Nataliya D. Pankratova (Ed.)

System Analysis and Information Technologies

18-th International Conference SAIT 2016
Kyiv, Ukraine, May 30 – June 2, 2016

Proceedings

Institute for Applied System Analysis
of National Technical University of Ukraine
"Kyiv Polytechnic Institute"

***Teodorescu H.N.*[1] *Bolea C.*[2]**
[1]*Romanian Academy, Iasi Branch, Iasi, Romania;* [2]*Institute of Computer Science, Romanian Academy, Iasi Branch, Iasi, Romania*

Analysis of probabilities of specified words' occurrences in SN messages related to catastrophes

Recently, social networks (SNs) have become an important source of information for social sciences, linguistics, marketing, and even for technology and for combating the effects of catastrophic events [1–5]. This interest in SNs requires improvements in the methods and tools for data acquisition on SNs [3–5]. Making decisions and forecasts based on data collected from SNs asks for high precision in relevant data collection and specific data discovery. Redundant data reduce the efficiency of information extraction, while missing data or data misinterpretation may compromise the decision.

Various types of catastrophic events, such as earthquakes, floods, aviation crashes and terrorist acts may produce huge increases of the traffic on SNs [3, 4]. Consequently, several research groups proposed methods for making use of this information [1–5]. While several authors have attempted the analysis of the vocabulary and the statistics of the words in messages related to such events for English and a few other major languages, we are not aware of any extended research on the statistics of the language (language model) specific to SNs related to catastrophic conditions, except a few very recent papers [6–8]. Moreover, such studies are almost completely missing for 'smaller' languages such as Romanian.

In previous papers, e.g., [7], we dealt with various issues related to the vocabulary of the Internet and social networks, including frequently used terms on specified SNs, in relation with natural disasters (earthquakes in [7], earthquakes and flood in other papers). Also, this research complements those results and the study of the jargon of Social Networks and in general on Internet; the latter study produced an *Internet Slang Annotated Dictionary* (ISAD) which is freely available at the address http://iit.academiaromana-is.ro/isad/.

The examination of the vocabularies of messages related to disasters on various SNs has several reasons. In the first place, it helps selecting the best keywords and search logical conditions; second, it provides clues on the main issues perceived by the public in relation with disasters, moreover on the sentiments and attitudes prevailing in these conditions. Third, from a linguistic point of view, it allows us study the specificity of the idiomatic expressions and jargon used on SNs in general, on individual SNs, and in relation with specified disaster conditions.

In the current investigation, we are specifically interested in the following:

- probabilities of the stopwords in specific categories of SN messages, i.e., in SNs messages that relate to a specific type of disaster, $p(w \in Stop | w \in m, m \to D)$, where $p(|)$ denotes conditional probability, w is a word, $Stop$ denotes the set of stopwords, m denotes a message, and $m \to D$ means that the message refers to a specific type of disaster D;
- probabilities of the most frequent words that are not stopwords,
- ratios of the probabilities of specified couples of synonyms, e.g., *earthquake* and *seism*, or *flood* and *inundation*, in messages on SNs;
- probabilities of co-occurrences of specific words, $p(w_1 \wedge w_2)$, for example $w_1 =$ *earthquake* and $w_2 =$ *collapsed*;
- probabilities of co-occurrences of pairs of words that comprise one geographic name, for example earthquake and *Vrancea*.

The present study refers only to earthquakes and to messages and posts (on social media) in Romanian related to seisms. Beyond the probabilities of synonyms, we are also interested in the analysis of the degree of synonymy and of the contexts that determine synonymy.

The authors acknowledge the partial support of the NATO Science for Peace and Security Programme under grant G4877.SPS grant G4877 (*Modeling and Mitigation of Social Disasters Caused by Catastrophes and Terrorism.*)
Authors' contribution: HNT conceived the research and methods, and wrote the abstract; CB gathered the data and performed the analysis with the help of HNT.

