LREC 2016 Workshop

EMOT:
Emotions, Metaphors, Ontology and Terminology during Disasters

PROCEEDINGS

Edited by

Khurshid Ahmad, Stephen Kelly, Xiubo Zhang

24 May 2016

Proceedings of the LREC 2016 Workshop
"EMOT – Emotions, Metaphors, Ontology and Terminology during Disasters"

24 May 2016 – Portorož, Slovenia

Edited by Khurshid Ahmad, Stephen Kelly, Xiubo Zhang

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 607691.

Organising Committee

- Khurshid Ahmad, Trinity College Dublin, Republic of Ireland*
- Stephen Kelly, Trinity College Dublin, Republic of Ireland
- Xiubo Zhang, Trinity College Dublin, Republic of Ireland

*: Main editors and chairs of the Organising Committee

Programme Committee

- Khurshid Ahmad, Trinity College Dublin, Republic of Ireland
- Gerhard Heyer, Leipzig University, Germany
- Bodil Nistrup Madsen, Copenhagen Business School, Denmark
- Maria Teresa Musacchio, University of Padova, Italy
- Henrik Selsoe Sorensen, Copenhagen Business School, Denmark
- Carl Vogel, Trinity College Dublin, Republic of Ireland

Preface

An unexpected event induces an emotive reaction and metaphorical use of language; natural disasters, a class of unexpected events, induces the use of disaster-specific terminology and ontological descriptions in addition to emotive/metaphorical use of language. Disasters are characterised equally well by the fact that the onset, duration, and aftermath of this event, all to a greater or lesser degree, involve greater demand of information in a situation where information is generally scarce. One has to use all modalities of communications, including written and spoken language, visual communications, and non-verbal communications especially gestures.

The authors contributing to this workshop have been working on how to specify, design and prototype disaster management systems that use social media, including microblogging and social networks, as one of the inputs. Their linguistics coverage includes English, German and Italian. Their focus is on extracting information from continuous data streams including text, speech and images. There are two papers on emotional expressions used in disasters: Busa and Cravotta have analysed gestures and speech of reporters working in areas threatened by natural disasters like floods and suggest that hand gestures and the modulation of voice is correlate with the severity of an impending disaster. Spyropoulou looks at the terminological and affect content of text messages and speech excerpts and finds that speech comprises more information about the sentiment of the public at large than, say, their text messages. Topic modelling is one of the essential techniques that can be used to automatically categorise the contents of a text data stream of messages – this technique uses machine learning and information extraction, and has been successfully used by Schlaf, Gründer-Fahrer and Jähnichen to analyse German social media texts, especially on Facebook and Twitter: they find that the machine discovered categories that correspond quite ‘naturally’ to the categories of texts used in disaster management – warnings before disasters, re-location information during disasters, and requests after the disasters.

Vogel presents a theoretical discussion of the relationship between emotions and metaphors, and how affect-based language, laden with sentiment, is used. The selective use of terminology and ontology play a key role in a methodology form detecting impending and current emergencies in social media streams which has been developed by Musacchio, Panizzon and Zorzi. Finally, we have a paper that deals with the automatic extraction of terminology and ontology from text especially social media by Zhang et al: These authors have designed, implemented and evaluated an ontology/terminology extraction system (CiCui).

The authors appear to be aware of the problems relating to the factual and ethical provenance of data, especially on social media, involving as it does issues such as privacy, dignity, copyright ownership, rumours and many other legal and ethical considerations.

The presenters wish to acknowledge the support of the EU FP7 Programme focussing on the impact of social media in emergencies. The research leading to these results has received funding from the European community’s Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL 2014:2017)

K. Ahmad May 2016

Programme

	Opening Session
09:20 – 09:30	Introduction
09:30 – 09:40	Khurshid Ahmad Emotion, Metaphor, Ontology and Terminology
	Emotion
09:40 – 10:05	M. Grazia Busa, Alice Cravotta Detecting Emotional Involvement in Professional News Reporters: an Analysis of Speech and Gestures
10:05 – 10:30	Maria Spyropoulou Emotive and Terminological Content in Disaster Related Messages
10:30 – 11:00	Coffee break
	Metaphor
11:00 – 11:25	Carl Vogel Emotion, Quantification, Genericity, Metaphoricity
	Ontology
11:25 – 11:50	Antje Schlaf, Sabine Gründer-Fahrer, Patrick Jähnichen Topics in Social Media for Disaster Management – A German Case Study on the Flood 2013
	Terminology
11:50 – 12:15	Maria Teresa Musacchio, Raffaella Panizzon, Xiubo Zhang, Virginia Zorzi A Linguistically-driven Methodology for Detecting Impending and Unfolding Emergencies from Social Media Messages
12:15 – 12:40	Xiubo Zhang, Raffaella Panizzon, Maria Teresa Musacchio, Khurshid Ahmad Terminology Extraction for and from Communications in Multi-disciplinary Domains
	Closing Session
12:40 – 13:00	Khurshid Ahmad

Table of Contents

1 Emotion

Detecting Emotional Involvement in Professional News Reporters: An Analysis of Speech and Gestures

Maria Grazia Busà, Alice Cravotta
University of Padua
Via B. Pellegrino, 26
35137 Padova
E-mail: mariagrazia.busa@unipd.it, alice.cravotta@unipd.it

Abstract

This study is aimed to investigate the extent to which reporters' voice and body behaviour may betray different degrees of emotional involvement when reporting on emergency situations. The hypothesis is that emotional involvement is associated with an increase in body movements and pitch and intensity variation. The object of investigation is a corpus of 21 10-second videos of Italian news reports on flooding taken from Italian nation-wide TV channels. The gestures and body movements of the reporters were first inspected visually. Then, measures of the reporters' pitch and intensity variations were calculated and related with the reporters' gestures. The effects of the variability in the reporters' voice and gestures were tested with an evaluation test. The results show that the reporters vary greatly in the extent to which they move their hands and body in their reportings. Two gestures seem to characterise reporters' communication of emergencies: beats and deictics. The reporters' use of gestures partially parallels the reporters' variations in pitch and intensity. The evaluation study shows that increased gesturing is associated with greater emotional involvement and less professionalism. The data was used to create an ontology of gestures for the communication of emergency.

Keywords: Non-verbal behaviour, gestures, speech analysis, emergency communication.

1. Introduction[1]

Studies have shown that in the communication of affect (i.e., emotional involvement), words account for under 10% of the meaning exchanged, while around 40% of the meaning is transmitted through paralinguistic features of the speakers' voice (e.g., pitch, voice volume), and about 50% through body language (Mehrabian & Wiener, 1967; Mehrabian & Ferris, 1967; also reported in Mehrabian, 1972). Different modalities lend themselves to representing certain kinds of information better than others. For example, the hands express shapes better than speech; the face expresses attitudes better than words.

Investigating how people communicate emotion through their use of non-verbal language is important on theoretical grounds as well as for extracting data that can be used in information analytics systems. Such systems must take into account that communication involves different modalities (written, spoken, non-verbal). This is of particular importance today since the social media provide an increasing amount of content through audios and videos.

While it seems feasible to extract multimodal data from audios and videos, there is a lack of a solid body of research on the relation between acoustic features of speech and gestures in emotive communication that would help getting information about the emotive state of the speaker.

Investigations of emotive vocal behaviour (Scherer, 1986, 2003; Juslin & Scherer, 2005; Juslin & Laukka, 2003) have associated certain changes in voice acoustic patterns to basic emotive states. As far as body and gestures are concerned, research has focussed more on facial expressions (Ekman & Friesen 1977; Ekman, 1993; Ekman et al., 2013) than on the expression of emotions through body postures and gestures. Facial expressions have been shown to be universally associated to a set of basic emotions, while gestures and body posture were left apart in defining the quality of an affective state (De Gelder, 2009). But there is evidence that emotions are manifested with a synchronized response incorporating physiology, speech, facial expressions, modulations of posture and gestures, affective vocal behaviours and actions.

The present study was carried out within the EU FP7 Security Programme sponsored Slandail project. The aim of the project is to make ethical use of the information available in the social media to enhance the performance of emergency management systems. As part of this project, research is carried out aimed at extracting and integrating text, image, speech and non-verbal data from the social media.

In this framework, the analysis of integrated speech and gesture data can be used (i) directly, as indicators of the speakers' emotional involvement as an effect or a reaction to a disaster and, (ii) indirectly, to obtain information about the gravity of the disaster.

This study aims to examine whether speech and non-verbal language can be used to extract data on speakers' emotional involvement. For this purpose, it investigates a corpus of videos on journalists reporting on natural disasters. The study consists of (1) a qualitative analysis of the gestures performed by the reporters in their reportings; (ii) an acoustic analysis of some speech characteristics (pitch, intensity) of the reporters' voices; (iii) an evaluation of the relation between the reporters' speech characteristics and their levels of body dynamism; (iv) an evaluation of the effects of the reporters' speech

[1] The research leading to these results has received funding from the European community's Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL)

and gestures on the audience public. The study also draws the ontology of the two gestures that appear most commonly in emergency communication.

2. Detecting Frequent Gestures in the Reporting of Disasters: A Qualitative Study

Journalists reporting on emergency situations must appear professional and not emotionally involved with the event they are reporting. However, while they may manage to do so with their verbal language (choice of words, discourse structure), they may communicate emotional involvement with their body gestures. It is in fact possible that the more they are involved in the situation they are reporting the more this involvement will leak out from their body movements.

This led to formulate our first set of hypotheses, that is: 1) reporters convey emotional involvement in a disaster situation through their non-verbal language; 2) when they are emotionally involved in the reported event, reporters will tend to use some gestures more frequently than others.

To test these hypotheses, the study reported below was carried out.

2.1 Methods and Materials

2.1.1. The iTVR corpus

We first proceeded to create a corpus of videos for the analysis. We decided to search for videos of journalists reporting on the flooding that took place in Liguria (Italy) in October and in November 2014 in the Italian news portals Skytg24, Rainews24 and RaiTv.

The following criteria were followed to include videos in the corpus:

- The reporter was well visible (at least the arms, hands and face);
- The video was of relatively good quality;
- The audio was of relatively good quality;
- The communicative situation was homogeneous (reporter talking to the camera, no interaction with people; scenery in the background);
- The reporter was describing and/or talking about a natural event, like a disaster;
- The reporter was in sight for at least 10 seconds without interruptions (e.g., there were no interruptions such as footage showing the scenery that was being referred to);
- There were at least three videos of each reporter reporting the same event in different moments.

38 TV-news reports of variable duration by 9 Italian journalists (4 men and 5 women) were chosen for inclusion in the corpus. The videos were captured from full screen streaming using the Camtasia Studio 8 screen recorder software, which also records the system audio directly from sound card preserving the original audio quality. This corpus will be referred to as iTVR (Italian TV Reports corpus).

The multimodal annotation software ELAN (Wittenburg et al., 2006) was used to analyse the videos.

2.2 Analysis

A detailed qualitative analysis of the reporters' body language was carried out. The analysis focussed on the reporters' hand and arm gestures, though a note was made of the reporters' gaze, posture and body movements when these were conspicuous.

On the basis of this inspection, we defined three categories of speakers' dynamism, aimed at reflecting the extent to which the speakers were moving during their reports:

'Level 1'. The reporter is relatively still, and she hardly moves her arms and hands. When she wants to point at something, she does it at most with a movement of her eyes and gaze.

'Level 2'. The reporter is relatively still, but she is moving one arm and hand to emphasize a part of her report or a particular point in the scenery to which she is drawing the audience's attention.

'Level 3'. The reporter is moving constantly, and she turns a part of or the whole body together with her arm and hand to emphasize her discourse or to point to particular points in the scenery.

2.3 Results

2.3.1. Identification of Frequent Gestures in Emergency Communication

The reporters in the videos were assigned to the three levels of dynamism on the basis of the differences in the reporters' extent of gesturing

We identified two classes of arm and hand gestures that are distinctive of the different levels of dynamism.

The first is beats, that is, the rhythmic beating of a finger, hand or arm to accompany speech. Typically, beats involve up-and-down or back-and-forth hand movements that coincide with spoken clauses, breaks, or sentence ends (fig.1). Beats are not present in Level 1 videos, instead they are rather distinctive of Level 2 and 3, where the reporter emphasises words, sentences or speech rhythm in general.

Figure 1: Beat gestures

The second type of gestures that occur frequently in the corpus is pointing gestures (or deictics), that is, gestures that the journalists use to point to some place that they are referring to. This pointing gesture may be done with different degrees of 'extensiveness': from pointing that is merely hinted at with the gaze; to pointing that is done with the arm and hand while the journalists' body continues to stand still and face the camera; and finally to pointing that makes the journalists turn away from the camera and towards the place/situation that they are describing (fig. 2).

Figure 2: Pointing modes

These different extents in degree of movement correspond to the three identified levels of dynamism.

2.4 Discussion

Reporters appear to alternate between moments in which they gesture less (level 1) to moments when they gesture more (levels 2 and 3). Two types of gestures occur most frequently in the reporting of disasters. These are beats and deictics. When reporters use fewer gestures, they make little to no use of beats and deictics; when they gesture more they make movements that involve the whole body and extensive use of beats and deictics.

The frequent use of beats and deictics in this type of communication can be explained.

Beats are frequent in politicians' speeches (McNeill, 1992), especially when they have a cohesion function (when they serve to mark different points which are supposed to be crucial and coherent). Also, beats can appear to mark the word or phrase which introduces new characters, summarises the action, introduces new themes, etc. Neither politicians' speeches nor news reports involve a dialogical exchange or a real interaction with the interlocutor; the speakers - especially in tv reports - cannot count on back-channel feedbacks from the addressee and they probably need to mark the structure of the speech for clarity. Also, it has been shown that radio broadcasting news, for instance, can be characterised by "circumflex" intonation (a regularity in the use of pitch contours) and a constant and regular emphatic stress on words and syllables (Rodero, 2013). This typical rhythm that people commonly associate with news reading may also enhance the occurrence of beats, whose main function is rhythm beating.

As for deictics, it is quite natural for journalists reporting from a disaster site to point to the places and situations they are referring to, whether for clarity or to suggest to the cameraman where to shot. Also, their speech often presents adverbs referring to places and locations (here, there, etc.) that are naturally accompanied by pointings.

Research has shown that gestures and speech are interconnected (e.g., Goldin Meadow, 2005; Kendon, 1980). According to McNeill (1992), gestures and speech are synchronous at the semantic level, as they are co-expressive of the same underlying meaning, at the pragmatic level, as they co-occur to express the same pragmatic function; and at the phonological level, as gestures are temporally coordinated with the phonology of the utterances. Gestures and speech may also be constrained by the same contextual factors, accounting for individual differences, speakers' emotional involvement, etc. This is, however, still largely unexplored. Finally, the use of gestures during speech (co-speech gestures) is largely unconscious (Cienki & Müller, 2008). This makes body movements a way to explore underlying thoughts and emotive states of the speaker.

It is thus possible that when reporters' are reporting on disasters, their gesturing reflects their emotional involvement in the situation. It is also possible that the reporters' involvement will also show in some of the characteristics of the reporters' voices, such as pitch and intensity. The investigation of these aspects is the object of our next study.

3. Relating Reporters' Speech and Gestures: A Pilot Experiment

As discussed in the previous section, it is possible that reporters' use of gestures while communicating a disaster may betray different levels of emotional involvement. Specifically, increased gesturing may reflect an increased level of involvement in the situation, as an effect of their reduced control over their body language (reporters are likely to have learned to control their body language as part of their professional training).

Since gestures are synchronised with speech, it is likely that the increase in the reporters' gesturing may parallel an increased variability in the reporters' speech characteristics, and particularly those that are related to affect, such as pitch and intensity. These acoustic cues, in particular, are known to correlate with the emotional states of the speaker's voice (Juslin & Laukka, 2003).

To investigate these issues, this study addresses the relation of the reporters' differences in body dynamism to the acoustic correlates of pitch and intensity in the reporters' voices.

The aim was to test the following hypotheses:

H1: Variations in the reporters' body gesturing are paralleled by variations in pitch and intensity;

H2: The variation in the reporters' gestures, pitch and intensity can be perceived by the viewer; thus, this variation can be interpreted as a signal of emotional involvement;

H3: Extensive body gesturing is inversely related to perceived professionalism.

These hypotheses were tested in a two-part study. The first part is an acoustic analysis of the reporters' pitch and intensity variation patterns. These are then related to the differences in body dynamism that we observed in the qualitative analysis. The second part is an evaluation study testing the perception of the observed variation in the reporters' speech and gestures.

3.1 Procedure

3.1.1. Materials

7 videos were selected randomly for each level of body dynamism (section 2.2) from the iTVR corpus. This created a corpus of 21 video samples. The duration of the videos was shortened to 10 seconds to include only the part that was most representative of the reporters' body dynamism and to allow the creation of an evaluation test

that was not too long.

