Intelligent Environments 2017
C. Analide and P. Kim (Eds.)

© 2017 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-796-2-49

A Knowledge-Based Approach to Social Sensors for Environmentally-Related Problems

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 can serve to contribute to situation awareness, which involves the perception and comprehension of the reality around us, so that future actions can be projected. For example, Twitter can be used as a real-time channel of communication to report on environmentally-related problems. The goal of this paper is to describe a knowledge-based system that is able to detect such problems, so that a protocol of action can be developed. The evaluation of the system demonstrated that our symbolic approach to problem detection can outperform supervised classification methods.

Keywords. Twitter, social sensor, topic categorization, 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. There are two types of sensors: electronic sensors and social sensors. As explained by Crooks et al. [1], social sensors operate in a manner comparable to electronic sensors: micro-bloggers are the sensors and the microcontrollers, since they collect the information that is important to communicate, whereas the actual 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, since "users sometimes misunderstand phenomena, sleep, and are not near a computer" [2], social sensors stand out for their low operating cost, wide geographical dissemination and immediate information transfer. Regarding our research, we focus on the automatic analysis of Spanish micro-texts from Twitter to develop a protocol of action to manage environmentally-related problems, such as overflowing rivers, waste discharge or wildfires, among many others. The remainder of this paper is structured 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 the methodology, 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 English texts. For example, Sakaki et al. [3][4] presented one of the first applications to use Twitter as a medium for social sensors to detect real-time events. They devised a 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 center of the event location. As a result, they developed a reporting system to promptly notify people of an earthquake in Japan. Liu et al. [5] described a tweet-based detection system used by the U.S. Geological Survey to rapidly detect widely felt seismic events. The algorithm essentially scans for significant increases in tweets containing the word "earthquake" or its equivalent in other languages and sends alerts with the detection time, tweet text, and the location where most of the tweets originated.

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 categorization and sentiment analysis. In comparison with English, there is a small number of studies on topic categorization or sentiment analysis in Spanish [6][7][8][9][10][11][12]. In this regard, two main approaches can be distinguished: a machine learning approach, which is usually implemented through a supervised method, and a symbolic approach, which is based on lexicons and rules. A supervised machine-learning method (e.g. KNN, Naïve Bayes or SVM) requires a training dataset, that is, a collection of text data that have 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 sufficiently large and representative. For example, Sidorov et al. [12] recommended having a training dataset containing at least 3,000 tweets. This requirement conflicts with the development of a system like ours, which was intended to classify new tweets on the ground of dynamically created categories of environmentally-related problems. The effort to expand a given training dataset to fit new categories makes applicability to new domains a non-trivial task. This fact actually became 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” [11]. For this reason, the solution was aimed at dealing with problem detection from an unsupervised lexicon-based approach.

4. Knowledge resources for Spanish

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 that were used in our research.

4.1. Multilingual Central Repository

The Multilingual Central Repository [13][14] integrates wordnets from six languages (i.e. Basque, Catalan, English, Galician, Portuguese and Spanish) into the same EuroWordNet framework. The Inter-Lingual-Index allows the connection from words in one language to equivalent translations in any of the other languages. 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 [15], WordNet Domains [16] and SUMO [17].

4.2. SFU-Review-SP-Neg

SFU-Review-SP-Neg [18] comprises 2,953 sentences, which 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. This resource focuses on syntactic-level negation, being annotated as (a) simple, expressed with a single particle (e.g. *Nunca han dado problemas* [They have never given rise to problems], *Nadie quedará decepcionado* [Nobody will get disappointed] or *Es un teléfono sin cobertura* [The phone is out of range]), or (b) complex, expressed with two or more particles— continuous (e.g. *Casi no llega* [He almost failed to arrive]) o discontinuous (e.g. *No vino nunca* [She never came]). It should also be noted that sentences that contain negative particles such as *a más no poder* [in the extreme] or *ni que decir que* [it goes without saying that] do not really express negation.

4.3. SentiWordNet

SentiWordNet [19][20] is the result of automatically annotating all synsets in English WordNet 3.0 according to their degrees of positivity, negativity and objectivity. Since WordNet synsets represent abstract concepts, 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 and opinion mining applications.

4.4. Spanish Emotion Lexicon

The Spanish Emotion Lexicon [21][22] 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. The database

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

$$KB = \left\{ \begin{array}{c} SYNSETS: [\{STEM, SYNSET\}], \\ POS: [\{SYNSET, TYPE\}], \\ RELATIONS: [\{RELATION, SYNSET1, SYNSET2\}], \\ NEGATION: [\{OPERATOR, TYPE, SCOPE\}], \\ SENTIMENTS: [\{STEM, POLARITY\}], \\ CATEGORIES: [\{CATEGORY, STEM, POS\}] \end{array} \right\}$$

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

SYNSETS, POS and RELATIONS were built from Multilingual Central Repository. SYNSETS holds all the synsets that are lexicalized in the Spanish WordNet, together with the stemmed words assigned to the synsets; stemming was performed with the SnowBall Analyzer. POS stores the grammatical category (i.e. noun, verb, adjective or adverb) linked to every synset. RELATIONS holds the semantic relations that can occur between two synsets; the only relations that were relevant for this project were *causes*, *has_hyponym*, *has_subevent*, *derived_from*, *near_synonym*, *pertains_to* and *related_to*.

