Intelligent Environments 2018
I. Chatzigiannakis et al. (Eds.)

© 2018 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:10.3233/978-1-61499-874-7-87

The Analysis of Tweets to Detect Natural Hazards

Carlos PERIÑÁN-PASCUAL[a,1] and Francisco ARCAS-TÚNEZ[b]
[a]*Universitat Politècnica de València (Spain)*
[b]*Universidad Católica San Antonio de Murcia (Spain)*

Abstract. During times of disasters, users can act as powerful social sensors, because of the significant amount of data they generate on social media. Indeed, they contribute to creating situational awareness by informing what is happening in the affected community during the incident. In this context, this article focuses on the text-processing module in CASPER, a knowledge-based system that integrates event detection and sentiment tracking. The performance of the system was tested with the natural disaster of wildfires.

Keywords. Twitter, social sensor, topic categorization, sentiment analysis.

1. Introduction

Hazards and disasters give rise to three main types of costs: (a) human cost, since they cause significant suffering and loss of lives, (b) economic cost, since they may result in damage and loss of property, and (c) environmental cost, since they can destroy natural habitats or release pollutants. Because of the increasing public concern on this issue, social media play an active role in disaster detection, tracking, response and assessment. In fact, results from an American Red Cross [1] survey indicated that half of the adults who use social media channels would report emergencies on these channels, and more than two-thirds of the respondents agreed that response agencies should regularly monitor and respond to postings on their websites. For example, *USA Today* reported that, after Houston city officials had warned in August 2017 that emergency services were "at capacity", flood victims decided to use Twitter to ask for help, as shown in the following message:[2]

(1) I have 2 children with me and the water is swallowing us up. Please send help.

As noted by Crowe [2], "initiating protocols and systems to monitor social media conversations—particularly during disasters—is critical for both emergency public information and situational awareness". In fact, for situational awareness, the collection and review of social media information at real time can help emergency managers provide an efficient and effective response to the incident by mobilizing in-situ stakeholders such as fire fighters, police officers or medical staff, among others.

In this context, our research led to the design and development of CASPER (CAtegory- and Sentiment-based Problem FindER). Indeed, this article continues previous research by the authors, where the system was primarily oriented to problem

---

[1] Corresponding author: Universitat Politècnica de València, Escuela Politécnica Superior de Gandía, Paranimf 1, 46730 Gandia (Valencia), Spain; E-mail: jopepas3@upv.es
[2] https://www.usatoday.com/story/news/nation-now/2017/08/27/desperate-help-flood-victims-houston-turn-twitter-rescue/606035001/

detection with Spanish tweets [3]. Following a symbolic approach to topic categorization and sentiment analysis, this new version of CASPER involves not only constructing further resources to analyze English micro-texts but also, and most importantly, enhancing the system to specifically detect hazardous and critical situations that could help guide emergency managers in decision making. According to the EU Vademecum on civil protection,[3] disasters fit into two broad categories: natural disasters (e.g. avalanches, earthquakes, floods, forest fires, hurricanes, storms, tsunamis, and volcanic eruptions) and man-made disasters (e.g. chemical spills, industrial accidents, marine pollution, war and terrorist attacks). This article evaluates the performance of CASPER in relation to the environmental hazard of wildfires. The remainder of this article is organized as follows. Sections 2 and 3 briefly describe some works related to social sensors and the approach of our research, respectively. Section 4 explores the knowledge base developed for the system, whereas Section 5 provides an account of the procedure to detect hazards from micro-texts. Finally, Section 6 evaluates the research, and Section 7 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 [4]. Sakaki et al. [5, 6] 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 center of the event location. As a result, they developed a reporting system to promptly notify people of earthquakes in Japan. Likewise, Liu et al. [7] described a tweet-based 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. It is important to note that most of these systems are trained to detect a single or a few events, e.g. grassfires and floods [8] or swine flu [9], among others.

3. The approach

In this research, hazard detection is going to be addressed as an issue of classification, being comprised of two complementary tasks: topic categorization and sentiment analysis. In this regard, researchers are likely to take one of the following two approaches: a machine learning approach, which is usually implemented through a supervised method, and a symbolic approach, which is primarily based on a knowledge base. A supervised machine-learning method (e.g. 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 hazard). This training dataset should not only be carefully tagged but also be sufficiently large and representative, which actually conflicts with the development of a system like CASPER, which is intended to classify new tweets on the ground of multiple hazards. The effort to expand a given training dataset to fit new categories makes the portability

[3] http://ec.europa.eu/echo/files/civil_protection/vademecum/index.html

of the system 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" [10]. For this reason, the solution was aimed at dealing with hazard detection from a knowledge-based approach.

