Monge: Geographic Monitor of Diseases

Monge: Monitor Geográfico de Enfermedades

Salud María Jiménez-Zafra, Flor Miriam Plaza-del-Arco, Miguel Ángel García-Cumbreras, María Dolores Molina-González, L. Alfonso Ureña-López, M. Teresa Martín-Valdivia
Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{sjzafra, fmplaza, magc, mdmolina, laurena, maite}@ujaen.es

Abstract: Monge is a prototype of a geographic monitor of diseases, based on tweets. After the recovering phase of tweets, located in different Spanish cities, these tweets are processed and filtered with techniques and tools of Human Language Technologies. Tweets are filtered with three criteria: location, language (Spanish and Catalan) and bag of words of diseases (generated using synonyms of WordReference and embeddings). The processed information is presented in an interactive way allowing to predict possible epidemic outbreaks of different diseases (e.g. flu, asthma). This demo could be very useful because the Centers for Disease Control and Prevention take between 1-2 weeks from the moment the patient is diagnosed until the data is available, while with this prototype a real-time monitoring of diseases is offered.
Keywords: Natural language processing, web application, social monitoring, Twitter, word embeddings

Resumen: Monge es un prototipo de un monitor geográfico de enfermedades basado en tweets. Recuperando tweets localizados en distintas ciudades españolas, tanto en español como en catalán, y procesando y analizando la información con técnicas y herramientas de Tecnologías del Lenguaje Humano, permite predecir posibles brotes epidémicos de distintas enfermedades de interés general (gripe, asma, etc.). Los tweets son filtrados utilizando tres criterios: localización, idioma y bolsas de palabras de enfermedades que han sido generadas utilizando sinónimos de WordReference y embeddings. Esta demo podría ser de gran utilidad porque los Centros para el Control y la Prevención de enfermedades tardan entre 1-2 semanas desde que se diagnostica al paciente hasta que los datos están disponibles, mientras que con este prototipo se ofrece una monitorización en tiempo real.
Palabras clave: Procesamiento del lenguaje natural, aplicación web, monitorización social, Twitter, word embeddings

1 Introduction and Motivation

Social media has clearly changed how we interact and communicate with each other. There are different mass media in which people published content, such as blogs, social networks, wikis or forums but, currently, social networks are the main one where people express their opinions and experiences. The web has been transformed from a static container of information into a dynamic environment in which users publish any type of information, including ailments and diseases.

In this work, we present *Monge*, a prototype of a geographic monitor of diseases that retrieves tweets located in different Spanish cities, written in Spanish or in Catalan, and allows predict possible epidemic outbreaks of different diseases (e.g. flu, asthma), making use of Human Language Technologies (HLT). This prototype was developed to participate in the *II Hackathon of HLT*[1] that was held on February 26, 2018 in Barcelona, as part of the *Four Years From Now* of the *Mobile World Congress*. It was awarded the second prize in the "General Corpora" category[2].