NEGATION serves as our polarity dictionary, which resulted from the analysis of the negative particles in SFU-Review-SP-Neg. These particles were expanded with synonyms from the Spanish WordNet and stored as operators, which were classified as *neg* and *noneg*. The scope of the operator can be to the left and/or right of the negative particle.

SENTIMENTS holds the stems of words associated with positive or negative polarity. On the one hand, positively-marked words 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 words were extracted from those terms that belong to the sentiment dimensions of anger, disgust, fear and sadness in the Spanish Emotion Lexicon. Moreover, complaint words not present in the Spanish Emotion Lexicon were detected from a corpus of 790 tweets. Manual validation of SENTIMENTS was required.

Finally, CATEGORIES is used to store the significant features, in the form of stems and their part-of-speech, that are semi-automatically discovered from a new category of environmentally-related problem.

6. Methodology

6.1. Registering categories

The system was designed to classify tweets, i.e. micro-texts with a maximum of 140 characters, on the basis of user-defined environmentally-related problems. Therefore, a new category implies a semi-automatic process of selecting significant features, that is, relevant words that identify the target event. First, the user decides a few seed terms that are representative of the new category. Second, the system presents the different senses of each seed term, so that relevant meanings (synsets) for the category can be selected. As each seed term becomes a feature, this results in the vector $C_i = (f_{i1}, f_{i2}, \dots, f_{ik})$, where every f_{ij} identifies a feature in the form of a synset assigned to the category C_i. Third, a relation-driven expansion of C_i takes place by means of RELATIONS. In particular:

- 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*, and in the relations *x-derived_from-y* and *x-pertains_to-y*, where f_{ij} can instantiate *x* or *y*; in both cases, each expansion is added to C_i as f_{ik+1}.
- For each f_{ij} in C_i, expand to other synsets involved in the relation *x-causes-y*, where f_{ij} instantiates *x*; each expansion is added to C_i as f_{ik+1}.
- For each f_{ij} in C_i, expand to other synsets involved in the relation *x-near_synonym-y*, where f_{ij} instantiates *x*; each expansion is added to C_i as f_{ik+1}.

Fourth, every f_{ij} in C_i is mapped into one or several stems (together with their grammatical categories) with SYNSETS and POS. The outcome is stored in CATEGORIES.

6.2. Collecting and processing data

With the aid of the Twitter API, the next step consisted in crawling tweets related to the target event. The tweets were processed as follows. First, some elements were removed from the micro-texts, such as hashtags (i.e. any word starting with #), references (i.e. usernames headed by @) and URLs. Second, texts were tokenized and tagged with their POS by using the Stanford POS Tagger. Third, each token was stemmed with the SnowBall Analyzer, where each stem was in turn refined, so that all the inflectional forms of a given word could be reduced to a single stem. At this point, a tweet is represented as the vector $T_m = (w_{m1}, w_{m2}, \dots w_{mp})$, 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, the stem, the part-of-speech, the topic and the sentiment, where the values of the latter two are discovered in the next step.

6.3. Discovering relevant stems

This stage consists in detecting significant stems with respect to the topic (i.e the target event or category) and the sentiment. On the one hand, the weight 1 was assigned to the attribute topic of every w_{mn} in T_m whose stem and part-of-speech was found as an f_{ij} in C_i, which was stored in CATEGORIES. On the other hand, the values *p* or *n* (i.e. positive or negative) were assigned to the attribute sentiment of every w_{mn} in T_m according to the polarity of the stem in SENTIMENTS.

In both cases, contextual valence shifters were taken into consideration. In other words, the scope of negation affects the topic and sentiment values of w_{mn}. In particular, when a stem is found within the scope of a negative particle (i.e. three words to the left and/or right) according to NEGATION, the stem is no longer significant for topic and sentiment, so the values of these attributes become 0 and *o* (i.e. objective) respectively.