4. The knowledge base

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 most important resources that were built for our research, i.e. HAZARD, EMERGENCY, SENTIMENT, NEGATION, MODIFIERS and ABBREVIATIONS.

4.1. Hazard, Emergency and Sentiment

CASPER has been designed for two scenarios, i.e. (i) problem detection in general, and (ii) hazard detection in particular. This article is concerned with the latter, which is more likely to take place when tweets are submitted to an emergency management agency, where they should be automatically classified on the basis of the type of incident and the level of emergency. The hazard-detection mode requires three types of lexicon, i.e. HAZARD, EMERGENCY and SENTIMENT, which are briefly described in the remainder of this section.

HAZARD holds lexical descriptors for each hazard (e.g. flood, hurricane, etc), so that their presence in micro-texts leads to topic categorization. For example, some of the descriptors of *wildfire* are *burn, flame*, *grassfire* or *inferno*.

EMERGENCY takes the form of a collection of words that are not specific to any given hazard but are commonly perceived as lexical triggers to activate an emergency response. This dataset was constructed from the keywords in CrisisLex [11] and EMterms [12] after stopwords were removed and was expanded by means of morphological derivation. For example, some of the words in EMERGENCY are *accident*, *dead* or *victim*.

SENTIMENT contains those words that are related to a single sentiment (i.e. positive or negative) regardless of the context in which they are used. This dataset originated from SentiWordNet [13, 14]. SentiWordNet is the result of automatically annotating all synsets (i.e. synonymous sets of words) in English WordNet 3.0 according to their degrees of positivity, negativity and objectivity, where each of the three scores ranges from 0 to 1 and the sum of the three scores is 1 for each synset. In particular, SENTIMENT was originally populated with (i) positively marked words extracted from those terms whose positive score is equal to or higher than 0.8 and the negative score is 0 in SentiWordNet, and (ii) negatively marked words extracted from those terms whose negative score is equal to or higher than 0.8 and the positive score is 0 in SentiWordNet. Those words semantically linked to the resulting synsets were also taken into consideration. Finally, we manually validated the dataset, because it cannot include ambiguous nor context-dependent polarity words. On the one hand, there are words whose polarity is ambiguous when considered out of context, since not all their meanings reflect the same type of sentiment. For example, this is the case of *lofty*, whose sense of "morally good" is positive but its sense of "arrogant" is negative, as illustrated in (2) and (3), respectively.

(2) She was a woman of large views and lofty aims.

(3) He has such a lofty manner.

On the other hand, there are words whose polarity depends on the context, rather than on the meaning. For example, *long* refers to “a large amount of time” in both (4) and (5), but it becomes a positively marked word in the former and a negatively marked word in the latter.

(4) The battery of this camera lasts very long.

(5) This program takes a long time to run.

Therefore, words such as *lofty* and *long* are not included in SENTIMENT. By contrast, some of the words that are actually found in this dataset are *admirably, glad, support* [positive] or *cruel, grief, wreck* [negative].

It should be pointed out that some of the words in HAZARD and some of the words in EMERGENCY can also be found in SENTIMENT. However, no word in HAZARD can be included in EMERGENCY, and no word in EMERGENCY can be included in HAZARD.

4.2. Negation and Modifiers

NEGATION and MODIFIERS compose the main source of knowledge for valence shifters [15], i.e. words and phrases that can affect the values of the hazard, emergency and sentiment attributes of the ngrams in the micro-text.

NEGATION holds negative cues, where most of them can invert the truth value of phrases or sentences (e.g. *lack of*); however, we also found a few of them that do not actually convey negation (e.g. *nothing but*). Therefore, negative cues are classified as negative or non-negative, in addition to specifying the direction of their scope (or impact region), i.e. following or preceding the valence shifter. Negative cues were extracted from different resources: the SFU review corpus [16], Morante’s [17] analysis of the negation cues that occur in the BioScope corpus [18], Morante et al.’s [19] analysis of the negation cues that occur in two Conan Doyle’s stories (i.e. *The Hound of the Baskevilles* and *The Adventure of Wisteria Lodge*), and NegEx triggers [20].[4]

The valence shifters in MODIFIERS are classified as intensifiers or diminishers, i.e. expressions that increase or decrease, respectively, the degree of polarity of the ngrams to which they modify (e.g. *barely*, *significantly* or *very*). The scope of modifiers must also be determined. Modifiers were collected from the English grammar [21].

