returned an R^2 score of 0.73917. In this case, no parameter was clearly relevant, although POS almost attained statistical significance, as shown in Table 19.

Table 19. Regression analysis for predicting Recall.

Variable	**Coefficient**	**T-value**	**P-value**
Spell	0.09142	1.78937	0.14806
Ngram	0.07276	1.42413	0.22751
POS	-0.12624	-2.47100	0.06887

However, it is worth noting the negative coefficient of *POS*, which implies that Recall is more likely to decrease if *POS* is taken into consideration.

Third, Table 13 demonstrates that the highest score was returned when considering both *Negation* and *Modifiers* (Test #1.1 and Test #1.2), and the lowest score when considering *Modifiers* without *Negation* (Test #1.7 and Test #1.8). Therefore, we conclude that the joint implication of these two parameters played an important role in the processing of micro-texts. We also found that the parameter *Irrealis markers* was irrelevant, since it did not affect the performance of the system. However, this claim requires further conclusive evidence, as our collection of micro-texts didn't include linguistic realisations that contained irrealis markers.

Fourth, the second experiment enabled us to evaluate CASPER in comparison with the statistical approach (i.e. Naïve Bayes and SVM). Most researchers agree on the high precision of machine-learning methods when dealing with training and test datasets that share the same domain. In such a context, the motivation of the second experiment was to test whether or not supervised learning can outperform our knowledge-based approach. It is clear that, because cross validation was performed on the same collection of micro-texts, training and test datasets shared the same domain in the machine-learning classifications shown in Tables 15-16. When we compare the F1 scores in Tables 14-16, it soon becomes apparent that Naïve Bayes provided slightly better results as a whole. However, more significant is the fact that, whereas machine learning outperformed in Recall, the knowledge-based approach outperformed in Precision. Therefore, the question is now to determine which metric is of greater significance to decision makers, considering that the ultimate goal of this research is to implement an emergency-response system for the detection of environmentally related problems. In this regard, since "possible" positive results indicate "possible" problems, the greater the confidence in Precision, the greater the confidence in the system, since an excessive number of false-warning messages can increase anxiety in decision makers, forcing them to allocate unnecessary resources to monitor problems that are not indeed actual problems. Therefore, the major challenge should consist in increasing Precision, as in the case of CASPER, so that false alerts can be minimised.

It should be noted that evaluation results vary dramatically depending on whether or not training and test datasets pertain to the same domain. As described in Section 2, machine-learning methods are rather ineffective in classifying texts whose domain is different from that of the training documents. Indeed, Fernández Anta et al. (2013) presented a comprehensive set of experiments for topic and sentiment classification by testing 2,000 tweets with 5,000 trained tweets, exploring different techniques from "the full spectrum of classification methods provided by WEKA" (Fernández Anta et al., 2013: 47). Their experiments showed that the largest accuracy obtained was 58.45 for topic categorisation (complement Naïve Bayes) and 42.38 for sentiment analysis (multinomial Naïve Bayes), concluding that "none of the techniques explored is the silver bullet for Spanish tweet classification" (Fernández Anta et al. 2013: 52). One of the reasons for these poor experimental results was attributed to the multiple domains found in the datasets e.g. economy, literature, music, politics, sports and technology, among others. In the context of the sentiment analysis of Spanish tweets, Sidorov et al. (2013: 10) also demonstrated that "training with a corpus that has a domain different from the target domain affects precision very negatively, namely, it is two or three times worse". Therefore:

> For building a reliable text mining system, annotated corpora are indispensable. Additionally, systems need to be trained on corpora of the same domain as the target domain in order to show a good performance. (Ellendorff et al., 2016: 3723)

We can thus conclude that supervised machine-learning systems cannot be easily adapted to multiple concurrent events, since they would require to be provided with a training dataset that should be sufficiently large and representative with respect to each topic category, not to mention that the dataset should also be carefully annotated. In contrast, CASPER is able to deal with a wide variety of environmental hazards (e.g. avalanches, chemical spills, earthquakes, floods, forest fires, hurricanes, pollution, storms, tsunamis, and volcanic eruptions), only by providing a list of features for each topic category (Table 3). Consequently, our symbolic approach to classification, as illustrated in Figure 1, contributes to the portability of the system and the reuse of the resources across different domains.