3.1.2. Methods

The audio signal was extracted from the 21 videos. An acoustic analysis was carried out with Praat (Boersma, 2001). For both pitch and intensity, the listings of all the values were saved as a .txt file for further analysis.
For the evaluation test, three sets of stimuli were prepared. The first set used the 10-second videos but muted (Mute condition). The second set had audio tracks but no videos (Audio-only condition). The third set used the videos with the audios (Video condition).
In all the videos, the background around the reporters was removed with Adobe Premiere, so that the speakers appeared to speak on a black background (fig. 3). This was done to ensure that the background scenery (with views of the disaster) did not influence the participants' evaluation.

Figure 3: Example of a 10-second video with black background

Each set of stimuli contained a randomised list of the 21 items and was preceded by 4 videos selected additionally to create a trial session.
The test was administered online through Google Forms. The participants in the experiment were supposed to express - on a 1 to 5 point scale – their evaluation of: (a) the reporters' degree of involvement in the situation reported; and (b) the reporters' degree of professionalism. Forty evaluators completed the test.

3.2 Analysis

In the acoustic analysis the pitch and intensity listings were obtained for the 10-second audios with Praat.
The values were used to calculate the means and standard deviation (SD) of both pitch and intensity. For both acoustic cues, the SD was divided by the mean to normalize the data and remove inter-subjects differences, to obtain a measure called Pitch Variation Quotient (PVQ) (Hincks, 2004) and, by analogy, an Intensity Variation Quotient (IVQ).
The mean values of PVQ and IVQ were calculated for each level of dynamism, and the values for each level were compared to verify whether the reporters' variations in pitch and intensity parallel the reporters' variations in gesturing (H1).
For the evaluation study, all the scores obtained for each stimulus were averaged by level of dynamism. To verify the perception of the reporters' variation in speech and gestures, the mean scores were related to (1) the three levels of body dynamism that we had identified in the qualitative study and (2) the results of the acoustic analysis (H2 e H3).
At this stage of the investigation no statistical analyses have been carried out. These will be carried out on a larger data sample.

3.3 Results

Table 1 shows the Pitch Variation Quotient (PVQ) and the Intensity Variation Quotient (IVQ) in relation to the three levels of dynamism that were presented in Section 2.2

Level	PVQ*	IVQ**
1	0.16	0.10
2	0.24	0.10
3	0.22	0.08

* Pitch Variation Quotient; ** Intensity Variation quotient.
Level **1**= "Idle"; Level **2** = Beats only (still body); Level **3** = Beats, Pointings, Body moves.

Table 1: PVQ and IVQ in relation to the three levels of speakers' body dynamism.

The results confirm H1 only partially. For pitch, the values of PVQ show that there is an increase in overall variation from Level 1 to Level 2, while the values for Level 3 are slightly lower than those of Level 2. As for intensity, the data do not provide support to our hypothesis, and show equal intensity values for Level 1 and 2 and slightly lower for Level 3.
Table 2 shows the mean values of the evaluators' ratings of the three sets of stimuli. The data show some interesting trends.

Level	M*		A**		V***	
	a	**b**	**a**	**b**	**a**	**b**
1	2.41	3.44	2.83	3.46	2.51	3.40
2	3.46	3.36	3.44	3.11	3.33	3.02
3	3.44	3.02	3.46	3.21	3.44	2.99

*Mute condition; ** Audio only condition; *** Video condition.
a= question about involvement; b= question about professionalism.

Table 2: Mean values of the evaluators' ratings of the three types of stimuli: Muted videos (M), Audio only (A), regular videos (V).

All reporters were considered to be less involved at Level 1 of body dynamism than in the other levels (Column a) in all conditions. In other words, reporters that move little were perceived as less involved than reporters that move more while reporting on emergency situations. This confirms H2. The fact that the tendency is true also in the audio only condition suggests that listeners are able to interpret speakers' voice qualities as involvement in the same way as they do with gestures and body movements. In fact, the data suggest that when no video is present, the audio data has a stronger effect than the video, at least at Level 1 of body dynamism. As far as professionalism is concerned (Column b) an opposite trend can be detected: the more a reporter moves the less he/she is perceived as professional – as shown by the values decreasing from Level of body dynamism 3 to level 1. This confirms H3.

3.4 Discussion

We predicted that variation in both pitch and gestures increases as speakers increase their involvement in the reported event. The results of our study provide only partial support to our hypothesis. However, it should be noted that the audio data extracted from the videos were rather noisy, due to the fact that the speakers were reporting from emergency scenes, and so the background noise may have affected the results of the acoustic analysis.

One of the aims of our evaluation experiment was to test whether greater gesturing is perceived as greater involvement. The results confirm our hypothesis and show that, when reporters gesture more or more extensively (see section 2.2), they are perceived as more emotionally involved in the situation they are reporting. Also, there is an inverse relation between perceived involvement, as is reflected by gesturing, and perceived professionalism: the more the reporters in the videos were using gestures the less professional they were judged. This result confirms our hypothesis that extensive body gesturing is inversely related to perceived professionalism.

4. Towards an ontology of gestures used in the communication of emergencies

As reviewed in the introduction, research on non-verbal language can be used to extract data on speakers' emotional involvement. In this work we identified two gestures that seem to be characteristic of emotional reporting in emergency communication. In this section the ontology of these two gestures is proposed.

4.1. Ontology of Beats and Pointing Gestures

Our analysis showed that the two gestures that are most commonly used in emergency communication are beats and deictics. To develop the ontology for these gestures we drew on influential nonverbal classification schemes (Efron, 1941; Ekman and Friesen, 1969; McNeill 1992; Kendon 2004). The ontology was then refined by reference to other notation systems and coding schemes that were recently developed in association with automatic human gesture recognition and synthesis (e.g., Bressem, 2008; Kipp, Neff & Albrecht, 2007).

The ontology that was developed for beat gestures consists of a decision tree (Figure 4) that can be considered a 'filter' for defining the gesture.

A beat gesture has to satisfy some necessary hierarchical conditions to be defined as a beat gesture:

- It has to be a repetitive movement that follows the speaker's speech rhythm. This movement has a certain frequency and duration that can be measured;
- It has to be a straight movement. It can be an up-down or a back-forth movement;
- The movement has to show a certain muscle tension. High-level labels (e.g., flat hand/spread or single finger/bent) represent hand shapes and orientation and can be assigned to any gesture (Figure 5).

A decision tree of the same model was developed for pointing gestures. We took into consideration the most prototypical pointing gesture, that is the one performed to with the hand or the arm, excluding other pointing strategies (e.g., gaze). The schema lists all the necessary conditions that must be satisfied to assign the pointing label to a gesture.

The following conditions have to be hierarchically satisfied:

- The gesture needs to trace a well-defined path. It means that it needs to be a movement that follows a clear direction, i.e., moves clearly towards something;
- The final part of the movement is usually linear (not circular or spiral);
- The gesture is usually held for a while at its furthest extent;
- Hands must have a certain shape or a certain muscle tension. High-level labels (e.g., index finger - palm down or index finger - palm up) represent hand shapes and orientation and can be given to any gesture.

5. Conclusion

This study aimed to examine whether speech and non-verbal language can be used to detect the speakers' emotional involvement. For this purpose, it analysed a corpus of videos on journalists reporting on natural disasters.

The study shows that, in their emergency reports, journalists vary their voice characteristics and gestures considerably. In particular, their voice may span from lower to higher levels of pitch and intensity; their bodies also show different levels of dynamism. Two gestures seem to be characteristic of this kind of reporting: beats and deictics. The results of an evaluation test showed that the reporters' variations in voice pitch and gesturing are perceived as a signal of emotional involvement and affect the perception of the reporters' professional image.

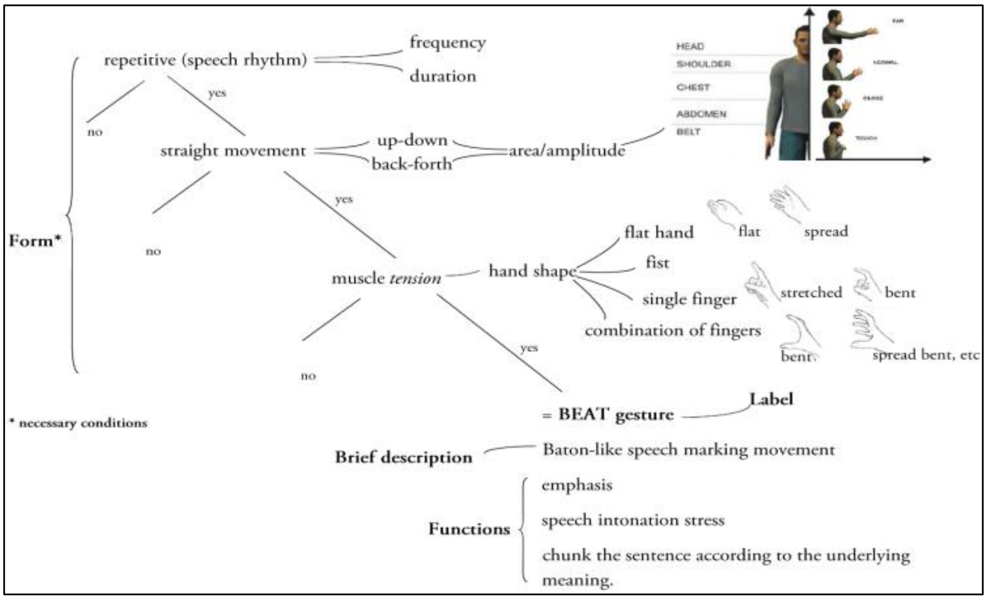

Figure 4: Beat gestures decision tree. Hands and body illustrations from Bressem (2008) and Kipp et al. (2007).

Our results should be considered preliminary, and more work will done to expand this research. We are planning to gather quantitative data on a wider repository of gestures (head, gaze, and other categories of hand gestures) and extract information about gesture occurrences, duration, latency, etc. Also, we will carry out additional acoustic analyses, including measurements of speech rate and vocal perturbation. With more data we will also run statistical analyses and investigate the correlation between gestures and voice cues.

Though preliminary, the results of this study show the importance of studying emotional involvement that can be expressed *beyond* the speaker's words.

Reporters play a key role in the delivery of information. In emergencies, reporters' voice or body language may reflect feelings of anxiety or fear that are not conveyed by the words alone. This may impact on the way the message is received by the public, with consequences on their actions or thoughts. Thus, this research can provide information that is useful for training end-users and spokespeople in emergency communication.

Finally, the evidence that speakers' modify their voice and

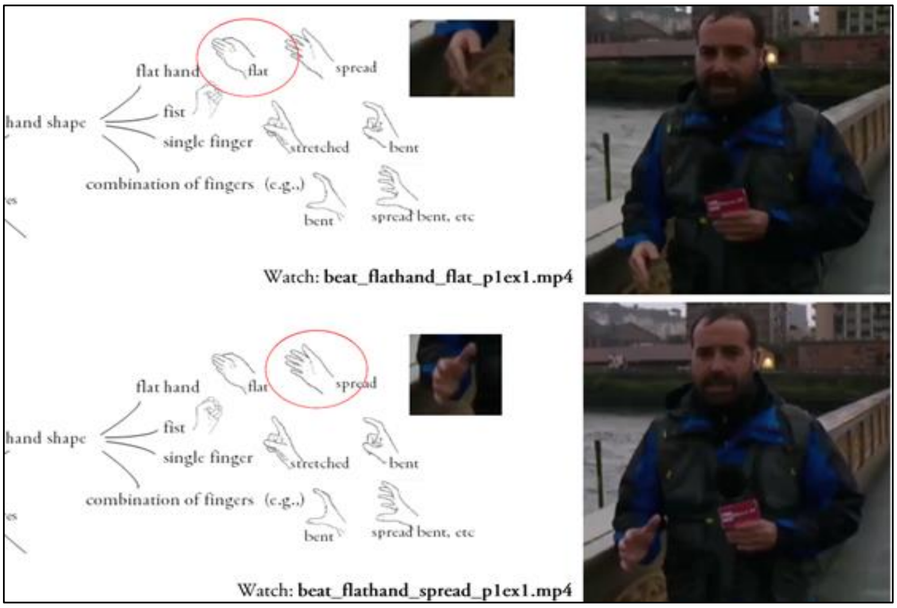

Figure 5: Beat Gestures. Hand shapes and gestures areas. Hands illustrations from Bressem (2008).

body language patterns as an effect or a reaction to a disaster can be used to obtain indirect information about the gravity of the disaster. These data can be used for improving the communication protocols in an emergency management system.

6. References

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341-345.

Bressem, J. (2008). Notating gestures—Proposal for a form based notation system of coverbal gestures. *Unpublished manuscript. Retrieved from http://www. janabressem. de/publications. html*.

Cienki, A., & Müller, C. (2008). *Metaphor and gesture.* John Benjamins Publishing.

de Gelder, B. (2009). Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society of London.Series B, Biological Sciences, 364*(1535), 3475-3484. doi:10.1098/rstb.2009.0190 [doi]

Efron, D. (1941). Gesture and environment.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*(1), 49-98.

Ekman, P., & Friesen, W. V. (1977). Facial action coding system.

Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings* Elsevier.

Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think* Harvard University Press.

Hincks, R. (2004). Processing the prosody of oral presentations. *InSTIL/ICALL Symposium 2004,*

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*(5), 770.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. *The New Handbook of Methods in Nonverbal Behavior Research,* , 65-135.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The Relationship of Verbal and Nonverbal Communication, 25*, 207-227.

Kendon, A. (2004). *Gesture: Visible action as utterance* Cambridge University Press.

Kipp, M., Neff, M., & Albrecht, I. (2007). An annotation scheme for conversational gestures: How to economically capture timing and form. *Language Resources and Evaluation, 41*(3-4), 325-339.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought* University of Chicago press.

Mehrabian, A. (1972). *Nonverbal communication* Transaction Publishers.

Mehrabian, A., & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology, 31*(3), 248.

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology, 6*(1), 109.

Rodero, E. (2015). The principle of distinctive and contrastive coherence of prosody in radio news: An analysis of perception and recognition. *Journal of Nonverbal Behavior, 39*(1), 79-92.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1), 227-256.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. *Proceedings of LREC, 2006* 5th.

Emotive and Terminological Content in Disaster Related Messages

Maria Spyropoulou
Trinity College Dublin
Lloyd Institute
Maria.Spyropoulou@scss.tcd.ie

Abstract

As the availability of information during a disaster is low, research has started to focus on other modes of communication that can complement text in information extraction and sentiment analysis applications. This paper attempts an initial estimation on what kind of results we should expect to get from disaster related text and speech data.

Keywords: disaster terminology, sentiment analysis, named entity recognition, Twitter

1. Background

The dissemination of information about a disaster is primarily done through text and secondarily through videos, images and audio data. The analysis of textual formal and social media data, like RSS feeds and Twitter tweets appears to be very useful for information extraction purposes before or after a disaster. But during an emergency the text data traffic is complemented by video and audio data that are equally easily and quickly captured and distributed. Video and audio data have the distinct advantage that they present perspectives on the real feel and real implications of a disaster, because they contain sound and image.

Emergency calls (911/999) in particular could play a huge part in disaster mitigation if their content was captured, analysed and searched for target terms. Looking at disaster data, the target terms cover two fundamental categories: named entities, for instance people, locations, equipment and objects, and affect terms, namely emotional states and positive or negative connotations. Our assumption is that text analytics provide the basis for named entity recognition, while speech data is perhaps more emotional and direct, thus being ideal for sentiment analysis.

For this paper, disaster related textual data from Twitter and transcripts of emergency calls or crisis related audio were compared and the results were analysed, to conclude which source of information helps for a more constructive evaluation for emergency situations. This work is tied to the research being carried out for Slándáil, an FP7 project focusing on the impact of social and formal media in emergency situations. One of Slándáil's main objectives is to create a framework under which data can be ethically harvested in order to helpemergency management agencies make more informed decisions and aid in disaster mitigation and recovery.

2. Motivation

Speech processing as a source of information in disaster management has been used in the past (Backfried et al., 2012). The basic concept is based on audio segmentation and automatic transcription of speech, which is then passed to the text analytics system for further analysis like topic generation and named-entity recognition. The use of radio and online audio as a source of information through automatic media monitoring was also investigated under the EU funded FP5 project ALERT[1]. Although the use of visual, textual and audio-visual modalities is the standard, aural resources are not being used to the same extent. Indicatively, in Federal Emergency Management Agency's (FEMA[2]) multimedia library, out of the approximately 41,000 available items, only 252 are audio, so 0.6% of the whole library. FEMA's audio files are uploaded as an archive.

An example of live audio feed is Broadcastify[3], the world's largest source of live audio stream on public safety. Broadcastify offers a communication platform between the emergency responders, primarily the fire brigade and police departments, and the public, by transmitting the signals from the police scanners in frequencies that the police departments agree to share. Police dispatch scanners can give information on police operations, fire brigade, highway patrol, emergency medical services and other emergency management agents' actions. All this information could be captured in times of emergency, processed, indexed and passed to a text analytics system. Police scanners could be used to add information about the place and the people involved in an event, as suggested by (Joslyn, Hogan & Robinson, 2014).