4.3. Abbreviations

ABBREVIATIONS holds the abbreviations (and their full forms) that are commonly used in social media, such as *btw* -> *by the way* or *thx* -> *thanks*.

[4] NegEx triggers can be downloaded from
https://github.com/mongoose54/negex/blob/master/negex.python/negex_triggers.txt

5. Discovering hazards with CASPER

This section describes the seven stages that take place in CASPER when trying to assign a score to a given tweet in relation to its degree of relatedness with hazards.

In the first stage, the tweets are pre-processed to produce clean texts for natural language processing: (i) reduction of a sequence of three or more repeated characters by means of regular expressions (e.g. gooooood -> good), (ii) spell checking with NHunspell,[5] a library that implements Hunspell [22] for the .NET platform, (iii) transformation of abbreviations into their full-word equivalent with the aid of ABBREVIATIONS, and (iv) removal of hashtags (i.e. any word starting with #), references (i.e. usernames headed by @) and URL links by means of regular expressions.

In the second stage, each micro-text is split into sentences, and then each sentence is tokenized and POS-tagged by using the Stanford Log-linear Part-Of-Speech Tagger.[6] 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 in the micro-text, the word form, the lexeme, the POS, the hazard (h), the emergency (e) and the sentiment (s), where the values of the latter three are discovered in the next stages. We employed the LemmaGen library for lemmatization [23].[7]

The third stage consists in detecting significant ngrams with respect to a given hazard. The weight 1 is assigned to the attribute h of every w_{mn} in T_m whose ngram is found in HAZARD, together with its corresponding POS. The default value is 0.

The fourth stage is aimed at discovering emergency-related ngrams. The weight 1 is assigned to the attribute e of every w_{mn} in T_m whose ngram is found in EMERGENCY, together with its corresponding POS. The default value is 0.

The fifth 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 s of every w_{mn} in T_m according to the polarity of the ngram in SENTIMENT, where the POS of the ngram is also taken into consideration. The default value is 0.

In the sixth stage, valence shifters are applied to neighbouring words within the micro-text. Negation cues make all the ngrams involved in their scope be no longer significant for hazard, emergency and sentiment, so the values of their attributes h, e and s are re-computed to 0. By contrast, intensifiers and diminishers change the degree of polarity of the ngrams involved by multiplying the values of the above attributes by 3 or 0.5, respectively. Whereas negation cues are applied to all the words within the scope, modifiers act only on the first polar expression that is found in the scope. The impact region of the valence shifters is three words, where the direction of this scope is determined by the information included in NEGATION and MODIFIERS.

In the final stage, a problem-relatedness perception index (PPI) is calculated not only to measure how reliable we can feel that a given tweet deals with a problem about a given hazard but also to set alert thresholds from which the severity of the problem could be rated. The computation of the PPI involves three steps. On the one hand,

---

[5] NHunspell was downloaded from https://sourceforge.net/projects/nhunspell/
[6] The Stanford POS Tagger was downloaded from https://sergey-tihon.github.io/Stanford.NLP.NET/StanfordPOSTagger.html
[7] LemmaGen was downloaded from http://lemmatise.ijs.si

considering that the lexical descriptors for a given hazard form a vector of features (i.e. f_1, f_2, ..., f_k), cosine similarity is used to assess the degree of relatedness between the tweet and the hazard. Since we deal with the binary values of the attribute h and the number of distinct hazard-related ngrams in the tweet T_m is equal to or less than the number of lexical descriptors for the hazard, the hazard-relatedness function can be simplified to the Eq. (1).

$$rel_h(T_m) = \frac{\sum_{n=1}^{p} w_{mn}}{\sqrt{\sum_{n=1}^{p} w_{mn}} \times \sqrt{\sum_{j=1}^{k} f_j}} \tag{1}$$

Therefore, a tweet is related to a given hazard if the similarity score is greater than 0. On the other hand, a logit scale is used to compute the sentiment score, as shown in the Eq. (2).

$$\begin{gathered} rel_s(T_m)' = \log\left(\frac{P+0.5}{N+2D+0.5}\right) \\ \text{if } rel_s(T_m)' < 0, \text{ then } rel_s(T_m)'' = -rel_s(T_m)' \\ \text{otherwise, } rel_s(T_m)'' = 0 \end{gathered} \tag{2}$$

where P and N refer to the total value of positively and negatively marked ngrams in T_m, respectively (calculated from the attribute s), and D refers to the number of emergency-oriented words in T_m (calculated from the attribute e). The normalized value is derived from the Eq. (3).