In short, the above two experiments led us to report the following findings:

a) When we analyse the components of text processing within CASPER, we can conclude that:

a.1- POS has a great impact on performance. In fact, if POS is taken into consideration, Precision is more likely to increase but Recall is more likely to decrease, and vice versa. However, in case of being concerned with a trade-off between Precision and Recall (i.e. F1 score), then POS should be ignored.

a.2- The best results are achieved when valence shifters are involved. In particular, negation cues together with modifiers (i.e. intensifiers and diminishers) play an important role in determining the polarity of micro-texts. The impact of the valence shifter is preferably focused on the first preceding word and/or the following word in the neighbouring context.

b) When we compare our knowledge-based approach with that of supervised machine learning, we can conclude that:

b.1- CASPER considerably reduces human workload, since it does not require manually annotated training data.

b.2- CASPER easily accommodates to multiple different domains, since it does not depend on domain-specific corpora.

b.3- CASPER outperforms in Precision, which is a priority in the development of emergency-response systems.

Finally, as a preliminary step to improve performance, this section ends with a deep analysis of the errors that gave rise to the misclassified tweets in the first experiment with CASPER. On the one hand, FPs were almost non-existent, which had a positive effect on Precision. This is primarily attributed to the fact that text processing in CASPER is grounded on exogenous sources of knowledge from which a custom-made database was developed specifically for problem detection (cf. Section 5). Particularly, only three FPs were detected out of 1,532 classifications. An instance of micro-text whose topic was not correctly categorised is shown in the example (5).

(5) La empresa cocakola en el Salvador, roba nuestro recurso agua, primero secó manto acuífero en el centro, ahora va por Nejapa.
[The Coca-Cola Company in El Salvador steals our water resource, first it drained the aquifer downtown, it is now going for Nejapa.]

These few FPs were generated because topic features can serve as lexical descriptors of multiple categories, even of those that were not defined by the user. For example, the category *water-resource depletion* can subsume the categories (a) *surface-water depletion* and (b) *underground-water depletion*. In this regard, *secó* [drained] in the example (5) contextually pertains to (b), whereas *secó* [dried up] in the example (6) pertains to (a), which is in turn related to DRO.

(6) Ministra Pohl informa que río Angue, en Metapán se secó por primera vez.
[Minister Pohl reports that the Angue River in Metapán was dried up for the first time.]

The problem is certainly compounded when topic features take the form of polysemous or homonymous words.

On the other hand, FNs were more numerous, which had a negative effect on Recall. In particular, 52 FNs were detected during sentiment analysis and 38 FNs during topic categorisation, which were distributed as follows: 9 in AQU, 22 in DRO, 4 in FLO and 3 in LAN. The main reasons behind this type of errors are presented below:

a) Errors in language standardisation. For example, *Sequia* [Drought] in the example (7) was not processed as a misspelling, because it was treated as a proper name. However, if the missing punctuation marks had been inserted, CASPER could have corrected the accented character (i.e. *Sequía*). Similarly, there is a missing space after the interrogation mark in the example (8), so *PAÍS?Derrame* [COUNTRY?Molasses] was not processed as two separate tokens.

(7) En Santa Ana hace 5 Dias no llueve Sequia???
[It hasn't rained in Santa Ana for 5 Days Drought???]

(8) OTRO CRIMEN AMBIENTAL CONTRA EL PAÍS?Derrame de melaza
[ANOTHER ENVIRONMENTAL CRIME AGAINST THE COUNTRY?Molasses spill]

b) Incompleteness of the database. For example, the verb *rebalsa* [overflows] in the example (9) and the noun *desbordamiento* [overflowing] in the example (10) were not included as topic features of FLO. Similarly, the collocation *lluvia deficitaria* [deficit rain] in the example (11) could have been considered as a topic feature of DRO.

(9) Laguna El Jocotal de San Miguel se rebalsa por lluvias
[The El Jocotal de San Miguel lagoon overflows from rain]

(10) Decenas de casas estan casi bajo el agua por desbordamiento de Laguna del Jocotal
[Dozens of homes are nearly under water by the overflowing of the El Jocotal lagoon.]

(11) La lluvia deficitaria seguirá golpeando a la región Centroamericana en los meses de agosto, septiembre y octubre
[Deficit rain will continue to impact the Central American region in August, September and October]

This type of errors is easy to avoid by performing further research oriented to enhance both SENTIMENTS and the user-defined dataset of topic features.

c) Lack of background knowledge. For example, both the examples (12) and (13) require relevant commonsense knowledge to help the system infer that (a) a simple gutter (i.e. *canaleta de agua* [gutter]) carries significantly less water than any typical river, or that (b) extreme drought is one of the most important causes of water shortage, respectively. This background knowledge would have made it possible to classify both micro-texts as DRO.