This type of streaming though should be carried out with caution, as the information shared through a police scanner is frequently calls that the call center receives from witnesses, information that is not intended for the public and needs to be investigated further. This can cause many false alarms as the social and sometimes even the formal media will spread rumours based only on the broadcasts from these police scanners without double-checking their validity (Franks & Evans, July 2015; Tapia, LaLone & Kim, 2014). Another problem that needs to be addressed is

[1] http://www.alert-project.eu/
[2] http://www.fema.gov/
[3] http://www.broadcastify.com/

highlighted in (Crampton et al., 2013): Twitter spam advertising bots might take advantage of trending hashtags that include the name of a police scanner related to a trending event to promote products, creating noise for information extraction engines.

Speech recordings of 911/ 999 calls from affected civilians that are isolated, trapped, injured or witness a disaster could be used as a source of information. Not only can the type, location and time of the particular catastrophe be used to inform emergency managers but so can the identification of victims or possibly endangered civilians. The calls that hospitals, police and fire brigades receive at their call centres and scanners could be recorded and processed through a system that analyses their speech and returns optional and necessary information for the emergency management agencies.

Automatic call summarization could prove to be faster and more accurate than human conversational call transcription. Among other research programs, FP7 SENSEI explored the efficacy of call centre conversation tabular summarization implementing both content and emotion mining (Favre, Stepanov, Trione, Béchet & Riccardi, 2015, September). Naturally, the majority of research on call center transcriptions data mining is based on business transactions (Clavel et al., 2013; Garnier et al., 2008; Mishne et al., 2005; Takeuchi, Subramaniam, Nasukawa & Roy, 2009). Another idea that has yet to be implemented is speech driven dialogues interfaces for crisis management (Sharma et al., 2003).

3. Case Studies

We chose to test disaster related text and speech data to get an estimation on how transcribed speech data would help emergency management agents in times of disaster. In the context of the Slándáil speech analytics module, an evaluation was performed on commercial state-of-the-art software suites to see how well they can handle good quality audio but also telephony quality audio, as a test to check their effectiveness for use in an emergency situation. Three ready-to-use speech recognition software suites (VoiceBase[4], VoxSigma[5], PopUpArchive[6]) were tested with approximately one hour of data, 28 minutes of which was telephony quality with a cut-off frequency at 4 KHz and 26 minutes of which were studio quality with a cut-off frequency of 10 KHz and above.

The data was taken from the FEMA audio library and contained speech excerpts from disaster related podcasts, announcements, guidelines and conference calls. Three highly dramatic emergency calls from the 2011 Texas floods were included as well, extracted from YouTube. Our initial hypothesis that speech recognition systems cannot handle telephony quality speech due to background noise was confirmed, as the Word Accuracy for these files was an average of 50% with a standard deviation of 0.17. For the higher quality files the average Word Accuracy was 87% with a standard deviation of 0.07. These results are not very encouraging at the moment for emergency call machine transcription, but higher quality audio feed could be extracted and used as an alternative source of information. As the generated transcriptions of spontaneous low quality audio excerpts were not very reliable, only the reference transcriptions were taken into account for this evaluation, approximately 8,000 tokens.

Regarding text data, a small Twitter corpus was extracted and compared with the speech corpus transcriptions to yield results of disaster and affect terms. 24,000 tweets were collected in the period 2/12/2015 to 01/02/2016 and were scanned for disaster and affect terms. The reason these months were chosen is because throughout December and January Ireland was hit by two major storms, Desmond and Jonas. As a result, the Irish counties were devastated by strong winds, heavy showers, flooding and power outages. The aftermath of the storms and the recovery efforts were a hot topic discussed on social and formal media for the two month period. The disaster terminology subcategories used to extract specialist terms included *avalanches, floods, earthquakes, storms and volcanoes*. Names of places and objects were also recognized. In total a number of 500,000 tokens were collected. Example material from both datasets used in the evaluation is presented in Table 1.

3.1 Processing of Data

The percentages of disaster and affect terms in all the files were summed and then standardized with a unity-based normalization from 0 to 1. The normalization function was x_i*-min(range)/ max(range)-min(range)*, where x_i is the individual value. The tweets were split into four equal periods and the averages of the normalized values were extracted for these periods. It is evident (Figure 1) that for the first three periods, the reference of disaster terms was equal or bigger to affect terms. Affect terms surpassed disaster terms only on the fourth period, when little disaster related activity was present. The speech material was split into eight files, half of which were mitigation related, while the other half were primarily financial recovery related.

[4] http://www.voicebase.com/
[5] http://www.vocapia.com/voxsigma-speech-to-text.html
[6] https://www.popuparchive.com/

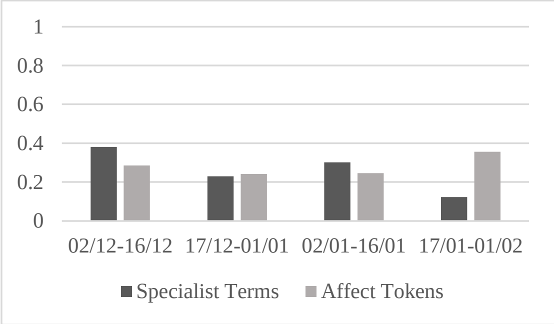

Figure 1: Normalized frequencies of specialist terms and affect tokens in Irish tweets

3.2 Results

File 1 in Figure 2 contained the transcription of three emergency calls and as shown, it had no reference to disaster terms as defined in the Disaster Terminology database, but a high number of affect terms, as it was rather emotional speech from affected civilians during the 2011 Texas floods. The civilians did not use any terms contained in the terminology database to refer to specific events during the disaster, but used expressive phrases like *'our house is down we're floating'* and *'we're running out of breathing room'*.

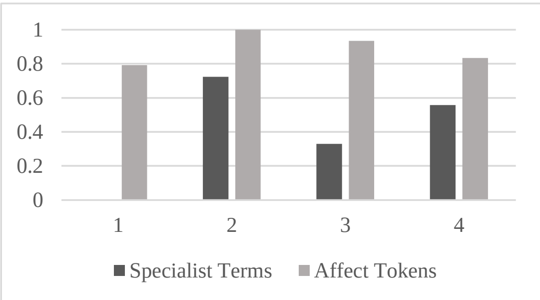

Figure 2: Normalized frequencies of specialist terms and affect tokens in mitigation related speech

These phrases are highly colloquial and only a very sophisticated information extraction software suite that could handle contextual inference would be able to categorize them as negative terms, since *breathing* and *floating* as individual terms have a rather positive connotation on their own. For this study it doesn't seem to be very important, as the affect terms are explored as a whole, but for a sentiment analysis tool such phrasing could lead to false inferences.

Files 2 and 3 from Figure 2 were read text, but they contained instructions on course of action in cases of emergency, which resembled the instructions given in the 911 calls. For instance, the 911 operators were telling people to keep away from the attic so that they don't get stuck there. Similarly, one instruction of file 3 was to plan two ways out of a room in case of a fire and move furniture that could block an escape path. Finally, file 4 of Figure 2 was a podcast on disaster relief legal planning in case of emergency, which contained a large amount of positive terms like *assistance, aid, understand, provide, protect* and *safety*.

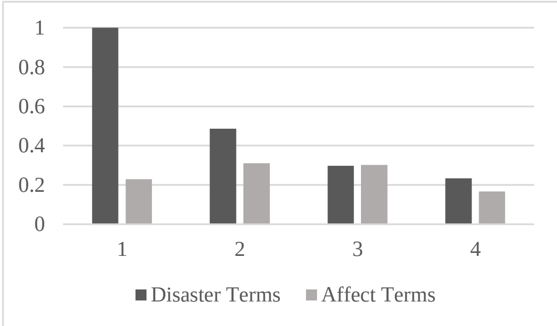

Figure 3: Normalized frequencies of specialist terms and affect tokens in recovery related speech

For reference, the disaster and affect averages of spontaneous and semi-spontaneous finance oriented files are shown in Figure 3. Specifically, file 1 of Figure 3 was an excerpt from a conference call on insurance policies from FEMA. Similarly, file 2 of Figure 3 was a webinar for stakeholders. Finally, files 3 and 4 of Figure 3 contained discussions on governmental funding for equipment, such as helicopters, and protection, such as fencing. The use of affect terms was in general equal or smaller in comparison to the use of disaster related terms, which seems to be expected for finance related transcriptions.

It should be noted as well, that a large number of the affect terms in files 2, 3 and 4 of Figure 3 were conversation related, such as *sorry* (*I'm sorry*), *let* (*let me check*), *thank* (*thank you*), *invite* (*invite you to comment*), *welcome* (*you're welcome*) and *clarify* (*I want to clarify*). The amount of such terms was a lot higher in these discussions than it was in the emergency related calls, where there was no time for courtesies, while *thank* (*thank you*) and *problem* (*no problem*) happened to be the only conversation related affect terms.

Table 1: Examples of Named Entity Recognition and Sentiment Analysis in speech and text

		Specialist Terms and Affect Tokens
Speech	**Mitigation and Emergency**	- Hays County 911 do you **need** police fire or EMS? - Hi I'm not sure where we are located. We're on the Blanco River in Wimberly and the water is **up** to the second story of the house. - Ok what's your address, the address of the **emergency**? - One hundred Deer Crossing Deer Crossing Deer Crossing. - Ok and what is your name? - Laura McComb.

	Financial Recovery	- Yes, so I'm **sorry**, when you say you have to go through your state, you're part of the urban area, you're part of the Detroit urban area? - Yes. - Yes, so I don't have a point of contact in front of me for the Detroit urban area working group, but do you participate in or does anybody in the county participate in the urban area working group? - I think, I'm pretty sure our director with homeland **security** does, yes.
Text	**Tweets**	rain **clearing** eastwards showers following most frequent in southern western coastal counties **windy** for a time relatively mild lows 6 9c **cold** at first with many places dry but some local **heavy** showers. troops from an chéad cath filling sandbags in Ballinasloe tonite to **support** flood **relief** efforts #stormdesmond **interesting** storm Desmond rainfall animation from NASA

4. Conclusion

From this preliminary study, it is implied that disaster related speech might be more suitable as a source for sentiment analysis. This is enforced by the fact that the text data tokens were immensely larger in number than speech data tokens, but speech data tokens still appeared to include a richer variety of affect terms.

Another point to be made is that the averages of affect speech in the disaster related textual data and financial recovery related speech data in their majority are 0.3 and less, which suggests that disaster related textual data is generally uninformative of the public feeling of a situation. In contrast, mitigation and emergency related speech appears to be more informative for sentiment analysis, even though the tokens of that part of the speech corpus were far less than the tokens of the text corpus.

5. Acknowledgements

The current paper was completed under the FP7 Project SLÁNDÁIL, a collaboration between academic and commercial partners with the goal to improve the use of social media in enhancing the response of disaster related agencies. The research leading to these results has received funding from the European community's Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL). I want to thank prof. Khurshid Ahmad (Trinity College Dublin) for his precious remarks and advice.

6. References

Backfried, G., Schmidt, C., Pfeiffer, M., Quirchmayr, G., Glanzer, M., & Rainer, K. (2012, August). Open source intelligence in disaster management. In 2012 European Intelligence and Security Informatics Conference (pp. 254-258). IEEE.

Clavel, C., Adda, G., Cailliau, F., Garnier-Rizet, M., Cavet, A., Chapuis, G., & Guillemin-Lanne, S. (2013). Spontaneous speech and opinion detection: mining call-centre transcripts. *Language resources and evaluation*, *47*(4), 1089-1125.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and geographic information science*, *40*(2), 130-139.

Favre, B., Stepanov, E., Trione, J., Béchet, F., & Riccardi, G. (2015, September). Call centre conversation summarization: A pilot task at Multiling 2015. In 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (p. 232).

Franks, P., & Evans, L. (2015, July). Social Media and Trust in North American Local Government law Enforcement. In *Proceedings of the 2nd European Conference on Social Media 2015: ECSM 2015* (p. 157). Academic Conferences Limited.

Garnier-Rizet, M., Adda, G., Cailliau, F., Gauvain, J. L., Guillemin-Lanne, S., Lamel, L., & Waast-Richard, C. (2008, May). CallSurf: Automatic Transcription, Indexing and Structuration of Call Center Conversational Speech for Knowledge Extraction and Query by Content. In *LREC*.

Joslyn, C., Hogan, E., & Robinson, M. (2014). Towards a topological framework for integrating semantic information sources. *Semantic Technologies for Intelligence, Defense, and Security (STIDS)*.

Mishne, G., Carmel, D., Hoory, R., Roytman, A., & Soffer, A. (2005, October). Automatic analysis of call-center conversations. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 453-459). ACM.

Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Brewer, I., ... & Sengupta, K. (2003). Speech-gesture driven multimodal interfaces for crisis management. *Proceedings of the IEEE*, *91*(9), 1327-1354.

Takeuchi, H., Subramaniam, L. V., Nasukawa, T., & Roy, S. (2009). Getting insights from the voices of customers: Conversation mining at a contact center. *Information Sciences*, *179*(11), 1584-1591.

Tapia, A. H., LaLone, N., & Kim, H. W. (2014). Run amok: group crowd participation in identifying the bomb and bomber from the Boston marathon bombing. *Proceedings of the 11th ISCRAM*

2 Metaphor

Emotion, Quantification, Genericity, Metaphoricity

Carl Vogel
Centre for Computing and Language Studies
Computational Linguistics Group
School of Computer Science and Statistics
Trinity College Dublin, the University of Dublin
vogel@tcd.ie

> "...the opposite of reason is not emotion; it is unreason. And the opposite of emotion is not reason; it is frigidity on the one hand and sentimentality on the other."
> (John Barth, *The Tidewater Tales*, p. 163 Methuen/Minerva 1987.)

A respected tradition in the analysis of emotions, and subcategories thereof, regards them as modulators of rational behavior, if not directly a species of irrationality (see Brady (2009)). For example, at first glance, the fact that people have emotive responses to situations described in works of fiction, and which they know to be fictional, appears to entail that people are acting irrationally when they indulge in such responses. However, it has been argued that fiction provides an occasion for people to engage in hypothetical reasoning and empathy (Dadlez, 1996), and in these terms, the claim of irrationality dissipates. Another alternative to ascribing irrationality to emotion is to study the logic of emotive expressions.

A modest path in studying the logic of emotive expressions is suggested by the program of truth-conditional semantics via formal logic. In the program truth-conditional semantics, one worries partly about the conditions in which expressions are true and partly about what other sentences have to be true if one commits to some sentence as being true. One accepts that the meaning of terms like "good", "love", "happiness" and "friendship" varies with interpretation and requires extensive analysis but notes that one has little hope of explicating "good" if one cannot explain the meaning of more mundane words like "and" or "not". Similarly, in the context of linguistic manifestations of emotion, it is helpful to focus for a time on categories of expressions that are not open, like "love" or "honor", but rather those which are relatively more closed, like "at all" or "somewhat". The expressions of primary focus here are therefore quantificational.

Dynamic selection of domains of quantification and predication is a very creative dimension of language use. It has been argued (Vogel, 2011) that metaphoricity and genericity in natural language are united as modulators of senses of open class terms, predicates, with the essential behavior of genericity being to restrict the meaning of extant predicates to smaller sets for which universal quantification holds true,[1] and with the essence of metaphor being the expansion of predicate denotations to larger sets.[2] Thus, it is argued that metaphoricity and genericity exist as mutual duals. Inasmuch as metaphoricity and emotion are commonly analyzed together, and as genericity is a species of hyperbole, it is sensible to reconsider these semantic phenomena in this framework. However, both metaphoricity and genericity hinge on open class categories, and thus do not exhaust the discussion of semantic analysis of emotion. The complement analysis derives from analysis of closed-class items. While the default closed category explored in the context of emotion is that of prepositions, the focus here is on determiners and negation marking, in a return to the phenomena associated with polarity items, items that require negative or positive contexts to be licit. The contrasts illustrated in (3)-(5) point out the sensitivity of some linguistic items (italicized) to being located in negative polarity contexts, while the contrasts of (6)-(8) show that positive polarity items also exist. Ladusaw (1980) generalized over constructions like these involving affective determiners which license polarity sensitive items as providing downwards entailment.[3]

(3) a. No politician *gives a damn*
    b. *Many politicians *give a damn*

(4) a. Few politicians have *any* desire to see the Euro collapse.
    b. *Most politicians have *any* desire to see the Euro collapse.

(5) a. Hardly any politicians have visited the region *in the longest time*.
    b. *Nearly all politicians have visited the region *in the longest time*.

(6) a. *No politicians have *already* embraced the arguments.
    b. Many politicians have *already* embraced the arguments.

[1] A restricted quantification analysis of generics is fairly standard, see Krifka et al. (1995).

[2] See Vogel (2001).

[3] Downwards entailment is entailment from the truth of sentences involving predications to the truth of sentences involving subsets of those predications – the truth of (1) entails (2).

(1) Leslie lacks a pencil.

(2) Leslie lacks a sharpened pencil.

(7) a. *Few politicians vote *unfortunately* along party lines.
b. Most politicians vote *unfortunately* along party lines.

(8) a. *Hardly any politicians are *somewhat* cagey.
b. Nearly all politicians are *somewhat* cagey.