6.4. Determining topic and sentiment

On the one hand, 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 normalized dot product) as a measure of semantic distance. In our case, since we deal with binary values for topic relatedness and the number of topic-related stems in the tweet T_m is equal to or less than the number of relevant features in the category C_i, the relatedness function between T_m and C_i can be reduced to Eq. (1):

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

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

On the other hand, a simple approach to sentiment calculation could have been to sum up the sentiment values of each stem in the text message. However, we chose to assess the degree of sentiment in a given tweet with a metric originally used to assess political positions in texts. Particularly, Lowe et al. [23] 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 stems in the tweet T_m into a point on the sentiment dimension S by means of Eq. (2):

$$rel(T_m, S) = \log(P + 0.5) - \log(N + 0.5) \tag{2}$$

where P and N refer to the number of the positively- and negatively-marked stems in T_m respectively.

6.5. Detecting the problem

In this context, detecting an environmentally-related problem implies to fix a minimum number of tweets where each T_m has been tagged with a given event category (i.e. $rel(T_m, C_i)$ returns a positive value) and categorized with a negatively-marked sentiment (i.e. $rel(T_m, S)$ returns a negative value) for some specific time and location. After the normalization of the sentiment score, we calculated the geometric mean to combine both values into a single measure, i.e. the problem-relatedness perception index (PPI), as shown in Eq. (3), which can be simplified into Eq. (4):

$$PPI(T_m, C_i, S) = \sqrt{rel(T_m, C_i) * \left(1 - \frac{1}{\log(-rel(T_m, S) + 2)}\right)} \tag{3}$$

$$PPI(T_m, C_i, S) = \sqrt{rel(T_m, C_i) - \frac{rel(T_m, C_i)}{\log(-rel(T_m, S) + 2)}} \tag{4}$$

In other words, the PPI serves to measure how reliable we can feel that a given tweet (T_m) deals with a problem (S) about an environmental topic (C_i). How to determine spatio-temporal indicators of the event categories in the tweet is out of the scope of this paper.

7. Evaluation

This research was evaluated with a corpus of 300 tweets, where 108 were manually categorized as flood (INU), 110 tweets as rain (LLU) and 74 tweets as landslide (DESL). In this experiment, a single seed term was used to expand the relevant features of each event category, as shown in Table 1.

Table 1. Relevant features of the categories DESL, INU and LLU.

Category	Seed term	Features
DESL	deslizamiento	deslizamiento
INU	inundación	aluvión, anegar, avalancha, avenida, diluvio, inundación, inundar
LLU	lluvia	aguacero, empapado, húmedo, llover, lluvia, lluvioso, mojado, pluvioso, precipitar, precipitación

Most evaluation metrics for two-category classification are built over a 2x2 contingency matrix, as illustrated in Table 2, where TP, FP, FN and TN denote the number of true positives, false positives, false negatives and true negatives respectively.

Table 2. Contingency matrix for binary classification.

		Expected	
		1	0
Predicted	1	TP	FP
	0	FN	TN

Typical evaluation metrics that come from information retrieval are Precision and Recall, as shown in Eq. (5):

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

One of the most popular measures that combines Precision and Recall is F1, which is presented in Eq. (6):

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \tag{6}$$

The evaluation was conducted with our knowledge-based system, which represents the symbolic approach, and with multinomial Naïve Bayes, which illustrates the machine-learning approach. In the context of processing Spanish tweets for sentiment analysis, a popular classification method is Naïve Bayes, because "it is a simple and intuitive method whose performance is similar to other approaches. NB combines efficiency (optimal time performance) with reasonable accuracy" [10]. In this case, a doc-ngram matrix was created from the corpus, whose binary-weighted features took the form of stems filtered by functional stopwords. A k-fold cross validation, where k = 10, was performed. Table 3 shows the values of Precision, Recall and F1 in the topic categorization of DESL, INU and LLU.

Table 3. Evaluation of topic categorization.

	Knowledge-Based			Multinomial Naïve Bayes		
	Precision	Recall	F1	Precision	Recall	F1
DESL	0.985	0.930	0.957	0.576	0.623	0.593
INU	0.986	0.724	0.835	0.537	0.628	0.570
LLU	0.960	0.818	0.883	0.586	0.590	0.580

With regard to sentiment analysis, the evaluation procedure was the same as that described above, whose results are shown in Table 4.

Table 4. Evaluation of sentiment analysis.

Knowledge-Based			Multinomial Naïve Bayes		
Precision	Recall	F1	Precision	Recall	F1
0.835	0.763	0.798	0.428	0.475	0.450

Finally, Table 5 shows the evaluation of problem relatedness once topic categorization and sentiment analysis are integrated by means of the PPI.

Table 5. Evaluation of problem relatedness.

	PPI		
	Precision	Recall	F1
DESL	0.983	0.830	0.899
INU	0.983	0.622	0.761
LLU	0.956	0.500	0.656
micro-average	0.976	0.638	0.771
macro-average	0.974	0.650	0.779

We conclude that our symbolic approach to problem relatedness can outperform classical methods of classification. Indeed, Fernández Anta et al. [9] presented a comprehensive set of experiments classifying 7,000 Spanish tweets according to topic and sentiment by means of different approaches and techniques. Their experiments showed that the largest accuracy obtained was 0.58 for topic categorization and 0.42 for sentiment analysis.