(12) Ya no parece rio. Parece una canaleta de agua solamente.
[It no longer looks like a river. It only looks like a gutter.]

(13) Estamos en una situación grave con el tema del agua. No hay vida sin agua. No hay desarrollo sin agua
[We are in a serious situation with the lack of water. There is no life without water. There is no development without water]

This type of errors is difficult to deal with, because they require to implement strategies to get a deeper understanding of micro-texts.

8. Conclusion

Micro-texts from Twitter and other social media can become very valuable for the real-time detection of problems that affect people, thus having a profound impact on the management of decision-making processes. For example, the automatic detection of such troublesome situations can be useful not only for citizens but also for emergency responders. In this research, we address the development of a system (CASPER) that analyses micro-texts for the detection of environmentally related problems, such as aquifers, drought, floods and landslides.

Most of the earlier works on this subject have focused on the analysis of English tweets to detect a single or a few natural disasters from a supervised machine-learning approach. Instead, our research is aimed at detecting multiple disasters and hazards from Spanish tweets. To provide a workable solution in a real-life scenario, we soon realised that statistical models bring a number of problems derived from the need of a training corpus. First, it is necessary not only to collect a large number of tweets for each topic category but also to label each tweet with one or more categories. Indeed, this is a non-trivial task, since the quality of results returned by supervised models largely depends on both the distribution of the categories in the training dataset and the topic similarity between the training and test datasets. Second, it is not always possible to have a sufficiently representative number of micro-texts, especially when a few people report the incident, e.g. neighbours in a residential area who complain about the high level of particulate matter generated by the construction of a nearby mall. In contrast, we chose to adopt a knowledge-based approach that requires no training corpus but a small set of words as distinctive descriptors of each topic category, thus facilitating the portability of the system to new topic domains. Moreover, an added value of our model is given by not only classifying tweets with respect to a variety of problems but also assessing each tweet with a score (PPI) that determines the reliability that a given tweet actually deals with a particular environmentally related problem.

CASPER was evaluated through two experiments. The first experiment showed that the most effective fashion to analyse Spanish micro-texts involves NLP techniques such as spell checking, stemming, POS tagging, and the application of valence shifters, such as negation and modifiers, whose scope should be one ngram to the left or right of the valence shifter. Indeed, these techniques were used in the second experiment, where we compared the performance of our symbolic approach with that of machine learning (i.e. Naïve Bayes and SVM). Whereas CASPER obtained 0.989 in Precision and 0.801 in Recall, the best results with the supervised models were 0.919 in Precision (SVM) and 0.917 in Recall (Naïve Bayes). Considering that minimizing false-positive alarms is critical in emergency-response systems, since mobilising fire fighters, police officers or medical staff during the response stage of a large-scale disaster involves a considerable cost and organisational overhead, it is therefore desirable that the model should yield the highest possible precision, as in the case of CASPER.

Today we are successfully moving this proof-of-concept workbench to a real-world application, which is included in a project developed for the Ministry of Environment and Natural Resources in El Salvador. In particular, the measurements generated by 124 weather stations that were constructed throughout the country are accompanied by the measurements obtained from social sensors. Complementing the information provided by the electromechanical sensors with that of social sensors is proving to be very helpful for the government to detect environmentally related problems, such as droughts, floods and hurricanes, among others.

Future research is mainly aimed at automatically discovering spatio-temporal indicators within the micro-text. In other words, in a sentence such as *Tras las fuertes lluvias, ayer se inundó Valencia* [Valencia was flooded yesterday after heavy rainfall], CASPER should be able to find "where" (e.g. *Valencia*) and "when" (e.g. *ayer* [yesterday]) information. Moreover, the complexity of natural language requires a deeper understanding of tweets, so that a message such as *Nuestros hijos tendrán el planeta que les dejemos. ¿Qué hacen los políticos al respecto?* [Our children will inherit the planet as we leave it for them. What are politicians doing about this?], which actually conveys negative polarity without including any sentiment-bearing word, can be correctly classified.