Properties of generalized quantifiers (Barwise and Cooper, 1981), are the subject of a very large literature, and there is no lack of attention to entailments that they support ((Kanazawa, 1994; Peters and Westerståhl, 2006)). This paper returns to the licensing of polarity sensitive times as through affective domains, with particular attention the the reference sets quantified over explicitly (or implicitly) and their complement sets (Moxey and Sanford, 1993). The work has relevance to the deep syntactic and semantic analysis required to attune sentiment analysis to valency shifts (Kennedy and Inkpen, 2006).

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

Brady, M. S. (2009). The irrationality of recalcitrant emotions. *Philosophical Studies*, 145(3):413–430.

Dadlez, E. M. (1996). Fiction, emotion and rationality. *British Journal of Aesthetics*, 36(3):290–304.

Kanazawa, M. (1994). Dynamic generalized quantifiers and monotonicity. In Makoto Kanazawa et al., editors, *Dynamics, Polarity and Quantification*, pages 213–49. Stanford: CSLI Publications.

Kennedy, A. and Inkpen, D. (2006). Sentiment classificatioon of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Krifka, M., Pelletier, F. J., Carlson, G. N., ter Meulen, A., Link, G., and Chierchia, G. (1995). Genericity: An introduction. In Gregory N. Carlson et al., editors, *The Generic Book*. University of Chicago Press.

Ladusaw, W. (1980). On the notion 'affective' in the analysis of negative polarity items. *Journal of Linguistic Research*, 1:1–16.

Moxey, L. M. and Sanford, A. J. (1993). *Communicating Quantities: A Psychological Perspective*. Lawrence Erlbaum.

Peters, S. and Westerståhl, D. (2006). *Quantifiers in Language and Logic*. UK: Clarendon Press.

Vogel, C. (2001). Dynamic semantics for metaphor. *Metaphor and Symbol*, 16(1 & 2):59–74.

Vogel, C. (2011). Genericity and metaphoricity both involve sense modulation. In Khurshid Ahmad, editor, *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology*, Text, Speech and Language Technology, pages 35–51. Springer: Dordrecht.

3 Ontology

Topics in Social Media for Disaster Management - A German Case Study on the Flood 2013

Antje Schlaf[1], Sabine Gründer-Fahrer[1], Patrick Jähnichen[2]
[1]Institute for Applied Informatics - Adjunct Institute of the University of Leipzig, NLP Group
Hainstrasse 11, 04109 Leipzig
[2]University of Leipzig, Department of Computer Science, NLP Group
Augustusplatz 10/11, 04109 Leipzig
E-mail: antje.schlaf@informatik.uni-leipzig.de, gruender@uni-leipzig.de, jaehnichen@informatik.uni-leipzig.de

Abstract

The paper presents the results of a German case study on social media use during the flood 2013 in Central Europe. During this event, thousands of volunteers organized themselves via social media without being motivated or guided by professional disaster management. The aim of our research was to show and analyze the real potential of social media for disaster management and to enable the public organizations to get into touch with the people and take advantage as well as control of the power of social media. In our investigation, we applied state-of-the-art technology from *Natural Language Processing*, mainly topic modeling, to test and demonstrate its usefulness for computer-based media analysis in modern disaster management. At the same time, the analysis was a comparative study of social media content in context of a disaster. We found that Twitter played its most prominent part in the exchange of current factual information on the state of the event, while Facebook prevalently was used for emotional support and organization of volunteers help. Accordingly, social media are powerful not only with respect to their volume, velocity and variety but also come with their own content, language and ways of structuring information.

Keywords: social media analysis, crisis informatics, topic modeling, flood 2013 Central Europe, comparative corpus linguistics

1. Introduction

When in June 2013 the rivers Elbe and Donau bursted their banks and floodwaters were threatening the cities and villages in east and south Germany and Austria, a huge wave of readiness to help throughout the country was also on its way. During these days, not only the water levels and amount of rain set new records at many places, but also the engagement of volunteers was the highest ever known. Notably, these volunteers organized themselves mainly via social media.

There had been no concrete call to action and no guiding from the side of official public emergency management. But the new social media channels had created an opportunity to channel social energy and put into real action a valuable potential that might have been lost otherwise. And the viral effects of the open network made the movement rapidly spread and grow.

For instance, the public Facebook site 'Fluthilfe Dresden' founded by a young man, Daniel Neumann, on its own initiative, got about 12.500 followers after half a day only and reached 2.4 million people in one week. Before he knew how, the young man and two of his friends found themselves in a position to receive and answer about 60-80 messages per minute and coordinate as much as 50.000 people in overall (EIJK, 2015).

The high potential of social media platforms for emergency management is due to a number of properties. For instance, everybody can give and gain information directly and without barriers or information hierarchies to cross. The access is simple, stable and possible from everywhere, reach and coverage very wide, dissemination extremely rapid. The social aspects are inherent in the medium itself in form of interactivity, network building and multiplication of effects.

But in face of examples as 'Fluthilfe Dresden' it becomes obvious, that the powerful movement of privat engagement via social media urgently needs to be coordinated and guided by professional emergency managers (i.e., EIJK, 2015). Beside its great positive potential, self-dependent organization of volunteers can also bring dangers, as examples of misplaced and faulty actions during the flood 2013 showed. Volunteers and emergency managers both agree (ibid.) that state agencies should keep or regain information sovereignty and decision-making authority and that they have to assume ultimate responsibility in case of a disaster. This is only to be achieved by working hand in hand with volunteer's organizations and by *'going where the people are'* (Rutrell, 2010). That means, disaster management has to be present for gathering and giving information and for interactive communication on social media.

Currently, the idea to become actively involved into exchange of information and interaction on social media places a number of great challenges for the responsible bodies in German disaster management, though (i.e., Kirchbach Kommission, 2013; BMI, 2014). Two of them have been in the focus of the study presented in this paper.

- *Information flooding:* The digital room of social media is containing an overwhelming amount of information of great diversity and dynamicity.
- *New Modes of communication:* 24 hours of digital communication flow in an interactive many-to-many setting brings up new forms of communication together with new types of content.

The two problems mentioned are obviously interrelated, as the situation of information flooding in the social media space so far has been a great obstacle to all systematic investigations into its content and communicative characteristics.

With respect to the first problem, modern computer-based methods from the fields of text mining and information retrieval can automatically reveal content in very large collections of text and can enable disaster managers to efficiently search, sort and analyze relevant information. In our case study, we applied an up-to-date method for content analysis and clustering in order to test and demonstrate its usefulness in the given context. At the same time, with respect to the second of the above problems, our investigation was aimed at answering the following research questions:

1. What kind of communicative content is currently distributed by social media in context of a disaster?
2. Are their differences in content and language among several types of social media?

Answering the first question will help to better assess the real potential of social media for disaster management. The results of our investigation into the second question will support disaster managers to find out the most promising place where to search for and where to place certain kinds of information and the most appropriate way to communicate with the people in each context.

2. Data and Methods

2.1 Data

In our case study on the flood 2013, we investigated messages from two big social media platforms - Facebook and Twitter. Covering the time span of the core event from May to July 2013, we collected German data from both platforms via their public API, respectively.

For the Facebook flood corpus we retrieved data from public pages or groups containing the words *Hochwasser (flood)* or *Fluthilfe (flood aid)* in their names. Our sample trainings corpus consisted of 35.6k messages (1.2M word tokens) from 264 public pages or groups. Preprocessing of the corpus included deletion of punctuation marks and stop words, tokenization and lemmatization; numbers were mapped on a generic reference (*num_ref*).

For the Twitter flood corpus we retrieved a current version of the research corpus of the project *QuOIMA* (QuOIMA, 2013) on basis of tweet IDs. The *QuOIMA* corpus had been collected from the public Twitter stream and filtered by 65 hash tags coming from research of the Austrian Bundesheer and by 29 names of manually chosen public accounts connected to disaster management and flood aid. The current version of the corpus comprises 354k tweets (4M word tokens). The preprocessing was the same as in case of the Facebook corpus but additionally included removal of Twitter names, URLs and Hashes from Hash tags as well as deletion of retweets.

2.2 Methods

The technique we applied to reveal the hidden thematic structure in our corpora was topic modeling. Topic models (e.g., Blei, Ng, & Jordan, 2003; Blei, 2012; Griffiths & Steyvers, 2002) are a family of statistical models based upon the idea that documents are mixtures of topics. Each topic is defined in form of a probability distribution over words. These weighted topic words, that pick out a coherent cluster of correlated terms, allow for an intuitive interpretation of the topics. Topic model techniques are a very useful new way to search, browse, summarize and cluster large collections of text.

The specific type of topic model we applied in our investigation of social media content was a *Hierarchical Dirichlet Process* in form of a *Chinese Restaurant Franchise Sampler* (HDP CRF) (Teh and Jordan, 2010). Unlike more standard algorithms like LDA (*Latent Dirichlet Allocation*), HDP CRF is a nonparametric algorithm which automatically uncovers the number of topics based on the data characteristics. The benefit of this is that we have greater flexibility in adapting to the peculiarities of social media data and are able to formally define Bayesian priors, even if we do not know, how the appropriate prior probability distribution should look like.

We took an additional step to improve interpretability of topics for the following reason. Common terms in the corpus often appear near the top of the ranked word list for multiple topics, making it hard to differentiate the meanings of these topics on basis of the most probable terms according to the topic model. In order to meet this problem, we used *relevance* (Sievert, Kenneth, 2014) as our method of ranking topic terms. *Relevance* is defined as the weighted average of a term's probability within a topic and its *lift* (Taddy, 2012). The *lift* generally decreases the ranking of globally frequent terms, being the ratio between a term's probability within a topic and its marginal probability across the corpus. The following definition of *relevance* by Sievert and Kenneth (2014) uses a parameter λ (where $0 \leqq \lambda \leqq 1$) which determines the weight given to the probability of a term w under topic k relative to its *lift* (measuring both on the *log* scale).

Relevance: $r(w, k \mid \lambda) = \lambda \log(\phi_{kw}) + (1-\lambda) \log(\phi_{kw} / p_w)$,

where ϕ_{kw} denotes the probability of the term w for topic k and p_w the marginal probability of w across the corpus.

In order to investigate inter-topic differences within a corpus, we computed the distances between topics (*Jensen Shannon divergence*) and then applied *Multi-dimensional Scaling* (*Principal Components*) to project the distances onto two dimensions (compare Chuang et al., 2012). As our visualization system, we used *LDAvis* (Sievert, Kenneth, 2014).

In order to complement the results of our main analysis using topic modeling we conducted two supplementary comparative studies.

First, we did mutual comparison of equal sized samples of the flood corpora from the two different social media, aiming at differences in content and language between them. We analyzed differences in relative frequency of

keywords from one corpus to the other using *Log-Likelihood Ratio Test* (Dunning, 1993).
Finally, we investigated emotional involvement in flood-related messages in Facebook and Twitter. We compared relative frequencies of sentiment words for each case using the *SentiWS* resource from the University of Leipzig NLP group (Remus et al., 2010), a list of German positive and negative sentiment bearing words.

3. Results

The results of our investigation to be discussed in the remainder of the paper, are presented in the following figures and tables.
Table 1 and Table 2 summarize the results of the topic model analysis for the Facebook and the Twitter flood corpora. We have proposed a title to each topic on basis of the words that received high probability under the topic. For each topic, the 8 most *relevant* topic terms, according to the above definition, are displayed in decreasing order (together with their English translation in brackets). We chose $\lambda=0.4$ for Facebook and $\lambda=0.3$ for Twitter to balance term's probability within the topic and its *lift*. For each corpus, the topics are ordered by their *topic frequency*, calculated as the sum of topic probabilities over all documents, normalized by the length of the respective document.
The relative size of all topics according to this measure is displayed in a topic wheel in Figure 1 and Figure 2.
Another global view on the topics in each flood corpus is revealed in Figure 3 and Figure 4 on the lefthand side in form of the results of multi-dimensional scaling. On the righthand side of these figures there are the most relevant terms for an example topic, namely topic 1 in each case, together with the corresponding term frequencies within this topic and overall term frequencies within the corpus.
Table 3 lists the 12 most significant terms of the mutual differential analysis in decreasing order. The terms in the first row characterize the Facebook content when compared to Twitter; the terms in the second row, are the differentiating lexical features of Twitter when compared to Facebook.
Finally, results of sentiment analysis are displayed in Figure 5, showing the relative frequency of sentences containing positive and negative sentiment markers (as a percentage), respectively.

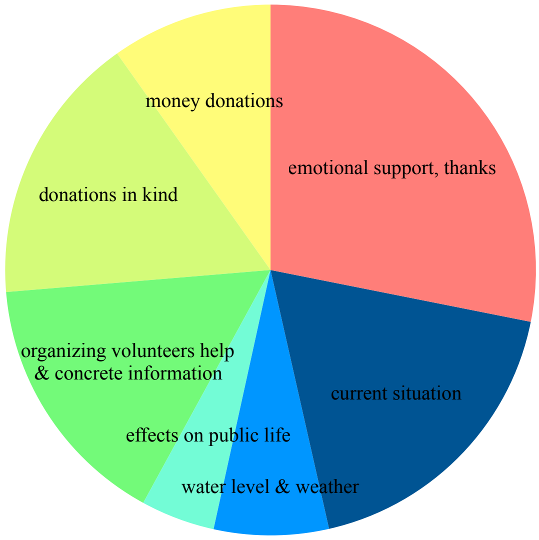

Figure 1: Facebook Topic Wheel Flood 2013

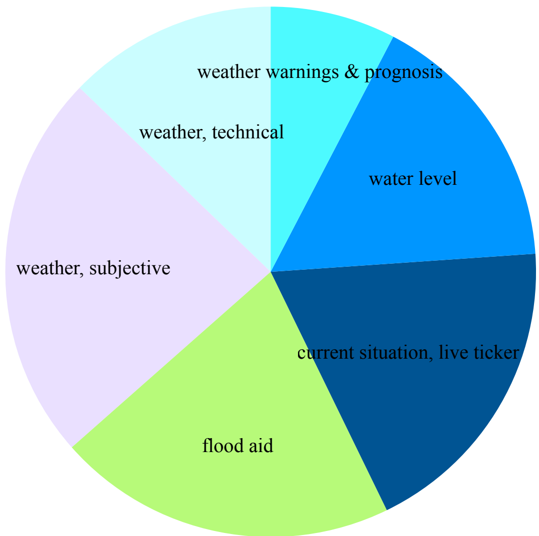

Figure 2: Twitter Topic Wheel Flood 2013

1	emotional support and thanks	*mal* (*times* or modal particle spoken language), *danke (thank you), vielen (many), Dank (thank), Menschen (people), Seite (site), schon (already* or modal particle spoken language), *viele (many)*
2	current state	*Wasser (water), Quelle (source),* num_ref, *Elbe* (river *Elbe), Deiche (dikes), Evakuierung (evacuation), Deich (dike), Landkreis* (administrative district),
3	donations in kind	*Sachen (things), Kleidung (clothes), melden (report, volunteer), Sachspenden (donations in kind), spenden (donate), bitte (please), abzugeben (to be given away)*
4	organizing volunteers help and concrete information	*Helfer (volunteers), jemand (somebody), Sandsäcke (sand bags), benötigt (needed), weiß (know), Hilfe (help), gebraucht (needed), tomorrow (morgen)*
5	money donations	*Euro,* num_ref, *Hochwasser (flood), Konto (bank account), BLZ (BIC), Soforthilfe (emergency aid), Betroffene (people affected), Deutschland (Germany)*
6	water level & weather	num_ref, *Uhr (o'clock), Pegelstand (water level), aktueller (current), html, Tendenz (tendency), steigend (rising), Pegel (water level)*
7	effects on public life	*Grundschule (primary school), Unterricht (classes), Straße (road), gesperrt (closed), Schulen (schools), Zwickau* (city in Saxony), *Mülsen* (town in Saxony), *ZOB (Central Bus Station)*

Table 1: Facebook Topics Flood 2013

1	weather, subjective	*Regen (rain), mal* (*times* or modal particle spoken language), *schon (already* or modal particle spoken language), *schön (nice), gut (good), Sommer (summer), endlich (finally)*
2	flood aid	*Hochwasser (flood), Hilfe (aid), Hochwasserhilfe (flood aid), Euro, helfen (help), Merkel* (German chancellor), *Helfer (volunteers)*
3	current situation, live ticker	*Hochwasser (flood), Lage (current situation), Magdeburg* (city in Saxony), *Hochwasser-Ticker (flood ticker), Webcam, Hochwasserlage (flood situation), Live-Ticker*
4	water level	num_ref, *Stand (state, level), Pegel (water level), gefallen (fallen), Pegelmw (mean water level), gestiegen (risen), Minuten (minutes)*
5	weather, technical	num_ref, *kmh* (metric measure of speed), *hPa* (metric measure of pressure), *temp* (abbreviation for temperature), *Wind (wind), Luftdruck (air pressure), /m² (per square meter)*
6	weather warning	*Starkregen (heavy rain), Unwetter (severe weather), Gewitter (thunderstorm), Wetterwarnung (weather warning), Unwetterwarnung (thunderstorm warning), schwere (heavy), Vorhersage (forecast)*