$$rel_s(T_m) = 1 - \frac{1}{\log(rel_s(T_m)'' + 2)} \tag{3}$$

Finally, as shown in the Eq. (4), the PPI is computed as the geometric mean of the values returned by rel_h and rel_s so as to reach a proportional compromise between topic categorization and sentiment analysis.

$$PPI(T_m) = \sqrt{rel_h(T_m) * rel_s(T_m)} \tag{4}$$

6. Evaluation

This research was evaluated with a corpus of 1,200 tweets posted during a devastating series of wildfires that occurred in Colorado throughout June, July and August 2012 [24].[8] The tweets in this dataset were labeled by crowdsourcing workers according to three parameters: informativeness (e.g. related and informative, related but not informative, not related, or not applicable), information type (e.g. affected individuals, infrastructure and utilities, donations and volunteering, caution and advice, sympathy and support, other useful information, not applicable, or not labeled), and information source (e.g. eyewitness, government, NGOs, business, media, outsiders, not applicable,

[8] The dataset was downloaded from https://github.com/sajao/CrisisLex/tree/master/data/CrisisLexT26

or not labeled). Table 1 presents the distribution of tweets with respect to informativeness, which is the only parameter relevant to the research in this article.

Table 1. Informativeness in the 2012 Colorado wildfires dataset.

	Related and informative	**Related but not informative**	**Not related**	**Not applicable**	**Total**
Tweets	685	268	238	9	1,200

At first sight, it might be thought that only "related and informative" tweets could really be useful for the task at hand, since they are supposed to be the only ones that help understand the crisis situation on the ground. However, this proved to be a rather subjective category, as shown in examples (6) and (7), which were manually categorized as "related and informative" and "related but not informative", respectively.

(6) Theres like 7 fires in colorado right now....

(7) Ack! A fire now in Boulder!

In this experiment, CASPER managed to identify 633 fire-related tweets, whose distribution with respect to informativeness and PPI scores is shown in Table 2 and Figure 1, respectively.

Table 2. Informativeness in the experiment results.

	Related and informative	**Related but not informative**	**Not related**	**Not applicable**	**Total**
Tweets	474 (74.88%)	146 (23.06%)	12 (1.90%)	1 (0.16%)	633 (100%)

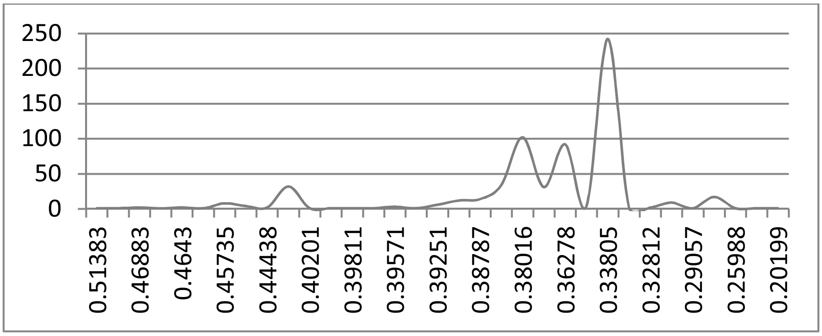

Figure 1. PPI scores in the experiment results.

We employed precision to evaluate the performance of the system, as formulated in Eq. (5).

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Precision is a key issue in the development of emergency-response systems, 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. The manual validation of the results revealed that precision was 0.8073. To prioritize hazardous and critical situations for effective emergency management, we chose to automatically rank tweets by arranging them from the highest PPI score (i.e. 0.51383) to the lowest PPI score (i.e. 0.20199), whose corresponding micro-texts are shown in the examples (8) and (9), respectively.

(8) Colorado fire: 41,140 acres burned, 1 dead: Firefighters were hoping to get control Tuesday of a fast-moving wildfire in northern Colorado

(9) Please RT! Help My Friends in CO .Great way to help support Colorado Fire

To this end, we employed five ranges (i.e. R1-R5) to organize the 33 distinct PPI scores. Figure 2 serves to illustrate the amount of tweets found within each range for each informativeness value.