In a new version of CASPER, we also intend to expand the system to analyse English tweets. CASPER is provided with two types of language-dependent resources: text-processing resources, i.e. tokeniser, stemmer, lemmatiser and POS tagger, and lexical resources, as illustrated by the four data stores in Figure 1. On the one hand, the task of integrating new text-processing modules to cover languages such as English, French or Italian is not expected to be time-consuming, since these modules are readily available in DAMIEN (Periñán-Pascual, 2017), a workbench that allows researchers to perform text analytics. On the other hand, the effort to develop further lexical resources should be primarily focused on the valence shifters (i.e. NEGATION and MODIFIERS) and the polarity lexicon (SENTIMENTS). In contrast, SYNSETS, GLOSSES, POS and RELATIONS are easily adaptable to English, since they were constructed from the Multilingual Central Repository, where the English WordNet is used as the backbone of the Spanish WordNet.

Acknowledgments

Financial support for this research has been provided by the Spanish Ministry of Economy, Industry and Competitiveness [grant number TIN2016-78799-P] (AEI/FEDER, EU) and by the Spanish Ministry of Education and Science [grant number FFI2014-53788-C3-1-P].

References

Aiala, R., Wonsever, D., Jean-Luc, M. (2010). Opinion identification in Spanish texts. Proceedings of the NAACL HLT Young Investigators Workshop on Computational Approaches to Languages of the Americas.

Álvarez, I. (2011). El ciberespañol: características del español usado en Internet. Luis A. Ortiz-López (ed.) Selected Proceedings of the 13th Hispanic Linguistics Symposium, Somerville (Mass.), Cascadilla Press.

Àlvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., & Rigau, G. (2008). Complete and consistent annotation of WordNet using the Top Concept Ontology. Proceedings of the 6th Conference on Language Resources and Evaluation.

Araque, O., Corcuera-Platas, I., Román-Gómez, C., Iglesias, C.A., & Fernando Sánchez-Rada, J. (2015). Aspect based sentiment analysis of Spanish tweets. Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN, Alicante, pp. 29-34.

Atserias, J., Comelles, E., & Mayor, A. (2005). Txala: un analizador libre de dependencias para el castellano. Procesamiento del Lenguaje Natural 35, pp. 455-456.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., & Vossen, P. (2004). The MEANING Multilingual Central Repository. Proceedings of the Second International Global WordNet Conference.

Ayala Hernández, D., Roldán Salvador, J.C., Ruiz Cortés, D., & Ortega Gallego, F. (2015). An approach for discovering keywords from Spanish tweets using Wikipedia. Advances in Distributed Computing and Artificial Intelligence Journal 4 (2), pp. 73-87.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the Seventh Conference on International Language Resources and Evaluation. European Language Resources Association, pp. 2200-2204.

Balbachan, F., & Dell'Era, D. (2012). Análisis automatizado de sentimiento en textos breves de la plataforma Twitter. Revista de Lingüística Informática, Modelizacion e Ingeniería Lingüística (Infosur) 6, pp. 3-14.

Budge, I. (1999). Estimating party policy preferences: From ad hoc measures to theoretically validated standards. Essex Papers in Politics and Government 139. University of Essex and Department of Government.

Carreras, X. (2007). Experiments with a higher-order projective dependency parser. Proceedings of the EMNLP-CoNLL 2007 Shared Task. Prague.

Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). FreeLing: An open-source suite of language analyzers. Proceedings of the 4th International Conference on Language Resources and Evaluation.

Cordobés, H., Fernández Anta, A., Chiroque, L.F., Pérez, F., Redondo, T., & Santos, A. (2014). Graph-based techniques for topic classification of tweets in Spanish. International Journal of Artificial Intelligence and Interactive Multimedia 2 (5), pp. 31-37.

Crooks, A. et al. (2013). #Earthquake: Twitter as a distributed sensor system. Transactions in GIS 17 (1), pp. 124–147.

Crystal, D. (2004). Language and the Internet. Cambridge, Cambridge University Press.

Díaz Rangel, I., Sidorov, G., & Suárez-Guerra, S. (2014). Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. Onomázein 29, pp. 31-46.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. Proceedings of the First ACM International Conference on Web Search and Data Mining, Stanford (California), pp. 231-239.

Ellendorff, T.R., Foster, S., & Rinaldi, F. (2016). The PsyMine Corpus - A corpus annotated with psychiatric disorders and their etiological factors. Proceedings of the 10th edition of the Language Resources and Evaluation Conference, pp. 3723-3729.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors 37 (1), pp. 32-64.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. Proceedings of the 5th Conference on Language Resources and Evaluation. Genoa: European Language Resources Association, pp. 417-422.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fernández Anta, A., Núñez Chiroque, L., Morere, P., & Santos, A. (2013). Sentiment analysis and topic detection of Spanish tweets: A comparative study of NLP techniques. Procesamiento del Lenguaje Natural 50, pp. 45-52.