Table 2: Twitter Topics Flood 2013

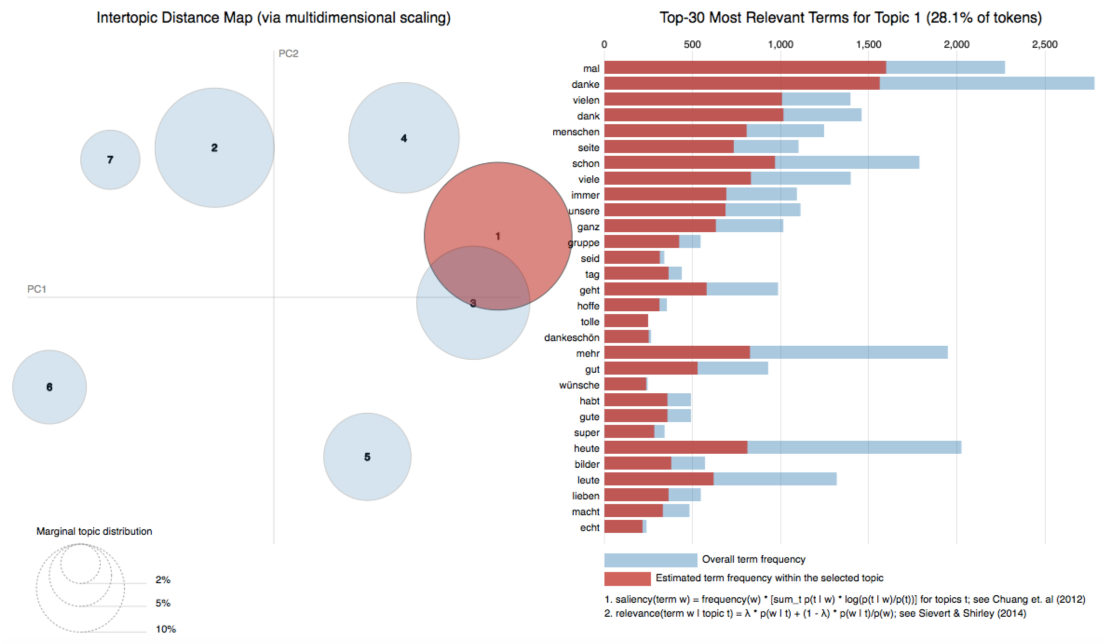

Figure 3: Facebook Topic Global View (λ = 0.4)

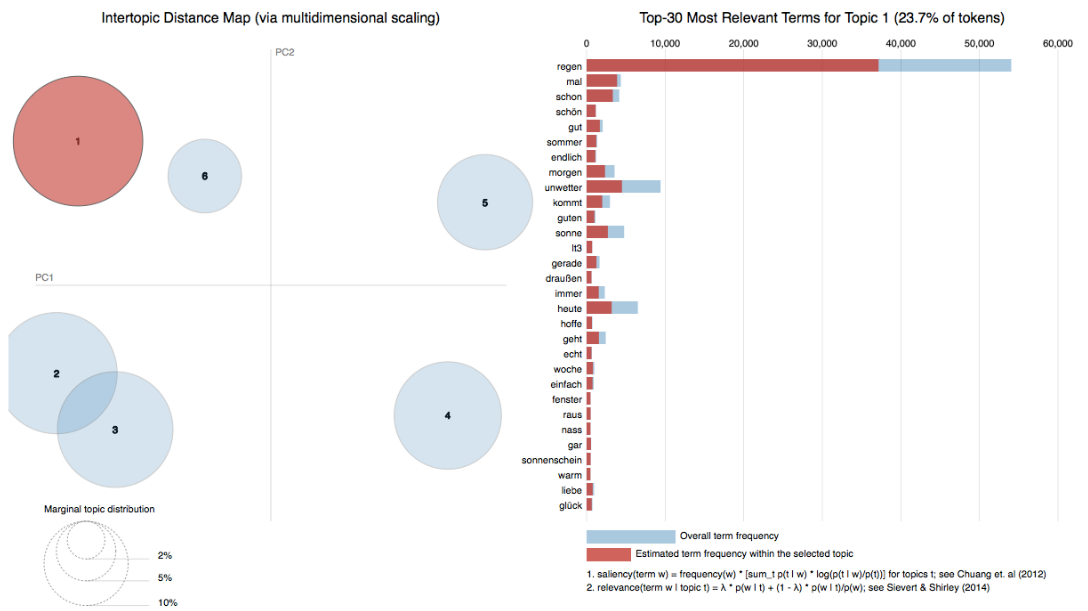

Figure 4: Twitter Topic Global View (λ = 0.3)

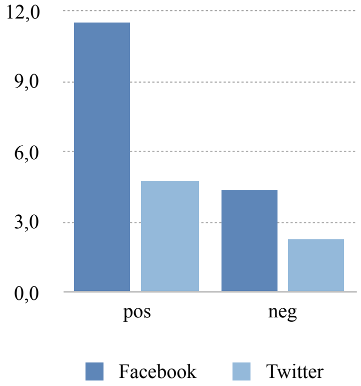

Figure 5: Comparative Sentiment Analysis

Facebook compared to Twitter	*helfen (help)* *bitten (please)* *melden (report, volunteer)* *bitte (please)* *gerne (gladly)* *benötigen (need)* *Sachspende (donation in kind)* *abgeben (give away)* *sache (thing)* *Gruppe (group)* *gebrauchen (use)* *Hilfe (help)*
Twitter compared to Facebook	*Hochwasser (flood)* *Stand (state, level)* *Pegel (water level)* location_city (location marker) *kmh* (metric measure of speed) *hpa* (metric measure of pressure) *Pegelmv (mean water level)* *Elbe* (river *Elbe)* *Wind (wind)* *Unwetter (severe weather)* *Regen (rain)* *temp* (abbreviation for *temperature*)

Table 3: Mutual Differential Analysis[1]

3.1 Facebook

The focus of the content in the Facebook flood corpus is on social interaction and help. In Figure 1, the socially relevant topics are market in red, yellow and green colors and take about three-quarters of the overall topic space of the corpus.

The most prominent topic is concerned with giving emotional support to the affected people and thank the helpers. Furthermore, there is a large topic on offering, asking for and coordinating volunteer's help and concrete information and two topics on organizing donations in kind and money donations, respectively. In Figure 1, the topics that include current information on the factual level of the flood event, i.e., general situation, water levels, weather conditions, are marked in blue colors and only take about one-quarter of the overall topic space. Even on the factual level, the social perspective occurs in topic 7 that is concerned with current effects of the event on public and social life.

In line with this consideration, the main division of lexical semantic space according to *Multi-dimensional Scaling* in Figure 3 appears between social topics on the right and factual topics on the left. For the vertical dimension one might suggest an interpretation in terms of formal content of the topics, with topics 5 and 6 ('money donations' and 'water level & weather') being he most formal ones and set apart from the other, more informal topics.

As the most dominant content of Facebook flood-related messages is related to empathy and social interaction, the prevalent language is emotionally involved and informal in style, often showing elements of spoken language (i.e., certain modal particles, as shown in Table 1, topic 1).

High empathy and emotional involvement is also proven by the results of *Sentiment Analysis* in Figure 5. The fact that especially high values appear in case of positive sentiment markers is reflecting the efforts of the public to provide emotional support, encouragement and motivation. In comparison, Twitter does not even come close to achieving similar high rates in positive sentiment.

The differentiating terms of Facebook compared to Twitter (in Table 3) all hint to social interaction and help.

In overall, the Facebook communication seems to be taking the perspective of the affected people.

3.2 Twitter

When moving from the Facebook to the Twitter flood corpus, the semantic focus just switches. As can be read off the topic wheel in Figure 2, the main focus in case of Twitter is on current information on the event.

The topics related to factual information with respect to the general situation, water level, weather, weather warnings and prognosis are marked in blue colors, and they are taking about 80 percent of the overall topic space. Social engagement and organization of help is also

1 For differential analysis, all location markers where mapped on the string 'location_city'.

present in Twitter but appears as subordinated.
Likewise, emotional involvement is relatively rare here (Figure 5), and it mostly appears in the prominent topic 1 were people show their sentiments with respect to the weather rather than in the context of empathy and emotional support as it was in Facebook.
The division of lexical semantic space according to *Multi-dimensional Scaling* seems to be between weather topics in the upper and topics related to flood and water level on the bottom. The vertical dimension may be interpreted as dividing between more formal language style on the right, and less formal and colloquial style on the left.
The prevalent language of Twitter is situative reporting and factual. It shows examples of colloquial style but also includes topics that show very technical language.
These very technical lexical features and the focus on the event itself also come out as the characteristic aspects of Twitter when compared to Facebook via differential analysis (Table 3). Surprisingly, Twitter hereby place itself in close proximity to professional reports rather than Facebook, as we have shown in Gründer-Fahrer, Schlaf (2015) on basis of *Latent Semantic Analysis*.
Generally, the communication in Twitter is mostly taking a factual perspective on the event.

4. Discussion & Conclusion

Our investigation we applied state-of-the-art techniques from *Natural Language Processing*, mainly topic modeling, in order to pursue a threefold aim.
First, we wanted to test and prove its usefulness in context of social media analysis for disaster management. The method generally revealed very meaningful and coherent clusters of content in the data examined. On basis of the inferred topic models, it is possible to filter, sort and semantically analyze huge amounts of social media data automatically and to open up efficient, computer-based access to this new source of information to professional disaster managers. For instance, the model can give an overview over the content of existing text collections and can point out the main topics under consideration there. Furthermore, new incoming text messages can be provided with topic labels and topic probabilities and can then be thematically ordered. This way of structuring and ordering the information makes it more comprehensible and available, and ready to be used according to different interests the managers may have when looking at the data. Additionally, the topic words can be used as search or filter terms for retrieval of data in each register. Finally, when tracking topics over time, one can get aware of overall developments and trends and of critical states in the factual or social dimension of the emergency situation. An implementation of the topic model tested here is supposed to become part of the text analytical module of the disaster management software prototype that will be the outcome of the European project 'Slándáil'. This software prototype will be available for further testing and possible application to end-users then.
Second, our investigation at the same time was an analysis of social media content and its potential for disaster management. From the point of view of emergency management, a crisis has three basic dimensions: 1. the real event; 2. the actions of the involved organizations; 3. the perception of the crisis. Taking the flood 2013 in Germany and Austria as our case study, we revealed high potential of social media content for disaster management in all three dimensions. With respect to the real event, social media were intensively used for sharing of up-to-date information and for spreading warnings. In this way social media contribute to improvement of situational awareness and timeliness of early warnings. In the dimension of the activities related to the event and resilience, social media played an important part in the organization of volunteers activities and donations. Here the general potential was in improving social connectedness and concrete support. As for the perception of the crisis and its emotional processing, social media provided a good and frequently used possibility to directly show empathy, provide emotional support or practice emotional (self-)management. Like this it was a valuable resource for psychological self-help.
Finally, the third focal point of our investigation was a comparative analysis of social media content and language in context of a disaster. In our analysis of German language data from the flood 2013, there appeared striking differences in content between the two big social media platforms under investigation, Facebook and Twitter. According to our observation, the focus of Facebook content is on empathy and social engagement. The conceptualization generally takes the perspective of the affected people and the prevalent language is emotionally involved and informal. Twitter, in contrast, is mainly used for exchange of current and concrete information on the event. It takes a more factual point of view, and the characteristic language is situative reporting and factual, stylistically ranging from quite technical to colloquial. This means, among the dimensions of a crisis mentioned in the previous paragraph, Facebook dominates the dimension that includes activities related to the event and to resilience as well as the dimension concerned with the perception of the crisis and its emotional processing. To the dimension of the real event, Twitter contributes more and more precise information.
In result of this study, it can be stated that the arrangement of information in social media in context of a disaster, rather than being chaotic or arbitrary, reveals a clear order. And, interestingly, the different platforms, Facebook and Twitter, rather than being just coexisting or competing, perform a systematic and effective division of labor. These results can serve as an orientation point for professional disaster managers who seek to more successfully search for relevant information in social media. At the same time, the outcome of our study can support them to place their own information in the right context and in an appropriate style as to gain people attention and trust. Moreover, our results should be of interest to communication and linguistic studies as well as

media research. It will certainly be interesting to extend this kind of analysis to different languages and social communities as to investigate commonalities and differences in the organization of public, collective and individual help, social behavior, sociolinguistics and media use.

5. Acknowledgements

The research leading to these results has received funding from the European community's Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL)."

6. Bibliographical References

Blei, D.M. (2012). Probabilistic topic models: Surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the ACM* 55(4), pp. 77–84.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, (3), pp. 993–1022.

BMI - Bundesministerium des Inneren (2008). Krisenkommunikation - Leitfaden für Behörden und Unternehmen. Online: http://www.bmi.bund.de/SharedDocs/Downloads/DE/Broschueren/2014/leitfaden-krisenkommunikation.pdf?__blob=publicationFile, 09.02.2016.

Chuang, J., Ramage, D., Manning, Ch. D., Heer J. (2012). Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. In: *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI'12).

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* , 19(1): 61-74.

EIJK - European Institute for Journalism and Communication Research (2015). Documentation of the conference on communication floods (Kommunikationsfluten), Leipzig. Online: http://konferenz.eijc.eu/dokumentation.html, 8.02.2016.

Griffiths, T.L., Steyvers, M. (2002). A probabilistic approach to semantic representation. In: *Proceedings of the 24th Annual Conference of the Cognitive Science Society,* pp. 381—386.

Gründer-Fahrer, S., Schlaf, A. (2015). Modes of Communication in Social Media for Emergency Management. In: *Proceedings of the 2nd Workshop on Computer Mediated Communication / Social Media at GSCL 2015*, pp.33—37.

Kirchbach Kommission (2013). Bericht der Kommission der Sächsischen Staatsregierung zur Untersuchung der Flutkatastrophe 2013 (Kirchbachbericht). Sächsische Staatskanzlei, E. Nowak, L. Rischke, H. Zorn (Eds.). Online: www.publikationen.sachsen.de/bdb/artikel/20534, 25.01.2016.

QuOIMA (2011). QuOIMA - Open Source Integrated Multimedia Analysis. Online: www.kiras.at/gefoerderte-projekte/detail/projekt/quoima-quelloffene-integrierte-multimedia-analyse, 25.01.2016.

Remus R., Quasthoff U., Heyer G. (2010). SentiWS – a Publicly Available German language Resource for Sentiment Analysis. In: *Proceedings of the 7th International Language Resources and Evaluation* (LREC), pp. 1168—1171.

Rutrell, Y. (2010). 5 ways social media is changing emergency management. Online: www.gcn.com/articles/2010/09/06/ social-mediaemergency-management.aspx, 09.02.2015.

Sievert, C., Kenneth, E. S. (2014). LDAvis - a method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70.

Taddy, M. A. (2012). On Estimation and Selection for Topic Models . In: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012), pp. 1184–1193.

Teh Y. W., Jordan M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In: N. L. Hjort et al. (Eds.), *Bayesian Nonparametrics*. Cambridge, UK: Cambridge University Press, pp. 114–133.

4 Terminology

A Linguistically-driven Methodology for Detecting Impending Disasters and Unfolding Emergencies from Social Media Messages

Maria Teresa Musacchio*, Raffaella Panizzon*, Xiubo Zhang, Virginia Zorzi***

*Department of Linguistic and Literature Studies, University of Padua, Italy
E-mail: mt.musacchio@unipd.it, raffaella.panizzon@unipd.it, virginia.zorzi@studenti.unipd.it

**School of Computer Science and Statistics, Trinity College Dublin, Ireland,
E-mail: xizhang@scss.tcd.ie

Abstract

Natural disasters have demonstrated the crucial role of social media before, during and after emergencies (Haddow & Haddow 2013). Within our EU project Slándáil, we aim to ethically improve the use of social media in enhancing the response of disaster-related agencies. To this end, we have collected corpora of social and formal media to study newsroom communication of emergency management organisations in English and Italian. Currently, emergency management agencies in English-speaking countries use social media in different measure and different degrees, whereas Italian National Protezione Civile only uses Twitter at the moment.
Our method is developed with a view to identifying communicative strategies and detecting sentiment in order to distinguish warnings from actual disasters and major from minor disasters. Our linguistic analysis uses humans to classify alert/warning messages or emergency response and mitigation ones based on the terminology used and the sentiment expressed. Results of linguistic analysis are then used to train an application by tagging messages and detecting disaster- and/or emergency-related terminology and emotive language to simulate human rating and forward information to an emergency management system.