[1] http://www.agendadigital.gob.es/tecnologias-lenguaje/

[2] https://goo.gl/bSqTcz

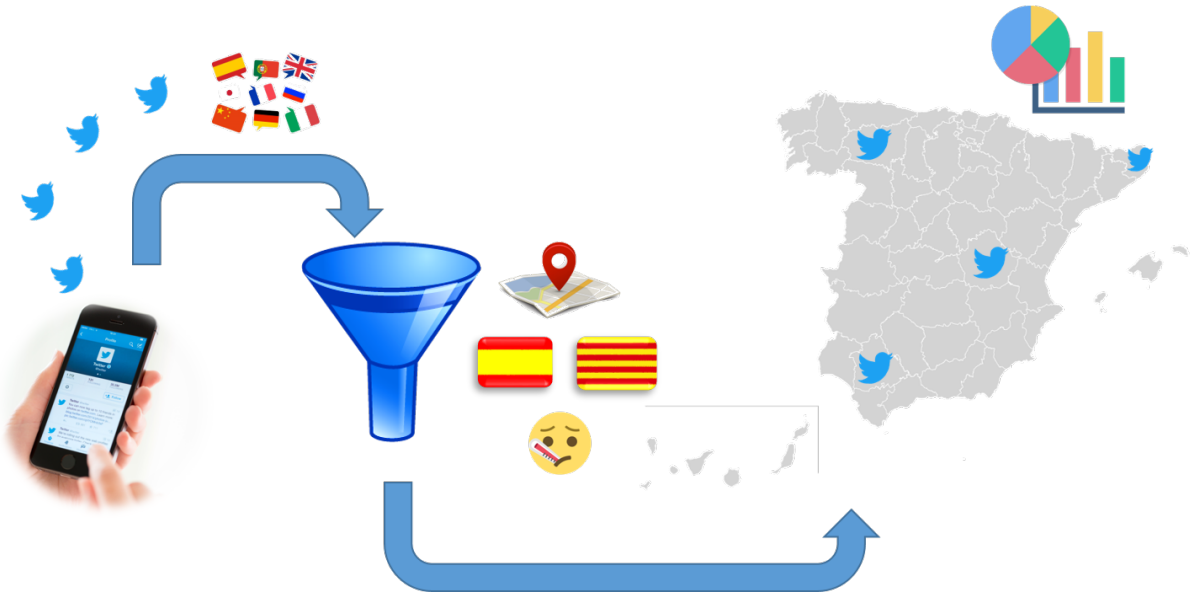

Figure 1: System flow

The reasons that led us to the development of this idea are the fact that most of the Spanish population is connected to the Internet (85% of the Spanish population[3]), performs searches related to health, being diseases one of the main concerns (60% of Spanish Internet users[4]), and uses social networks as way of expression. Therefore, we decided to recover and analyze the publications made in Twitter related to ailments and diseases, because it is one of the main platforms in which people share opinions and experiences (Vilares et al., 2017).

We think that this prototype could be useful because reducing the impact of seasonal epidemics is of vital importance to public health authorities. Studies have shown that effective interventions can be carried out to contain epidemics if there is an early detection. However, the traditional approach used by the Centers for Disease Control and Prevention usually has a delay of 1-2 weeks between the time the patient is diagnosed and the time when the data are available (Achrekar et al., 2011). With this prototype, intoxications, an epidemic or any illness could be detected and monitored in real time at a specific location.

Twitter has been used for real-time notifications such as large-scale fire emergencies, downtime on services provided by content providers (M. Motoyama and Savage, 2010) and live traffic updates. Moreover, there have been efforts in utilizing Twitter data for predicting national mood (Mislove, 2010), currency tracing and performing market and risk analysis.

2 *System description*

This prototype is composed by a back-end module, that deals with the retrieval, filtering and processing of tweets, and the front-end monitor, that shows the analyzed data with different interactive elements, such as a map, a line graph, etc.

2.1 Back-end

The back-end module was developed in Python, and it consists of the following steps (Figure 1):

1. *Seed words selection.* For this prototype we have selected some diseases as seed words, and for each one we have built different Bag of Words (BoW) to filter tweets. These are the 16 diseases selected: ebola, flu, cold, cancer, asthma, hepatitis, otitis, diabetes, caries, anorexia, obesity, Alzheimer's, AIDS, varicella, measles and appendicitis. The system works with two lists of diseases, one with Spanish words and another one with Catalan words.

2. *Tweets retrieval.* This module uses the streaming API of Twitter to recover

[3] https://goo.gl/9AcJtF
[4] https://goo.gl/CJjgmF

tweets in real time that satisfy the three filtering criteria. For this, the tweepy Python library[5] has been used.

3. *Filtering.* Tweets are filtered following these criteria:

   - *Location.* We have defined geoboxes of specific locations (seven cities from Spain) to analyze and locate only the tweets published in these cities.
   - *Language detection.* The system detects the language of each tweet and processes it only if is in Spanish or Catalan. We have used the langdetect Python library[6] that ports of Google's language detection module.
   - *Filtering by disease.* In this final step, tweets are filtered using different BoW that have been generated from the seed words and have been enriched following two approaches:
     - *Using WordReference.* We have created a BoW using the WordReference API[7] to extract synonyms of each disease selected as seed word in the first step. This BoW has been revised manually in order to filter the words related to the human disease. For instance, the synonyms related to the seed word cold are cold, congestion, flu and catarrh.
     - *Using word embeddings.* In this case, the initial list of diseases has been enriched with the 30 most similar words to each of the seed words using two models based on word embeddings. On the one hand, it was used a model generated using a dump of the Spanish Wikipedia (Montejo-Ráez and Díaz-Galiano, 2016). On the other hand, we used the model developed by Cardellino (Cardellino, 2016) with the Spanish Billion Words Corpus and Embeddings[8], which consists of a collection of texts from various corpora and sources (e.g. Wikipedia, Ancora, OPUS Project) with a total of almost 1.5 billion words.

4. *Tweets preprocessing.* Tweets that meet the filtering criteria are preprocessed as follows: they are tokenized using the TweetTokenizer of NLTK[9], all letters are converted to lower-case, and stopwords and punctuations are removed.

5. *Indexing.* At last, the final set of tweets is indexed using ElasticSearch[10].

2.2 Front-end

We have used Kibana[11] to implement the monitor. It is an open-source tool belonging to Elastic, which allows us to visualize, explore and analyze data in real-time that are indexed in ElasticSearch. Kibana is also known for the ELK stack[12] (Elasticsearch, Logstash, Kibana). In this tool, users can create visualizations in the form of tables, charts, maps, histograms, among others. It is useful to create dashboards and helps query data in real time. Dashboards are nothing but an interface for underlying JSON documents. They are used for saving, templating, and exporting. They are simple to set up and use, which helps us play with data stored in ElasticSearch in minutes (Gupta, 2015). In this case, we have created our own dashboard, modifying the default configuration of Kibana.