Synonyms and synonymic expressions are important because they are frequently used. Messages with synonyms must be interpreted accordingly, without loosing information and without erroneous interpretation. Their use increases the difficulty of the information extraction task. When pairs of words are synonyms only in specific contexts, the complexity of their role in the analysis increases significantly; this is because of the need of added context analysis.

Among others, the analysis of synonyms on SNs is aimed to determine the context where two words are truly synonyms, respectively the contexts making them different. In a previous research (HNT, unpublished, Apollonia Congress, Iaşi, March 4, 2016), the first author has shown that synonymy and related meanings obey context-dependent rules, providing as a typical example the Romanian words *cutremur* and *seism* which are synonyms only when the first is a noun. Although the word *seism* obeys S. Marcus' rule that the literary and scientific languages tend to converge, it seems that it enters the common language only in some restricted contexts. The statistics performed tends to support this hypothesis. In another direction of research, also of interest was the differential use of synonyms under normal circumstances and under surges (spikes) of SNs' activity (see [9]), due to disasters.

For example, after a destructive earthquake, finding where are located the large collapsed buildings and where are located victims is crucial. But such information can be expressed with a large number of almost synonym words. In fact, one can say about a building that "s-a dărâmat" (collapsed), "s-a prăbuit" (downfall), "s-a năruit" (fell), "s-a ruinat" (ruined). Similarly, casualties may be described with many words, such as casualty victim decease fatality injured person wounded person loss dead person (in Romanian, 'victime, decedati, raniti, pierderi, morti etc.).

Supplementary materials, including bags of words, can be found on the SRoL site described in [10] and at the above mentioned web address for ISAD.

References. **1.** Pankratova N.D., Modelling and mitigation of social disasters of various nature. Int. Conf. on System Analysis and Information Technology SAIT 2015, Kyiv, Ukraine, June 22–25, 2015, ISBN 978-966-2748-69-7, pp. 13-14. **2.** Pankratova N.D., Pyvovar V.V., Big Data Overview. Int. Conf. on System Analysis and Information Technology SAIT 2015, Kyiv, Ukraine, June 22–25, 2015, ISBN 978-966-2748-69-7, pp. 35. **3.** H.N.L. Teodorescu, On the Responses of Social Networks to External Events. Proc. ECAI 2015 7th IEEE Int. Conf. on Electronics, Computers and Artificial Intelligence, 25-27 June, 2015, Bucharest, Romania, pp. 13-18, DOI: 10.1109/ECAI.2015.7301138. **4.** H.N. Teodorescu, Using analytics and social media for monitoring and mitigation of social disasters. Procedia Engineering Vol. 107C, (2015) pp. 325-334, DOI 10.1016j.proeng.2015.06.088, Reference: PROENG17426. **5.** Pankratova N.D., Savastiyanov V., Foresight Process based on Text Analytics. Int. J. "Information Content and Processing", Vol. 1, Number 1, 2014, pp. 54-65. **6.** Petic M., Cojocaru S., Vocabulary enriching for text analysis. 17-th Int Conf. on System Analysis and Information Technology SAIT 2015, Kyiv, Ukraine, June 22–25, 2015, pp. 37-38. **7.** S.C. Bolea, Vocabulary, synonyms and sentiments of hazard-related posts on social networks. SPED 2015, 8th IEEE Int. Conf. on Speech Technology and Human-Computer Dialogue, Oct 14-17, 2015, Bucharest, Romania. **8.** M Pirnau Tool for monitoring Web sites for emergency-related posts and post analysis. SPED 2015, 8th IEEE Int. Conf. on Speech Technology and Human-Computer Dialogue, Oct 14-17, 2015, Bucharest, Romania. **9.** H.N.L. Teodorescu, Emergency-Related, Social Network Time Series – Description and Analysis. In: I. Rojas, H. Pomares (Eds.), Advances in Time Series Analysis. Series Contributions to Statistics, Springer, 2016. **10.** SM Feraru, HN Teodorescu, MD Zbancioc, SRoL-Web-based resources for languages and language technology e-Learning. International Journal of Computers Communications and Control 5 (3), pp. 301-313.