Keywords: social media, linguistic analysis, system training

1. Introduction

During natural disasters, communication plays a central role in successfully managing mitigation, response and recovery operations and in limiting damage to people and property. Traditional media – also known as legacy media, i.e. websites of conventional newspapers, news agencies and broadcasting corporations – have long been the main source of information for the population before, during and after emergencies, and have typically allowed for monodirectional, top-down and mostly asynchronous communication. The tremendous growth and spread of social media websites, including social networking systems like Facebook and microblogging systems like Twitter, has led to forms of communication that are bi- or multidirectional, dialogical, and synchronous. People now have the chance to report and spread information about events within seconds to a very broad network of interconnected users, thus acting like 'social sensors'. In other words, individuals and groups are capable of generating relevant and relatively reliable knowledge on a given issue or situation. In the specific case of emergencies, people can gather timely and updated information thus becoming a sort of unofficial early warning system (Avvenuti et al., 2014).
The use of social media for disaster management purposes started in recent times with superstorm Sandy (October 2012) when US FEMA first resorted to social media to spread real-time, validated information and organise and direct aids (Haddow and Haddow, 2013). It soon became clear that social media provided useful means for sharing information between agencies and the public at large at all times. The public can now receive information and announcements from agencies and respond to such information as well as provide and circulate crowd-sourced information that can play a vital role in mitigating the impact of, and recovery from, a disaster event. The role of social media in emergency management, particularly during natural disasters, has been studied extensively in recent years. For example, an analysis of social media during severe weather events shows that "social media data can be used to advance our understanding of the relationship between risk communication, attention, and public reactions to severe weather" (Ripberger et al., 2015). One of the challenges of using social media information is how to handle the large volumes and variable quality of messages published by non-authoritative sources.

2. Methodology

Methods to filter, organise, and analyse the wealth of information available play a crucial role now that natural disasters seem to occur more and more often. The key objective of the Slándáil system is to "integrate an emergency management system with a social media system that is capable of processing text (and image) data while taking into consideration the ethical and factual provenance of data, thus removing the burden of manual search and interpretation of latent information contained in social media data"[1]. An essential part of the project, and the topic of this paper, is the analysis of the communication techniques of agencies through social media in order to establish a linguistically-driven methodology for the analysis of messages with a view to extracting content features and instances of sentiment.

[1] www.slandail.eu

The methodology we propose in this paper consists first in looking at measures that can provide insights into the communicative features of the texts described above by looking at their complexity. This can be accounted for by measuring the readability of texts – or what level of education is required for readers to process a text –; type/token ratio (TTR) and standardised type/token ratio (STTR), which can shed light on the level of lexical variation present in the text; lexical density, which provides an indication of the content words present in a text; multidimensional analysis (Biber, 1988), which allows for the observation of features such as how informative or persuasive a text is, whether contents are more or less abstract and the level of re-elaboration of a text; and sentiment analysis, through which the attitude of writers towards events can be detected.
Our results have been used to train one of the software applications specifically designed for the project, CiCui, so that it can detect potential emergencies. The CiCui system is a text analysis system that transforms raw text into structured form. Initially conceived as a tool for exploring and extracting word collocations in text, it was later expanded into a more general text analysis platform. Its core function is to build a positional inverted index for raw input documents which come with no syntactic or semantic annotations (Zhang and Ahmad, 2014). Syntactic annotations typically identify the grammatical role that a word or other lexical units play in a sentence, e.g. nouns, verbs, phrases, etc.; semantic annotations hint to the semantic categories a lexical unit belongs to, e.g. organisation names, locations, date-time references, etc. Subsequent analyses can then use this index to extract additional information such as the lexicon of the corpus, prominent terminologies and word collocations, taxonomy and ontologies, quantified content, latent topics, and more.

3. Dataset

The methodology proposed is tested on a number of sample corpora in English and Italian. From the corpora we have collected for investigation within the Slándáil project we have selected components that are comparable because of their linguistic features. For most of our analysis, English and Italian datasets consist of fact sheets – short information manuals written by agencies for the general population –, Facebook posts and comments, and tweets. For sentiment analysis we used a larger corpus to retrieve as many collocations as possible. Because the Italian *Protezione Civile* (Civil Protection) does not have an official Facebook account, posts and comments were extracted from semi-institutional accounts set up by regions, provinces and municipalities. In the case of Twitter, also 'authoritative' accounts, i.e. those giving reliable information but not associated with any agency, were included (Table 1). The purpose here was to somewhat balance the sizes of the English and Italian corpora available, since local branches of Civil Protection have started using social media systematically only in very recent times (hence have produced fewer data) and are currently using only Twitter. Corpus composition and size are outlined in Table 1 below. Facebook posts from institutional and semi-institutional sources were collected with respective comments in order to account for input coming from non-institutional users as well and to look at communication produced by common people. Considering our focus on institutional communication, only Facebook posts were then analysed.

	Doc. type	**Tokens**	**Total**
Fact Sheets En	institutional	40,667	40,667
Fb Posts En	institutional	96,153	96,153
Twitter En	institutional	60,500	60,500
Fact Sheets Ita	institutional	65,755	65,755
Fb Posts Ita	institutional	124,280	128,248
	semi-institutional	3,968	
Twitter Ita	institutional	18,087	147,104
	semi-institutional	129,017	

Table 1: Detail of the corpus.

Other corpora were included to provide a wider picture of sentiment in disaster communications: a news corpus extracted from the portal Lexis Nexis consisting of 466,945 tokens for English (keywords: *weather, emergency, disaster*) and one of 176,597 tokens for Italian (keywords: *maltempo, emergenza, disastro*) (see section 4.4). Finally, we also collected and filtered 150.000 tweets with geolocation UK and Ireland from the Twitter API for the training of the CiCui system (see section 5).

4. Linguistic Analysis and Results

The analysis of portions of these corpora aims to shed light on the communicative strategies adopted by agencies when talking to the population before, during and after disasters across different media with inherently different constraints. Direct observation of the texts has shown that information is scaled depending on the medium used. Twitter provides minimum information because of its character limit but often includes links to a website where more in depth knowledge can be gained. Facebook does not impose such restriction, however users seem to be less prone to reading one lengthy message than a sequence of short messages. Therefore, even if posts tend to be longer than tweets, they are often accompanied by complementary or explanatory images or links to external sources just as in Twitter. In this way users can see a synthesised version of the message and decide whether they want to read more about it or not. This is also why FEMA often uses infographics in peace times: users can immediately see a summarised and simplified version of the message and can remember it more easily. Finally, fact sheets are also publicly available online but are in the form of a downloadable pdf file, which is often no longer than two or three pages. They provide factual information on disasters and what to do when faced with one. The medium chosen appears to have a direct impact on the language used to communicate with the public in terms of lexical choices and syntactic structure. In communication through social media before, during and after disasters, emergency agencies need to engage with the population as

partners. This implies a shared language so emergency operators need to adapt their language to the requirements for information and knowledge of disasters of the population at large. In linguistic terms, this means considering how difficult texts are to cognitively process when people are under pressure as they typically are in an emergency.
When looking at social media communication the stances of emergency agencies towards the topics they present and the people they communicate with need to be considered with a view to establishing what feelings are expressed as emergency operators may strive for objectivity in an attempt not to spread panic while the population may reflect in their messages their needs and concerns. Sentiment describes the writers/speakers' attitudes to the topics they deal with in their texts, but also how readers/listeners provide their own assessments of the situations and opinions on disaster response work.

4.1 Readability Analysis

The readability level of English texts was analysed through the Flesch Reading Ease (FRE) and the corresponding educational attainment by means of the Flesh-Kincaid Grade Level (FKGL) indices, whereas that of Italian texts was measured through the Gulpease index. All of them rely exclusively on textual factors such as the number and length of words and sentences present in the text. These are density-like measures, thus being independent of text length (Gervasi and Ambriola in Castello 2008). Both FRE and Gulpease indices scores range over a 100 point scale, where high values relate to ease of processing. In the former, the minimum score for the language used in the text to qualify as 'Plain English' is 60, which roughly corresponds to 50 in the latter. Additionally, the FKGL and corresponding grades for Italian were used to integrate the previous measures with corresponding levels of education by indicating how many years of school a person needs to have in order to process the text easily. The FKGL measures readability by comparing it to the US grade level or the number of years of education required to understand the text (e.g. 10.9 would mean ten years, nine months).

TEXT TYPES	READABILITY INDICES	
	English	
	Flesch Reading Ease*	Flesch-Kincaid Grade Level**
Fact sheets	44.3	10.9
Facebook	48.6	10.9
Twitter	55.7	7.7
	Italian	
	Gulpease	Grade level
Fact sheets	50.9	12
Facebook	56.1	10
Twitter	58.8	9

Table 2: Readability indices for a sample of fact sheets, Facebook posts and tweets in English and Italian. The indices score over a 100-point scale with high values relating to ease of understanding and corresponding level of educational attainment.

The readability analysis above shows that neither fact sheets nor social media messages are written in Plain English; their language is best understood by high school graduates, with Twitter being also within the reach of 10- to 12-graders. Conversely, the corpora of Italian texts present a narrower variation in readability as to Facebook and Twitter messages. Fact sheets in Italian prove to be more difficult than Facebook or Twitter messages even though they proved to contain a considerably higher number of words belonging to the core vocabulary. This may indicate that the inherent difficulty in fact sheets does not lie in their lexis (or only partly) but in the greater length and complexity of their sentences. Also, almost 60 per cent of the tokens in Twitter messages are low-frequency words, which may point to the fact that people tend to report events in a more succinct manner, thus using more context specific language. However, all these methods take into account only easily measurable aspects of texts, which may not be accurate measures of syntactic complexity (e.g. sentence length), or word difficulty (e.g. syllable count). Hence these values should be taken only as rough predictions of textual features (Castello, 2008). This is why the data on readability have been integrated with further measures accounting for text complexity.

4.2 TTR, STTR and Lexical Density

Texts were tagged, which helped to identify lexical items (i.e. content words). However, a second round of manual checks was required in order to exclude all non-relevant instances that were not captured by tagging (e.g. typical social media metalanguage such as 'com' or 'https').

	Fact sheets EN	FB posts EN	Twitter EN
Tokens	38,272	24,864	8,951
Types	4,270	3,064	1,561
TTR	17.34	19.99	23.11
STTR	42.23	42.03	35.24
STTR St. Dev.	56.61	53.34	56.08
Lexical Items	24,192	14,282	5,548
Lexical Density	63.21%	57.44%	61.98%

Table 3: Features of lexical complexity of English fact sheets, Facebook posts and tweets.

	Fact sheets IT	FB posts IT	Twitter IT
Tokens	14,113	16,036	12,291
Types	2,844	3,252	1,994
TTR	30.28	31.03	21.50
STTR	44.19	41.56	33.60
STTR St. Dev.	53.04	53.35	57.79
Lexical Items	8,264	9,375	8,896
Lexical Density	58.55%	58.46%	72.37%

Table 4: Features of lexical complexity of Italian fact sheets, Facebook posts and tweets.

Type/token ratio (TTR) is one of the measures that accounts for vocabulary and lexical diversity in a given corpus by indicating how often, on average, a new word form appears in the text. The higher the value, the greater the number of different lexical items used in the text. An issue with this measure is that all instances of the text – i.e. both grammatical and lexical words – are counted, and thus equally contribute to lexical diversity. By excluding grammatical words, we obtained a value closer to the actual density of the text. Since TTR decreases with text length and we examined long strings of text, the final value was calculated on samples of 1000 words each with the final measure resulting from their average (STTR). Results in Table 3 and 4 suggest that there are differences both between text types and languages. As for English, the highest STTR is found in fact sheets (44.19) and the lowest in tweets (33.60), which is also in line with the readability measures found in Table 2. Italian texts seem to follow a slightly different pattern in that fact sheets are comparable to Facebook but not to Twitter, whereas readability measures identified substantial differences between fact sheets and Facebook posts. The values extracted also allowed for the calculation of lexical density following Ure’s (1971) method, summarized in the formula:

$$\text{LD} = \frac{n.of\ lexical\ items}{total\ n.of\ tokens} \text{ x } 100$$

Written texts tend to have a density of over 40% (Castello, 2008), which seems to be in line with the results from our study. The lexical density of corpora in Italian ranges across a relatively large interval (58.46%-72.37%). A comparison of text types across the two languages (Tables 3 and 4) indicates that Italian tweets and Facebook posts are denser than the English ones, although fact sheets in English are much denser than Italian ones. Italian tweets are denser than any other text type across languages, which may indicate that Italian institutions tend to provide users with more ‘packed’ information, which can usually be ‘unpacked’ by following the URLS that appear in the tweet. Conversely, Facebook messages are the least dense in each language, which is due to their communicative features. The relatively low lexical density can also be attributed to the fact that posts are often complemented with images, infographics or redirect to a newspaper article or blog post. As for fact sheets, the values of English and Italian are not so close, with a 4.66% difference. The language used is generally more detailed and varied than the one found in social media but it also needs to be easy enough to be understood by people with an average level of education.

4.3 Multidimensional analysis

A multidimensional analysis based on Biber's dimensions was carried out in order to investigate six key features (or dimensions) of our texts and thus integrate the analyses presented above. Only English texts could be analysed due to the current lack of a set of parameters for corresponding dimensions in the Italian language. Results are summarised in Table 5 below.

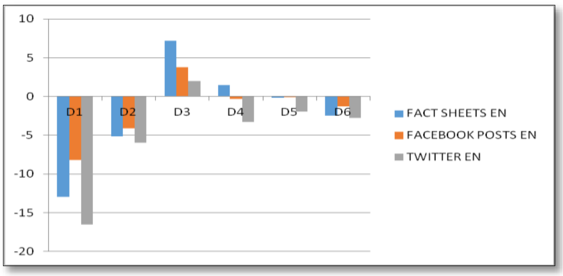

Table 5: Results of multidimensional analysis for English and Italian fact sheets, Facebook posts and tweets.

Dimension 1 (D1) refers to 'involved and informational discourse' where negative scores correlate with high informational density and a high number of nouns, long words and adjectives, while positive scores indicate that the text is affective and interactional and is characterised by a high number of verbs and pronouns. All three corpora examined present negative scores, thus being predominantly informational, with a cline going from Facebook posts, to fact sheets and finally to tweets, which appear the most informationally dense and least affective. In Dimension 2 (D2) – narrative and non-narrative concerns – low scores indicate that the text is non-narrative, while high scores show that the text presents many past tenses and third person pronouns, as in fiction. The three corpora present on average negative scores, suggesting that messages are related to the reporting of factual information and current events, with few references to the past. Dimension 3 (D3), or context-independent and context dependent discourse, accounts for the fact that the text is dependent on the context and presents many adverbs (low scores) or is context-independent and has many nominalisations (high scores). Fact sheets appear to provide information that does not require knowledge on a specific situational context, while Facebook posts and especially tweets tend to be closer to 0, so they may occasionally refer to more specific events, a prerogative of social media messages. Dimension 4 (D4), overt expression of persuasion, indicates whether texts express the author’s point of view as well as their assessment of likelihood and/or certainty on facts. High scores and modal verbs are an indication of such features. Fact sheets appear mildly persuasive, which may be justified by their inherent intent to guide people's behaviour during emergency situations. Conversely, Facebook posts and tweets are again more closely related to informational contents and current events. Dimension 5 (D5) – abstract and non-abstract information – accounts for the level of technical, abstractness and formality of a text, presenting passive clauses and conjuncts. The low scores registered indicate that the information provided in three corpora is non-abstract, popular and relatively informal. Finally, Dimension 6 (D6), or on-line informational elaboration, assesses whether the information given was produced under certain

time constraints, as for example in speeches. The data indicate that the texts were not produced under time constraints, with a low number of post-modifications.
The text type closest to the features present in fact sheets and tweets is 'learned exposition', i.e. texts that are formal and focused on conveying information (as in official documents, press reviews or academic prose). Conversely, Facebook posts appear to be closer to the 'general narrative exposition' type, i.e. texts that use narration to convey information as in press reportages, press editorials, biographies, non-sports broadcasts and science fiction.

4.4 Sentiment Analysis

In order to investigate sentiment in our corpora we proceeded to draw a list of verbs drawn from dictionaries and thesauruses accounting for the following categories in English and Italian: declarative, comment, judgement, predict, thanking, affect and request. The presence, frequency and polarity (positive or negative) attached to these verbs was coupled with observations on collocations for the most common natural disasters (Tables 6 and 7).

Natural hazard	**Collocations**
earthquake	catastrophic, devastating, extremely strong, large, major, massive, multiple, powerful, severe, significant, small
flood, flooding	(every) big, catastrophic, critical, dangerous, deadly, debilitating, destructive, devastating, heavy, historic, large, lethal, major, massive, minor, moderate, rapid, serious, severe, significant, widespread, sustained, widespread
(thunder)storm, hurricane storm superstorm cyclone storm	big, conventional, dangerous, deadly, destructive, devastating, difficult, disruptive, super-duper, epic, fearsome, furious, historic history-making, horrific, huge, awe-inspiring, intense, killer, large, lethal, major, massive, mighty, multiple, nasty, raging, severe, once-in-a-lifetime, perfect, powerful, once-in-a-long-time, small, strong, terrible, life-threatening, trouble, unprecedented, unrelenting, vicious, violent

Table 6: Examples of collocations for English.

Natural hazard	**Collocations**
valanga, slavina	grossa
nubifragio	violento
nevicata, neve	abbondante, bella, debole, forte molta, tanta
grandinata, grandine	forte, intensa, straordinaria, violenta
alluvione inondazione esondazione	devastante, grande, grave, tragica, tremenda disastrosa, drammatica, massiccia grave
tromba d'aria ciclone tornado	violenta devastante devastante, di debole/forte intensità
maremoto, tsunami	onde di maremoto/tsunami

Table 7: Examples of collocations for Italian.