Our dashboard, shown in Figure 2, is composed of the following elements:

- Spanish map of geographical dispersion of diseases.
- Bar chart with the distribution of diseases by city.
- Table with all the tweets recovered.
- Bar chart with the distribution of diseases by date.
- Line chart with the number of tweets by date and illness.
- Cloud of words most used in tweets.

[5] http://www.tweepy.org/
[6] https://pypi.python.org/pypi/langdetect
[7] http://api.wordreference.com/
[8] http://crscardellino.me/SBWCE/
[9] http://www.nltk.org/api/nltk.tokenize.html
[10] http://elastic.com/
[11] https://www.elastic.co/products/kibana
[12] https://www.elastic.co/elk-stack

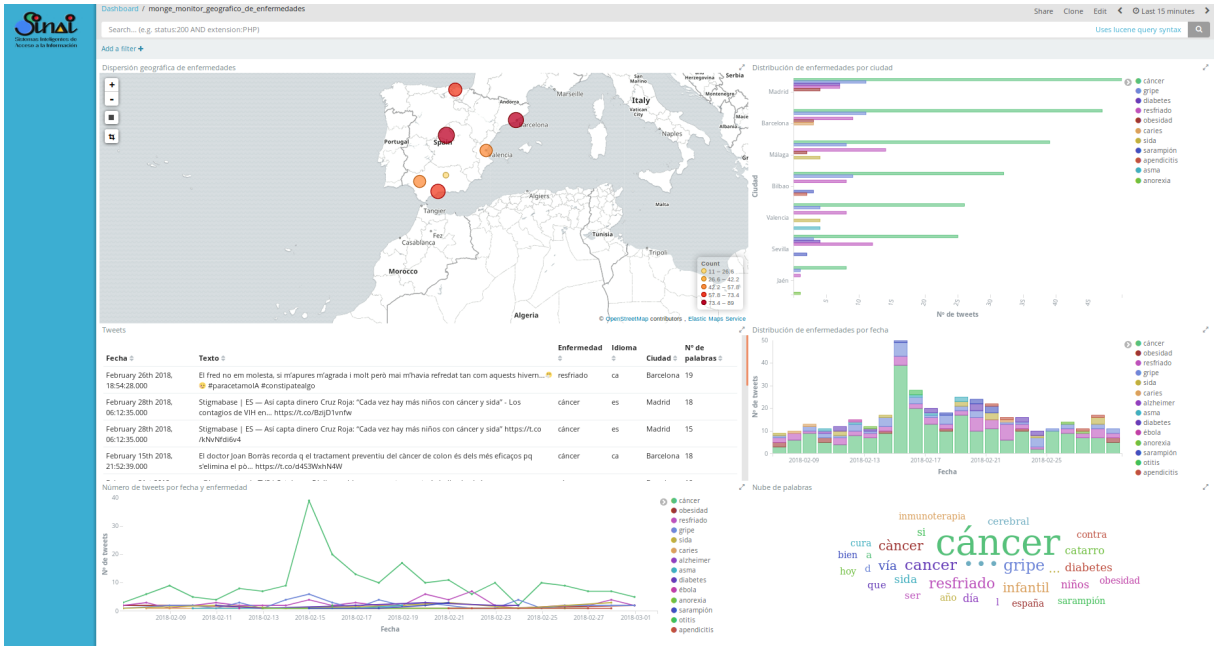

Figure 2: Dashboard

3 Conclusions and Future Work

In this paper, we have presented Monge, a prototype of a geographic monitor of diseases, based on tweets.

With the proposed method we can obtain in real time data to track and predict the appearance and spread of an epidemic in a population. In addition, some public health problems can be monitored, such as intoxications, complaints or, in short, any illness that affects a population in a specific location.

As future work, we want to improve the filtering of diseases in tweets and include the analysis of the presence of negation to improve our results (Martí et al., 2016). Moreover, we want to adapt our system to other different languages.

Acknowledgments

This work has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

Achrekar, H., A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. 2011. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March.

Gupta, Y. 2015. *Kibana Essentials*. Packt Publishing Ltd.

M. Motoyama, B. Meeder, K. L. G. M. V. and S. Savage. 2010. Measuring online service availability using twitter. In *Workshop on online social networks*.

Martí, M. A., M. T. Martín-Valdivia, M. Taulé, S. M. Jiménez-Zafra, M. Nofre, and L. Marsó. 2016. La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural*, 57:41–48.

Mislove, A. 2010. *Pulse of the nation: U.S. mood throughout the day inferred from twitter*.

Montejo-Ráez, A. and M. C. Díaz-Galiano. 2016. Participación de sinai en tass 2016. In *TASS@ SEPLN*, pages 41–45.

Vilares, M., E. S. Trigo, C. Gómez-Rodríguez, and M. A. Alonso. 2017. Tecnologías de la lengua para análisis de opiniones en redes sociales. *Procesamiento del Lenguaje Natural*, 59:125–128.