Fernández Vítores, D. (2017). El español: una lengua viva. Informe 2017. Technical report, Instituto Cervantes. https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2017.pdf

Gamallo, P., García, M., & Fernández-Lanza, S. (2013). TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets. Proceedings of the Workshop on Sentiment Analysis at SEPLN, pp. 126-132.

Gambino, O.J., & Calvo, H. (2016). A comparison between two Spanish sentiment lexicons in the Twitter sentiment analysis task. Advances in Artificial Intelligence - IBERAMIA 2016. Lecture Notes in Computer Science, vol. 10022. Springer, Cham.

Gonzalez-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual Central Repository version 3.0: Upgrading a very large lexical knowledge base. Proceedings of the Sixth International Global WordNet Conference.

Goswami, S. et al. (2016). A review on application of data mining techniques to combat natural disasters. Ain Shams Engineering Journal. http://www.sciencedirect.com/science/article/pii/S2090447916000307.

Grefenstette, G., Qu, Y., Shanahan, J., & Evans, D. (2004). Coupling niche browsers and affect analysis for an opinion mining application. Proceedings of RIAO 2004.

Hatcher, E., Gospodnetic, O., & McCandless, M. (2010). Lucene in Action. Greenwich: Manning.

Hogenboom, A., Van Iterson, P., Heerschop, B., Frasincar, F., & Kaymak, U. (2011). Determining negation scope and strength in sentiment analysis. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pp. 2589-2594.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of SIGKDD.

Huang, Q., & Xiao, Y. (2015). Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery. ISPRS International Journal of Geo-Information 4 (3), pp. 1549-1568.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013a). Extracting information nuggets from disaster-related messages in social media. Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management, Baden Baden.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013b). Practical extraction of disaster-relevant information from social media. Proceedings of the 22nd International Conference on World Wide Web, New York: Association for Computing Machinery, pp. 1021-1024.

Jimenez-Zafra, S.M., Valdivia, M.T.M., Camara, E.M., & Urena-Lopez, L.A. (2017). Studying the scope of negation for Spanish sentiment analysis on Twitter. IEEE Transactions on Affective Computing.

Juršič, M., Mozetič, I., Erjavec, T., & Lavrač, N. (2010). LemmaGen: Multilingual lemmatisation with induced Ripple-Down rules. Journal of Universal Computer Science 16 (9), pp. 1190-1214.

Karimi, S., Yin, J., & Paris, C. (2013). Classifying microblogs for disasters. Proceedings of the 18th Australasian Document Computing Symposium, New York: Association for Computing Machinery, pp. 26-33.

Khandelwal, A., Swami, S., Sarfaraz Akhtar, S., & Shrivastava, M. (2017). Classification of Spanish Election Tweets (COSET) 2017: Classifying tweets using character and word level features. Second Workshop on Evaluation of Human Language Technologies for Iberian Languages at SEPLN, pp. 49-54.

Kim, H., & Fording, R.C. (2002). Government partisanship in western democracies, 1945–1998. European Journal of Political Research 41 (2), pp. 187–206.

Kim, S., & Hovy, E. (2004). Determining the sentiment of opinions. Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics, pp. 1367-1374.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, N.A. (2014). A dependency parser for tweets. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha. Association for Computational Linguistics, pp. 1001-1012.

Konstantinova, N., De Sousa, S.C.M., Cruz, N.P., Maña, M.J., Taboada, M., & Mitkov, R. (2012). A review corpus annotated for negation, speculation and their scope. Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, pp. 3190-3195.

Kralj Novak, P., Smailovic, J., Sluban, B., & Mozetic, I. (2015). Sentiment of emojis. PLoS ONE 10 (12). e0144296. https://doi.org/10.1371/journal.pone.0144296

Lowe, W., Benoit, K., Mikhaylov, S., & Laver, M. (2011). Scaling policy preferences from coded political texts. Legislative Studies Quarterly 36, pp. 123-155.

Magnini, B., & Cavaglià, G. (2000). Integrating subject field codes into WordNet. Proceedings of the Second International Conference on Language Resources and Evaluation.