These collocations indicate how natural disasters are described in the two languages: English uses a considerably higher number of adjectives, which qualify the perceived or estimated strength of the phenomenon, while Italian has a much narrower range of collocations of this kind and the evaluation seems to be of a less subjective nature and much more standardised through the use of words such as *allarme*, *allerta*, *pericolo*, *rischio* (alarm, alert, danger, risk). When combined with the language of prediction and forecast on the one hand, and disaster relief on the other, for example damage claims, sentiment can be leveraged to distinguish pre-disaster messages from during and post-disaster ones.

5. System Training

The analysis presented above has allowed us at Slándáil to gain a greater understanding of the textual features characterising our corpora and to pinpoint the main features of the communication of ages and the sentiment expressed both by them and by news outlets and the public at large. This body of knowledge informed the second stage of the study, which involves the training of a classifier using the CiCui system to recognise instances of messages relevant to emergency management. The training was done only on English texts but will be extended to Italian and subsequently to German as well.
The first stage of the system training consisted in the selection of potentially relevant messages from all our corpora through the use of keywords relating to common disasters, i.e. *earthquake*, *flood*, *snow* and *storm*. The sample available was reduced to 828 tweets and 644 Facebook posts and comments to allow for manual coding by three independent raters. Messages were assigned on- or off-topic status depending on whether they were deemed to be related to a disaster or not. For example, messages such as [1] 'Now a thunder storm. Typicall scotish [sic] weather' or [2] 'There's been a tectonic shift in UK politics; the SNP earthquake and Salmond's eruptive roar. It's a great day for Earth Science &the UK.' were both coded as off-topic since the first reports on what are considered 'normal' weather conditions and the second uses the keyword *earthquake* in a figurative sense. On the other hand, in cases like [3] 'That red spot was over us either last night or this morning...It was a bad storm!' or [4] 'Send some of that to Oklahoma for the next snow storm' were considered on-topic. Because social media messages refer to ongoing or temporary events and typically lack further context, the instances analysed often proved cryptic. For this reason, if URLs were provided, they were manually inspected for further clarification. In case of disagreement or uncertainty among the three raters, the message was coded following the rating agreed by two raters.
The training consisted first in treating raw texts with natural language processing techniques such as tokenization, lemmatization, part-of-speech tagging and dependency parsing;

an inverted positional index is created to record the occurrences of these lexical constructs and is stored as relational databases that can be later queried with standard SQL language. In this experiment, we also leveraged the system's ability to register the occurrences of linguistic patterns defined in a user-supplied dictionary; the patterns can be defined using a flexible regular-expression-like syntax powered by TokenRegex (Chang and Manning, 2014) , which allows for the capturing of sophisticated linguistic structures that are otherwise impossible to catch using traditional dictionary-based content analysis. Two types of features were extracted (Table 8): (1) traditional lexical features such as unigrams and bigrams, and (2) domain features such as sentiment, domain terminology, and social media specific constructs.

Feature Type	Feature Name	Description
Lexical Features	n-gram	Number of occurrences of individual unigrams, bigrams, etc. The frequency vectors of the n-grams were weighted using tf-idf and then optionally treated with Latent Semantic Analysis to reduce the dimension of the feature.
Domain Features	Sentiment	Number of occurrences of positive and negative words from the General Inquirer dictionary. When multiple word senses are encountered, the most common sentiment category of the probable sense was chosen. Two taboo words were also added to the dictionary to account for colloquial language usages on social media.
	Terminology	Number of occurrences of disaster-related terms as defined in the Slandail Terminology Wiki. These include both single-word and multi-word terms.
	Mentions, Tags, and Links	Number of occurrences of mentions, tags, and URL links in the message respectively.
	Locations and Date/Time	Number of occurrences of location names and date/time expressions as recognized by Stanford's Named Entity Recognizer.
	Media Type	A binary feature indicating whether the source of the message is Twitter or Facebook.

Table 8 Summary of features used to train the classifier.

All the features were scaled between -1 and 1 and cantered around the mean. A linear SVM with stochastic gradient descent training from the sklearn package (SGDClassifier) was used as the classifier. Various configuration of the features were experimented to test their impacts on the performances of the classification task. For each configuration, a stratified 10-fold cross validation was performed and the average accuracy, precision, recall, and F1-score across the folds are tabulated in Table 9, sorted by average accuracy.
The average accuracy, average precision, average recall, and average F1 scores are all in percentages. The highest value in each measure are put in bold. The baseline accuracy of the learning task is around 60% due to the imbalance between 'relevant' and 'irrelevant' labels in the dataset. All average accuracies for the configurations are significantly different from the baseline ($p < 0.01$). The worst performing configurations measured by average accuracy are those whose lexical features were exposed to extreme dimension reduction (LSA Dimension = 50). Using bigrams alone generally sees the same level of average accuracy as configurations using unigram or both, but bigram-only models suffer from low recall. Among the top ranking configurations differences are quite small. Interestingly, the use of domain features did not seem to have much impact on the performances of the classification: it seems that the lexical features on their own, when condensed with LSA, would yield sufficient discriminating power to distinguish disaster-related messages from irrelevant ones.

The present classification method will be further integrated in order to better specify features of social media messages and help the system improve its accuracy. All the messages assigned on-topic status will be further classified on the basis of the following criteria (adapted from Starbird et al., 2010):

- emergency phase: messages will be classified according to the phase of the emergency they refer to, i.e. pre, during or post disaster;
- source: is the sender of the message institutional or non-institutional?;
- re-sourced, retweeted, shared, follow@: is the information provided taken from another source (re-sourced), has it been retweeted or shared from another user, does it suggest that other users should follow a given account?
- providing or seeking information: is the message giving other users helpful information or is it looking for it?
- expressing support, humour, fear, celebrating, hopeful, and educational: what sort of sentiment does the message convey overall?

The five parameters listed above will inform further studies and will also allow to gain a more complete picture of institutional and non-institutional communication in the field of emergency management.

Domain feature	Lexical Feature	LSA Dimension	avg. accuracy	avg. precision	avg. recall	avg. F1
Yes	Unigram, bigram	500	**70.9**	**63.8**	60.6	62.0
Yes	Unigram, bigram	250	70.2	61.9	63.6	**62.5**
No	Unigram	250	70.2	62.7	60.8	61.6
No	Unigram, bigram	500	70.1	62.4	61.1	61.5
No	Unigram	500	69.6	62.1	59.4	60.6
Yes	Unigram	250	69.4	60.8	62.4	61.5
No	Unigram, bigram	250	68.8	60.5	60.1	50.1
Yes	Unigram	500	68.5	60.6	58.2	59.2
No	Unigram	n/a	67.4	59.3	56.3	57.4
Yes	Unigram	n/a	66.8	57.8	57.7	57.6
No	Bigram	500	66.6	64.8	33.3	43.7
Yes	Bigram	500	66.6	63.0	36.4	45.9
Yes	Unigram, bigram	n/a	66.0	56.8	57.0	56.8
No	Unigram, bigram	n/a	65.2	55.6	56.3	55.9
Yes	Unigram	50	63.5	53.3	57.0	54.9
No	Unigram, bigram	50	63.4	53.0	**64.3**	57.8
Yes	Unigram, bigram	50	62.3	51.9	59.6	55.3
No	Unigram	50	61.9	51.3	58.5	54.5

Table 9 Cross-validation results of the classification under various feature configuration.

6. Conclusions

Communication plays a central role in natural disasters. Emergency management agencies have now the means to communicate with the public at large through social media during disasters as well as in peace times. The Slándáil project aims at using the potential of social media communication to support emergency management agencies by filtering relevant messages. In order to do this linguistic analysis and training of a software system were necessary.

The linguistic analysis carried out aimed at accounting for, comparing and contrasting text complexity and readability of corpora in English and Italian of fact sheets, Facebook posts and comments, and tweets. Lexical density varies both across languages and text types, being higher in fact sheets than in social media for English but appearing considerably higher in Italian Twitter messages than in the other text types. However, all text types seem to aim much more at informing the public in a neutral way, rather than expressing judgement or trying to persuade.

Preliminary results from the software training suggest that the syntax used in the posts and tweets is more informative than the meaning carried by the domain features (e.g. sentiment and domain knowledge). However, this may be related to the fact that there are far more general language features than domain-specific features, thus the impact of the domain features is lower. This issue will be investigated in our future work. That being said, we have nevertheless shown that linguistic characteristics of text messages can be used to identify disaster-related communications on social media during emergency situations. The methodology proposed can be used to highlight good practices in social media communication, which in turn ca be used to provide guidelines for emergency operators.

7. Acknowledgments

The research leading to these results has received funding from the European community's Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL).

8. References

Avvenuti, M., Cresci, S., La Polla, M., Marchetti, A., & Tesconi, M. (2014). Earthquake emergency management by social sensing. In *2014 IEEE International Conference On Pervasive Computing And Communication Workshops (PERCOM WORKSHOPS)*, pp. 587--592.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Castello, E. (2008). *Text complexity and reading comprehension tests*. Bern: Peter Lang.

Chang, A. X., & Manning, C. D. (2014). *TokensRegex: Defining cascaded regular expressions over tokens* (Stanford University Technical Report).

Haddow, G., & Haddow, K. (2013). *Disaster Communications in a Changing Media World*. Burlington: Elsevier Science.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,. . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ripberger, J., Silva, C., Jenkins-Smith, H., & James, M. (2015). The Influence of Consequence-Based Messages on Public Responses to Tornado Warnings. *Bulletin of American Meteorology Society*, 96(4), pp. 577--590.

Slándáil Magazine (2015). http://slandail.eu/slandail-2015/

Starbird, K., Palen, L., Hughes, A., & Vieweg, S. (2010).

Chatter on the red. In *Proceedings of the 2010 ACM Conference On Computer Supported Cooperative Work - CSCW '10*. New York: ACM, pp. 241--250.

Ure, J. (1971). Lexical density and register differentiation. In G. Perren and J.L.M. Trim (Eds.), *Applications of Linguistics*, London: Cambridge University Press, pp. 443--452.

Zhang, X., & Ahmad, K. (2014). Ontology and Terminology of Disaster Management. In *DIMPLE: DIsaster Management and Principled Large-scale information Extraction Workshop Programme.*

Terminology Extraction for and from Communications in Multi-disciplinary Domains

Xiubo Zhang*, **Raffaella Panizzon**†, **Maria Teresa Musacchio**†, **Khurshid Ahmad***

*Trinity College Dublin
Dublin, Republic of Ireland
zhangx6, khurshid.ahmad@scss.tcd.ie

†University of Padova
Padova, Italy
raffaella.panizzon, mt.musacchio@unipd.it

Abstract

Terminology extraction generally refers to methods and systems for identifying term candidates in a uni-disciplinary and uni-lingual environment such as engineering, medical, physical and geological sciences, or administration, business and leisure. However, as human enterprises get more and more complex, it has become increasingly important for teams in one discipline to collaborate with others from not only a non-cognate discipline but also speaking a different language. Disaster mitigation and recovery, and conflict resolution are amongst the areas where there is a requirement to use standardised multilingual terminology for communication. This paper presents a feasibility study conducted to build terminology (and ontology) in the domain of disaster management and is part of the broader work conducted for the EU project Sládáil (FP7 607691). We have evaluated CiCui (for Chinese name 词萃, which translates to *words gathered*), a corpus-based text analytic system that combine frequency, collocation and linguistic analyses to extract candidates terminologies from corpora comprised of domain texts from diverse sources. CiCui was assessed against four terminology extraction systems and the initial results show that it has an above average precision in extracting terms.

Keywords: terminology extraction, software evaluation, multilingual communication

1. Terminology and Ontology of Social Media Streams

The existence of a term relies on textual evidence, i.e. the statistically significant occurrence of a term in a number of randomly sampled texts within a domain (Ahmad et al., 1994; Ahmad, 2001). Term extraction can be exploited to meet specific needs such as glossary compilation, translation, information retrieval (IR), ontology or conceptual map generation among others. The extraction of terms and of the information associated with them (i.e. definitions, synonyms, related concepts, etc.) from domain-specific corpora has received considerable attention, thus encouraging the study and development of a variety of automatic or semi-automatic extraction methods and tools – especially in fast growing disciplines such as biotechnology or computer science – and the broadening, update or harmonisation of existing termbanks and glossaries. A number of different systems are currently available both as freeware and proprietary software for a number of purposes such as research in linguistics, improvement of professional translators' performance or social media trend monitoring. Term extraction methods also facilitate the extraction of candidate ontologies (Ahmad and Gillam, 2005). The rise of Internet-based communications, and especially social media, has expedited the development of multidisciplinary terminology through the availability of large volumes of specialist texts in a variety of domains (Ahmad et al., 2006). Term extraction methods have been used in one of the traditional foci of social media analytics – film reviews and viewer sentiment (Manek et al., 2016). Term extraction and ontological mapping have made considerable progress since pharmaceutical companies have started to use social media to monitor reports of adverse reactions to drugs, also called pharmacovigilance — to do this the precise terminology of the domain needs to match a layperson's (patient administered a drug) use of the term. Terminology extraction techniques have been used in the pharmacovigilance domain with some success and are based on statistical machine learning techniques (Nikfarjam et al., 2015). Terminology extraction has been put to use in the surveillance of automotive component failure as reported by the 'buzz' in social media (Abrahams et al., 2013).

Multidisciplinary subjects, especially disaster management that involves a large number of agencies with different objectives but focussed on disaster mitigation and recovery, are characterised by terminology that is essentially a federated collection of terms from different constituent domains. A terminology collection in a multidisciplinary domain has to be carefully prepared and terms need to be elaborated fully. Emergency management and disaster relief organisations have developed and maintained terminology concerning natural hazards. The Sládáil project, whose goal is to ethically improve the use of social media in enhancing the response of disaster related agencies, has surveyed existing terminology resources, which usually take the form of glossaries containing entries with a term and its definition. These usually take the form of glossaries containing entries with a term and its definition. FEMA has an online glossary[1], Australia's EMA also has an emergency management glossary[2], while New Zealand's Ministry of Civil Defence and Emergency Management has local emergency management plans with short glossaries[3]. Similarly, Germany's BBK has a glossary[4] in

[1] https://www.fema.gov/rules-tools/glossary-terms

[2] https://www.ag.gov.au/EmergencyManagement/Tools-and-resources/Publications/Documents/Manual-series/manual-3-australian-emergency-glossary.pdf

[3] http://www.gdc.govt.nz/assets/Files/Civil-Defence/Glossary-Abbreviations-2009.pdf

[4] http://www.bbk.bund.de/SharedDocs/Downloads/BBK/DE/Publikationen/Praxis_Bevoelkerungsschutz/Band_8_Praxis_BS_BBK_

German and Italy's Protezione Civile has an official, concise one[5] on its webpage, while more extended glossaries are available in the webpages of voluntary organizations more or less closely associated with the national one[6]. While these glossaries reflect the need of emergency management organizations to communicate their terminology in a concise or more extended form, there is little if any information on how they were developed. Glossaries in PDF format are difficult to search and update. There are also some bilingual glossaries – Canada's English-French[7], Italy's Italian-English – both with definitions – and Germany's German-English, which provides only English equivalents. A trilingual glossary was found in Italy's South Tirol (Zivilschutz Glossary[8]). International organizations include some emergency management and natural hazards terms in their glossaries – cf. EIONET and IATE – while UNISDR developed a glossary of 53 terms which was extended to 80 in 2015 [9]. UNISDR also offers some information about terminology development[10], as it clearly states that terms were identified in a corpus of 35,000 documents and then validated by a group of experts. There seems to be still ample scope for automating term extraction, populating term entries and establishing conceptual relations through ontologies. Table 1 provides further details about each glossary, in particular how term entries are structured and how many terms are included as well as whether terms and other relevant information are linked to each other (hypertextuality), whether the glossary can be searched and navigated (interactivity) and if their format can be easily updated.

2. Sublanguages of Specialist Domains and Social Media

The use of language by humans in every domain of enterprise shows that some words and some linguistic structures are used more frequently than others. A sublanguage is "a specialized language or jargon associated with a specific group or context"[11] where words describing key objects/events/ideas and key activities are used almost exclusively for key objects and activities: 'bank' in financial transactions is not the same 'bank' as used in river engineering, and the activity where we rely or bank on others is confused by someone going to 'bank' money in a bank. The notion of sublanguages was propounded by Zellig Harris, a pioneering figure in modern linguistics, who has looked at the language used by mathematicians and biochemists and noted subtle differences (Schwartz et al., 2013) in language use between the two domains and between the general everyday language and the sublanguages in the two domains (Harris, 1991). Since Harris and others, there has been much work on the translation of sublanguage texts (Grishman and Kittredge, 2014).