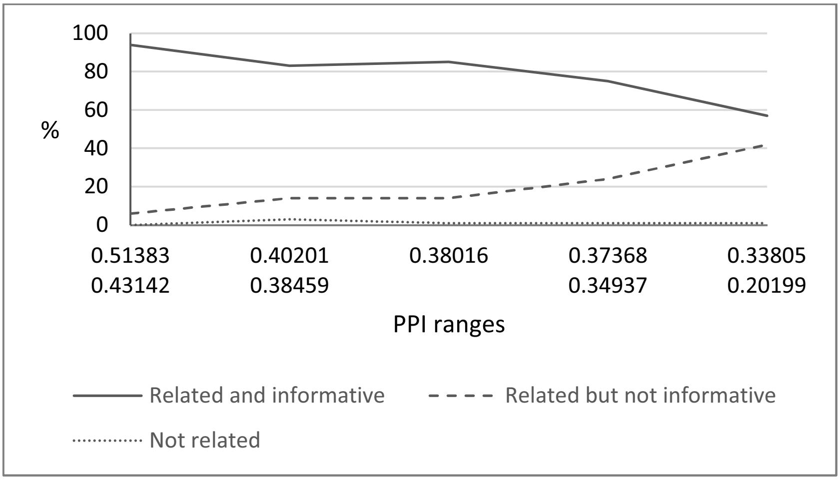

Figure 2. Informativeness in PPI ranges.

It can be noted that the graph lines in Figure 2 reflect a gradual distribution of informativeness, which is in line with the discriminating power of positioning critical situations at the top of the rank, while minor or non-existing problems concentrate closer to the bottom of the list. This is demonstrated in Table 3, which shows the cumulative precision along the ranges, together with the number and percentage of tweets in each range.

Table 3. Cumulative precision in PPI ranges.

Range	Precision	Tweets
R1	0.9074	54 (8.53%)
R1-R2	0.8984	128 (20.22%)
R1-R3	0.8913	230 (36.33%)
R1-R4	0.8627	357 (56.40%)
R1-R5	0.8073	633 (100%)

In this manner, for example, when CASPER retrieves the top-ranked 128 tweets, i.e. about 10% of the 1,200 tweets analyzed, precision is near 0.9, which contributes to developing an effective notification system for emergency managers.

Figure 3 displays the duration of the ten wildfires that occurred in Colorado throughout June and July 2012 (horizontal bars). The dashed line represents the average PPIs derived from the tweets submitted on each date (vertical bars). This chart demonstrates that the peak areas of PPI are located in the first halves of the two most destructive fires: High Park (9 June–30 June) and Waldo Canyon (23 June–8 July).

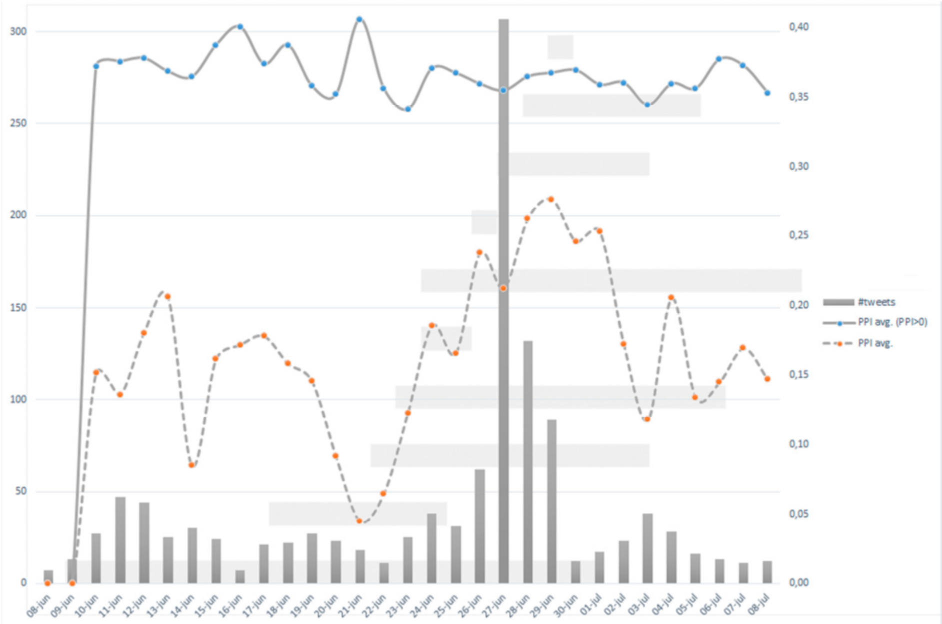

Figure 3. PPI scores over time.