Martí, M.A., Martín-Valdivia, M.T., Taulé, M., Jiménez-Zafra, S.M., Nofre, M., & Marsó, L. (2016). La negación en español: Análisis y tipología de patrones de negación. Procesamiento del Lenguaje Natural 57, pp. 41-48.

Miller, H., Kluver, D., Thebault-Spieker, J., Terveen, L., & Hecht, B. (2017). Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, pp. 152-161.

Moreno-Ortiz, A., & Pérez Hernández, C. (2013). Lexicon-based sentiment analysis of Twitter messages in Spanish. Procesamiento del Lenguaje Natural 50, pp. 93-100.

Nemeth, L., Tron, V., Halacsy, P., Kornai, A., Rung, A., & Szakadat, I. (2004). Leveraging the open source ispell codebase for minority language analysis. Proceedings of SALTMIL Workshop at the Language Resources and Evaluation Conference 2004: First Steps in Language Documentation for Minority Languages.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., & Marsi, E. (2007). Maltparser: A language independent system for data-driven dependency parsing. Natural Language Engineering 13 (2), pp. 95–135.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. Proceedings of the Language Resources and Evaluation Conference 2012. European Language Resources Association.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL 2002

Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 79-86.

Pease, A., Niles, I., & Li, J. (2002). The Suggested Upper Merged Ontology: A large ontology for the semantic web and its applications. Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web.

Periñán-Pascual, C. (2017). Bridging the gap within text-data analytics: a computer environment for data analysis in linguistic research. Revista de Lenguas para Fines Específicos 23 (2), pp. 111-132.

Pla, F., & Hurtado, LF. (2018). Spanish sentiment analysis in Twitter at the TASS workshop. Language Resources and Evaluation 52 (2), pp. 645-672.

Plaza del Arco, F.M., Martín-Valdivia, M.T., Jiménez Zafra, S.M., Molina-González, M.D., & Martínez Cámara, E. (2016). COPOS: Corpus of Patient Opinions in Spanish. Application of sentiment analysis techniques. Procesamiento del Lenguaje Natural 57, pp. 83-90.

Polanyi, L., & Zaenen, A. (2004). Contextual valence shifters. Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. Menlo Park (California): the AAAI Press, pp. 106-111.

Popescu, A., & Etzioni, O. (2005). Extracting product features and opinions from reviews. Proceedings of the Human Language Technology Conference and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 339-346.

Rill, S., Scheidt, J., Drescher, J., Schütz, O., Reinel, D., & Wogenstein, F. (2012). A generic approach to generate opinion lists of phrases for opinion mining applications. Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, pp. 7:1–7:8.

Sakaki, T., & Matsuo, Y. (2012). Earthquake observation by social sensors. Earthquake Research and Analysis—Statistical Studies, Observations and Planning, pp. 313-334.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. Proceedings of the 19th International Conference on World Wide Web ACM.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Transactions on Knowledge and Data Engineering 25 (4), pp. 919-931.

Segura-Bedmar, I., Quirós, A., & Martínez, P. (2017). Exploring convolutional neural networks for sentiment analysis of Spanish tweets. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 1, pp. 1014-1022.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., & Velásquez, F. (2013). Empirical study of machine learning based approach for opinion mining in tweets. Proceedings of the 11th Mexican International Conference on Artificial Intelligence, pp. 1-14.

Sixto, J., Almeida, A., & López-de-Ipiña, D. (2016). Improving the sentiment analysis process of Spanish tweets with BM25. Natural Language Processing and Information Systems. NLDB 2016. Lecture Notes in Computer Science, vol. 9612. Springer, Cham.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics 37 (2), pp. 267-307.

Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A., & Anderson, K.M. (2011). Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency. Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, Barcelona.

Vieweg, S., Hughes, A.L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 1079-1088.

Vilares, D., Alonso, M.A., & Gómez-Rodríguez, C. (2015a). A linguistic approach for determining the topics of Spanish Twitter messages. Journal of Information Science 41 (2), pp. 127-145.

Vilares, D., Doval, Y., Alonso, M., & Gomez-Rodríguez, C. (2015b). Lys at TASS 2015: Deep learning experiments for sentiment analysis on Spanish tweets. Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN, Alicante, pp. 47-52.

Vossen, P. (1998). Introduction to EuroWordNet. Computers and the Humanities 32 (2-3), pp. 73-89.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, Uppsala, pp. 60-68.

Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. IEEE Intelligent Systems 27 (6), pp. 52-59.

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. Political Communication 29 (4), pp. 205-231.