8. Conclusion

Micro-texts from Twitter and other social media have become very valuable for the real-time detection of the concern that affects people. The automatic detection of such events can be really useful not only for citizens but also for emergency responders. In this research, we address the development of a system that exploits Twitter users as social sensors for the detection of environmentally-related problems (e.g. floods, droughts, landslides or pollution, among many others). The paper primarily focuses on the external knowledge resources that were required, the processing of Spanish tweets, the discovery of relevant features and the detection of the problem as a two-fold task: topic categorization and sentiment analysis. As our system enables the user to dynamically create categories upon which new tweets should be classified accordingly, supervised machine-learning methods turn out to be impractical, since they largely depend on training datasets that should be sufficiently large and representative with respect to the new categories as well as being carefully annotated. In fact, the evaluation of our research demonstrated that our symbolic approach provides better results than a supervised classification method (i.e. multinomial Naïve Bayes), not only in text categorization but also in sentiment analysis.

Acknowledgments

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

References

[1] A. Crooks et al., #Earthquake: twitter as a distributed sensor system, *Transactions in GIS* **17-1** (2013), 124–147.

[2] T. Sakaki and Y. Matsuo, Earthquake Observation by Social Sensors, *Earthquake Research and Analysis—Statistical Studies, Observations and Planning* (2012), 313-334.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, *Proceedings of the 19th international conference on World Wide Web ACM* (2010).

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

[5] SB. Liu, B. Bouchard, DC. Bowden, M. Guy, P. Earle, USGS tweet earthquake dispatch (@USGSted): using twitter for earthquake detection and characterization, *AGU fall meeting abstracts* **1** (2012).

[6] R. Aiala, D. Wonsever, M. Jean-Luc, Opinion Identification in Spanish Texts, *Proceedings of the NAACL HLT* (2010).

[7] E. Martínez Cámara, M.T. Martín Valdivia, J.M. Perea Ortega, L.A. Ureña López, Técnicas de Clasificación de Opiniones Aplicadas a un Corpus en Español, *Procesamiento del Lenguaje Natural* **47** (2011), 163-170.

[8] F. Balbachan, D. Dell'Era, Análisis Automatizado de Sentimiento en Textos Breves de la Plataforma Twitter, *Revista de Lingüística Informática, Modelizacion e Ingeniería Lingüística (Infosur)* **6** (2012), 3-14.

[9] A. Fernández Anta, L. Núñez Chiroque, P. Morere & A. Santos, Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques. *Procesamiento del Lenguaje Natural* **50** (2013), 45-52.

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

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

[12] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez & F. Velásquez, Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets, *Advances in Artificial Intelligence* (2013), 1–14.
[13] J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini and P. Vossen, The MEANING Multilingual Central Repository, *Proceedings of the Second International Global WordNet Conference* GWC (2004).
[14] A. Gonzalez-Agirre, E. Laparra and G. Rigau, Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base, *Proceedings of the Sixth International Global WordNet Conference* (GWC 2012).
[15] J. Àlvez, J. Atserias, J. Carrera, S. Climent, E. Laparra, A. Oliver and G. Rigau, Complete and Consistent Annotation of WordNet using the Top Concept Ontology, *Proceedings of the 6th Conference on Language Resources and Evaluation LREC* (2008).
[16] B. Magnini and G. Cavaglià, Integrating subject field codes into WordNet, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)* (2000).
[17] A. Pease, I. Niles and J. Li, The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications, *AAAI* (2002).
[18] M. A. Martí, M. T. Martín-Valdivia, M. Taulé, S. M. Jiménez-Zafra, M. Nofre, & L. Marsó, La negación en español: análisis y tipología de patrones de negación, *Procesamiento del Lenguaje Natural* **57** (2016), 41-48.
[19] A. Esuli, F. Sebastiani, SentiWordNet: a publicly available lexical resource for opinion mining, *Proceedings of the 5th Conference on Language Resources and Evaluation. Genoa, Italy: European Language Resources Association* (2006), 417-422.
[20] S. Baccianella, A. Esuli and F. Sebastiani, SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC European Language Resources Association (ELRA)* (2010), 2200-2204.
[21] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, Se. Suárez-Guerra, A. Treviño and J. Gordon, Empirical Study of Opinion Mining in Spanish Tweets, *LNAI* **7629**, (2012), 1-14.
[22] I. Díaz Rangel, G. Sidorov, S. Suárez-Guerra. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español, *Onomázein* **29** (2014), 31-46.
[23] W. Lowe, K. Benoit, S. Mikhaylov and M. Laver, Scaling Policy Preferences from Coded Political Texts, *Legislative Studies Quarterly* **36** (2011), 123-155.