7. Conclusion

During and after disasters, people use microblogging services (e.g. Twitter or Facebook) to communicate actionable information that can help emergency responders gain situational awareness. In this regard, we described both the knowledge base and the natural language processing techniques that allowed us to develop a system that serves not only to classify micro-texts according to particular types of hazards but also to compute a score (PPI) for each micro-text to assess the disaster impact (i.e. damage to people, property or environment). Indeed, the evaluation of the research demonstrated that PPI scores can be used to effectively select the most relevant tweets to emergency response and recovery.

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] American Red Cross. *Social Media in Disasters and Emergencies*, 2010 Available at:

http://i.dell.com/sites/content/shared-content/campaigns/en/Documents/Red-Cross-Survey-Social-Media-in-Disasters-Aug-2010.pdf.
[2] A. Crowe, *Disasters 2.0. The Application of Social Media Systems for Modern Emergency Management,* Boca Raton: CRC Press, 2012.
[3] C. Periñán-Pascual and F. Arcas-Túnez, *A knowledge-based approach to social sensors for environmentally-related problems*, Workshop Proceedings of the 13th International Conference on Intelligent Environments, IOS Press, Amsterdam, 2017, 49-58.
[4] C. Castillo, *Big Crisis Data. Social Media in Disasters and Time-Critical Situations*, New York: Cambridge University Press, 2016.
[5] 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.
[6] 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.
[7] S.B. Liu, B. Bouchard, D.C. Bowden, M. Guy and P. Earle, *USGS tweet earthquake dispatch (@USGSted): using twitter for earthquake detection and characterization*, AGU Fall Meeting, 2012.
[8] S. Vieweg, A.L. Hughes, K. Starbird and L. Palen, *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, 2010, 1079-1088.
[9] A. Signorini, A.M. Segre and P.M. Polgreen, The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS ONE 6 (5) (2011).
[10] A. Moreno-Ortiz and C. Pérez Hernández, *Lexicon-based sentiment analysis of twitter messages in Spanish*. Procesamiento del Lenguaje Natural 50 (2013), 93-100.
[11] A. Olteanu, C. Castillo, F. Diaz and S. Vieweg, *CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises,* Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM'14). AAAI Press, Ann Arbor, MI. 2014.
[12] I. Temnikova, C. Castillo and S. Vieweg, *EMTerms 1.0: A Terminological Resource for Crisis Tweets,* Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM'15). Kristiansand, 2015.
[13] A. Esuli and 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.
[14] 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.
[15] L. Polanyi and A. Zaenen, *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, 2004, 106-111.
[16] N. Konstantinova, S. de Sousa, N. P. Cruz, M. J. Maña, M. Taboada and R. Mitkov, *A review corpus annotated for negation, speculation and their scope*, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC). Istanbul, Turkey, 2012.
[17] R. Morante, *Descriptive analysis of negation cues in biomedical texts,* Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), European Language Resources Association (ELRA), Valletta, 2010.
[18] V. Vincze, G. Szarvas, R. Farkas, G. Mora and J. Csirik. *The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scope,* BMC Bioinformatics, 9 (11): S9, (2008).
[19] R. Morante, S. Schrauwen and W. Daelemans, *Annotation of Negation Cues and their Scope Guidelines v1.0*. Computational Linguistics and Psycholinguistics Research Center, University of Antwerp, 2011. In: https://www.clips.uantwerpen.be/sites/default/files/ctrs-n3.pdf.
[20] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper and B.G. Buchanan, *A simple algorithm for identifying negated findings and diseases in discharge summaries*, Journal of Biomedical Informatics, 34 (2001), 301–310.
[21] COBUILD, *Collins COBUILD English Grammar*, Glasgow: HarperCollins, 2005.
[22] L. Nemeth, V. Tron, P. Halacsy, A. Kornai, A. Rung and I. Szakadat, *Leveraging the open source ispell codebase for minority language analysis,* Proceedings of SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages, 2004.
[23] M. Juršič, I. Mozetič, T. Erjavec and N. Lavrač, *LemmaGen: multilingual lemmatisation with induced Ripple-Down rules*, Journal of Universal Computer Science, 16 (9) (2010), 1190-1214.
[24] A. Olteanu, S. Vieweg and C. Castillo, *What to Expect When the Unexpected Happens: Social Media Communications Across Crises,* Proceedings of the ACM 2015 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '15). ACM, Vancouver, BC, Canada, 2015.