There is also an equally important sublanguage that is shaped by the medium used – we had telegraphese in the 19th-20th century due to telegraphy technology, and the 140 character language, complete with shriek symbols and @ signs, i.e. Twitter, for the 21st century. It has been suggested that this is "what people say in social media to find distinctive words, phrases, and topics as functions of known attributes of people such as gender, age, location, or psychological characteristics" (Schwartz et al., 2013). A combined study of the sublanguage of a specialist domain and that of the (micro) blogging and social networking has been deployed to understand how patients are reacting to diseases like breast cancer (Elhadad et al., 2014). Sublanguage studies have been used in retrieving and analysing 'hazard related' posts on social media networks (Bolea, 2015).

3. Terminology Extraction Method

The terminology used in a sublanguage plays a crucial role in the characterisation of the conceptual composition of the corpus. Such information provides insights into the key issues and concerns in the domain of emergency management. The CiCui system implements a machine learning based automatic terminology extraction procedure. The system first extracts preliminary term candidates (TCs) by matching preprocessed text against predefined linguistic patterns; it then further refines the resultant TCs using statistical classifiers trained on previously labelled data. The workflow of the CiCui system is summarised in Figure 1.

3.1. Extracting Preliminary Term Candidates

The system first treats the input documents with natural language processing techniques; running texts are tokenised and tagged with part-of-speech information using the Stanford CoreNLP package (Manning et al., 2014). We employed the TokenRegex facility in Stanford CoreNLP to extract preliminary term candidates (TCs); TokenRegex allows matching word sequences using regular expressions specified at a token level (instead of at a character level as in normal regular expressions). The linguistic pattern used in our method is: `[word:/[a-zA-Z-]+/; tag:/NN|NNS|JJ/]+? [word:/[a-zA-Z-]+/; tag:/NN|NNS/]+`. The pattern matches noun sequences optionally modified with adjectives; all words in the term must consist of only letters and hyphens. Word sequences that match the above pattern are kept as preliminary TCs. Frequencies, document frequencies, tf-idf scores, and weirdness scores are computed for each word in the vocabulary of the corpus. The weirdness score for a certain word is a keywordness measure defined as the ratio between the word's relative frequency in a domain corpus and its relative frequency in a reference general corpus; in this case, the frequencies of

Glossar.pdf

[5] http://www.protezionecivile.gov.it/jcms/it/glossario.wp

[6] http://www.proingpa.it/wp-content/uploads/2011/10/Glossario-protezione-civile-rev1.pdf

[7] https://www.sdc.gov.on.ca/sites/mgcs-onterm/Documents/Glossaries/EMOGlossaryEN-FR.htm

[8] http://www.provinz.bz.it/zivilschutz/service/veroeffentlichungen.asp

[9] http://www.unisdr.org/we/inform/terminology

[10] http://www.preventionweb.net/files/45462_backgoundpaperonterminologyaugust20.pdf

[11] http://www.oxforddictionaries.com/definition/english/sublanguage

Name	Lang.	Term Entry	# of Terms	Hypertextuality	Interactivity	Updatable Format
(FEMA) Glossary of Terms – 2015	ENG	term (and acronym), definition	154	×	✓ (navigation from one letter to the other)	✓ (web page)
(EMA) Australian Emergency Management Glossary – 2016	ENG	term (and acronym), abbreviation, definition(s), synonyms, related terms, references	ca. 1100	×	×	× (PDF)
(NZ MCDEM) Glossary/Abbreviations – 2009	ENG	term, abbreviation, definition, references	66	×	×	× (PDF)
BBK-Glossar: Ausgewählte zentrale Begriffe des Bevölkerungsschutzes (glossary of selected key concepts of emergency management) – 2011	GER	term (and acronym), abbreviation, definition, synonyms, related terms, notes, references	176	×	×	× (PDF)
(Protezione Civile) Glossario (glossary) – n.d.	ITA	term (and acronym), definition, synonyms, related terms	278	✓ (some related terms are hyper-linked)	✓ (navigation from one letter to the other)	✓ (web page)
(Palermo Engineers' Association for Civil Protection and Emergency Management) Nuovo Glossario di Protezione Civile (new civil protection glossary) – 2012	ITA	term (and acronym), abbreviation, definition, related terms, references	459	× (only one hyperlink to an external source)	×	× (PDF)
(Emergency Management Ontario) English-French Emergency Management Glossary Of Terms – 2011	ENG FR	term (and acronym), definition, synonyms, references	123	×	×	✓ (web page)
(Bolzano Province) Civil Protection Glossary – 2013	ITA GER ENG	term(and acronym), synonyms	357	×	×	× (PDF)
(UNISDR) Terminology on Disaster Risk Reduction – 2009	ENG	term (and acronym), definition, synonyms, related terms, notes, references	53	×	×	× (PDF)
(UNISDR) Proposed Updated Terminology on Disaster Risk Reduction: A Technical Review – 2015	ENG	term (and acronym), definition, synonyms, related terms, notes, references	80	×	×	× (PDF)

Table 1: Features of glossaries from emergency management and disaster relief organisations

words in the COCA corpus (Davies, 2008) were used to derive the reference relative frequencies.

3.2. Refining Term Candidates with Statistical Methods

3.2.1. Model Training

The preliminary TCs extracted then undergo additional filtering using statistical classifiers. These classifiers were trained using a set of labelled terms extracted from another disaster management corpus separately prepared. The training corpus contained 2015 announcements and documents published by various disaster management agencies around the globe[12], totalling 418,513 words. From the list of 11,721 preliminary TCs which were sorted descendingly according to their tf-idf scores, five batches of TCs each containing 500 entries were sampled evenly every 2000 items; the sampled TCs were manually evaluated by two human raters and divided into three categories: 'Green' (G), 'Amber' (A), and 'Red' (R). 'Red' indicates that the TC does not constitute a valid term; 'Amber' indicates the TC is not identified as a validate term as a whole but may contain one; and the 'Green' category signifies a valid term. The distribution of the manually labelled categories is shown in Table 2. The distribution of the classes in the training set appears unbalanced; as a result, the baseline accuracy for classifications on the dataset will be 65%, which is achieved when all TCs are labelled with 'Red'.

Features For training the classifiers, the following features were used:

[12]The sources include FEMA, NASA Earth Observatory Natural Hazards, CDC Emergency Preparedness and Response, GDACS, and ReliefWeb.

Batch	R	A	G
1-501	164	205	132
2001-2501	294	186	21
4001-4501	350	144	7
6001-6501	422	75	4
8001-8501	364	136	1
Total	1594	746	165
Percentage (%)	64	30	6

Table 2: RAG distribution in training set grouped by batch

- The frequency, document frequency, and tf-idf of a TC: these statistics characterise the usage of a preliminary TC in the corpus as a whole. *TC frequency* is the number of times the TC was encountered in the text; *TC document frequency* counts the number of documents in which the TC ocurred; *TC tf-idf* is the tf-idf score calculated from TC frequency (tf) and TC document frequency df:

$$tf\text{-}idf = tf \times \log \frac{|D|}{df}$$

  where $|D|$ is the total number of documents in the corpus.
- The mean, standard deviation, maximum, and minimum of the frequencies, document frequencies, tf-idf scores, and weirdness scores of the constituent words for each TC: these features summarise the characteristics of the individual words in each TC.
- The length of a TC: unusually long word sequences can be a result of misclassification of part-of-speech or other issues such as malformed sentences.

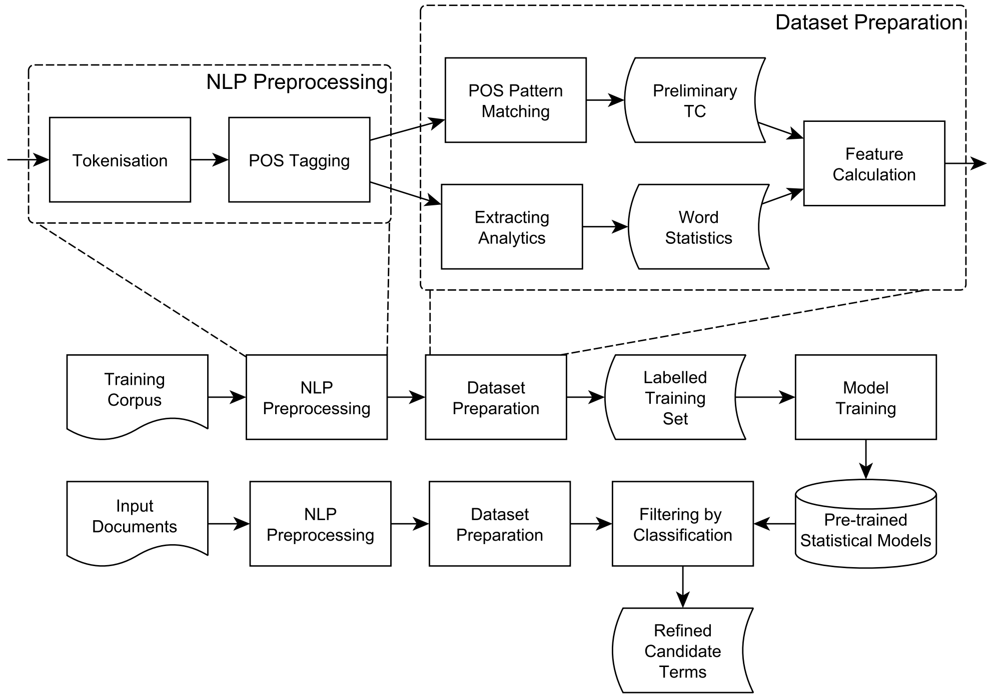

Figure 1: CiCui's Term Extraction Workflow

- The proportion of nouns in the TC: TCs comprised of mostly nouns are more likely to be valid terms.
- A binary feature indicating whether or not the first word in a TC is an adjective, and also the weirdness score of the first word of a TC.

All numerical features except for the proportion of nouns and the binary indicator were first logarithmically transformed and then normalised between 0 and 1 before further processing.

Model Performances We cross-validated a number of different classifiers on the training set using implementations provided by a data mining platform called KNIME (Berthold et al., 2007). The results of the cross-validation are tabulated in Table 3. Among the classifiers tested, random forest, neural network, SVM, and radial basis function network performed significantly better than the baseline, with the best being the random forest, which conferred a 7.47% increase in accuracy compared to the baseline.

3.3. Classification of Preliminary Term Candidates

Features for preliminary TCs extracted from new documents were prepared in the same way during the training. Based on the results from the training session, we classified new preliminary TCs by consulting an ensemble of three classifiers: a random forest, a SVM, and a multilayer neural network, all trained using the set-up described in Section 3.2.1. Each of the three classifiers votes for the preliminary TCs independently. A preliminary TC is kept if and only if it satisfies any of the following two criteria: (i) it received no 'R' label and at least one 'G' label; (ii) it received exactly one 'R' vote and two 'G' votes.

4. Term Extraction Evaluation

In this section, we present our evaluations of four commercial and open-source terminology extraction systems by looking at their extraction methods (statistical, linguistic or hybrid). The performances of these systems on a test corpus were evaluated and compared to the results returned by CiCui on the same test corpus. The performance scores of each are analysed and individual features are discussed. A summary of these systems is presented in Table 4.

Synchroterm is a Canadian-based statistical term extractor from Terminotix designed for professional translators and terminologists needing to create and manage translation memories. Though created to work with parallel texts (two languages), it can also extract terminology from unilingual documents. The testing has been conducted by selecting compound nouns from 2 to 8 elements without any stop list (though the option is available). The software allows a user to import lists of terms and expressions to be ignored during an extraction, and to create and modify a list of deleted items that are then ignored during all further extractions.

TaaS has been created within the EU project Accurat and uses a hybrid method, thus combining linguistic analysis (part of speech tagging, morpho-syntactic patterns, etc.) enriched by statistical features (e.g., frequency score). It supports all EU languages and Russian.

TermoStat Web 3.0 is a term extractor that uses both linguistic and statistical methods taking the potential terms'

Classifier	mean acc.	s.d.	mean diff.	t-statistic	p-value
Random Forest (100 Decision Trees)	71.47	2.02	7.47	11.71	< 0.01
Neural Network (2 layers, 10 nodes per layer)	70.27	2.35	6.27	8.43	< 0.01
SVM (linear kernel)	69.82	2.88	5.82	6.40	< 0.01
Radial Basis Function Network (Weka 3.7)	69.66	3.43	5.66	5.21	< 0.01
Logistic Regression	65.93	4.93	1.93	1.24	0.25
Multinomial Naive Bayes (Weka 3.7)	65.53	1.57	1.53	3.08	0.01
Naive Bayes	63.77	3.69	-0.23	-0.20	0.85
Fuzzy Rule Learner	63.00	2.67	-1.00	-1.18	0.27
Ordinal Logistic Regression (R, `MASS` package)	60.83	8.22	-3.17	-1.22	0.25

Table 3: Each classifier listed in the table was trained and tested with 10-fold cross-validation. The *mean acc.* and the *s.d.* column show the mean and the standard deviation of accuracies from the 10 classifications respectively. The *mean diff.*, *t-statistic*, and *p-value* columns show the result from a one-sample t-test (degree-of-freedom $= 9$) with the null hypothesis being that the average accuracy of the 10 classifications does not differ from the baseline (i.e. 64%).

structures and relative frequencies into account in the corpus analysis. It compares the specialised corpus provided by users with an in-built reference corpus for each of the languages it can process (French, English, Italian, Spanish and Portuguese). Users can choose to analyse mono and/or polylexical units. The English reference corpus has approximately $8{,}000{,}000$ words; half of it consists of news articles from the daily *The Gazette* published between March and May 1989 while the other half is taken from the British National Corpus (BNC). The corpus submitted is first tagged using TreeTagger and then, using regular expressions, simple or compound words are matched with predefined syntactic matrices.

Vocabgrabber is a software system that analyses text and generates lists of words and their use in context. No indication is provided as to the method applied to select and rank TCs though it is likely to be based on statistical frequency which can be sorted by subject (geography, people, social studies, etc.); ordered alphabetically, by relevance or "familiarity", i.e. frequency in general language.

Terminology Extraction System	Year	Target	Method
Synchroterm	2014	Translators, Terminologists	statistical
TaaS	2016	Companies	hybrid
TermoStat	2010	Linguists	hybrid
Vocabgrabber	2016	General public	statistical
CiCui	2014	Linguists, Companies	hybrid

Table 4: Summary of Terminology Extraction Systems

A test corpus consisting of $326{,}319$ words and comprising texts from FEMA fact sheets, handbooks from different emergency management agencies and news items extracted from LexisNexis (keywords: weather, emergency, disaster, Sandy, hurricane, superstorm, storm) has been used to test the systems described. In some cases only part of the corpus was analysed due to restrictions applied by systems. The results provided by each software system have been manually evaluated by assigning RAG labels to the top 100 TCs produced by each, and then compared with CiCui's performance.

Because an evaluation of recall requires to know all the terms present in the corpus in advance, only precision scores were calculated. The formula to calculate precision is:

$$P = \frac{A}{A + C} \times 100\%$$

where A is the number of accepted terms (i.e. 'Green') and C is the number of discarded results ('Amber' and 'Red'). The four extractors presented above have been tested on the same corpus, though in the case of TaaS and Vocabgrabber restrictions were applied, so that only part of the corpus was analysed. For the former $100{,}000$ tokens were processed, while for the latter a threshold was set at $200{,}000$ characters. Below are the results for precision and manual (RAG) evaluation calculated on the top 100 CTs returned by each system. Although 'Green' terms and precision express the same measure, i.e. the percentage of validated terms, both have been included in the table below for the sake of completeness.

The evaluation presented above allows for the comparison of the automatic term extraction system (CiCui) with currently available software performing similar tasks. It has been observed that term extractors have been designed with different users in mind (professional translators, businesses, linguists or general users), which strongly influenced the quality of their output. It also shows that automatic term extraction can greatly benefit from the adoption of machine learning techniques.

5. Conclusions

Successful disaster mitigation and recovery would not be feasible without the collaboration of experts from a variety of domains, who are bound to use more or less overlapping terminology. Therefore, all those involved in emergency management can greatly benefit from the standardisation of multidisciplinary and multilingual terminology. The evaluation presented above allows for the comparison of the Sláindáil automatic term extraction system (CiCui) with four currently available terminology extraction tools (two commercial and two open-source) performing similar tasks. It has been observed that term extractors have been designed with different users in mind (professional translators, businesses, linguists or general users), which strongly influenced the quality of their output. The comparison was carried out by calculating their performance on precision scores and by manually evaluating the top 100 TCs ranked by frequency.

System	# of TCs extracted	Precision	RAG evaluation		
			Red (%)	Amber (%)	Green (%)
Synchroterm	14791	56%	30	14	56
TaaS	345	42%	48	10	42
TermoStat	4082	64%	20	16	64
Vocabgrabber	2501	14%	78	14	14
CiCui	602	77%	6	17	77
Average			36.4	14.2	50.6
Standard Deviation			27.8	2.7	24.1

Table 5: RAG evaluation between terminology extraction systems

To this end, a test corpus was created and used to trial all five tools. Results highlighted substantial differences among them, with hybrid systems generally performing better in terms of precision and in the number of potentially invalid TCs. CiCui's term recognition showed above average results for precision and segmentation, although extensive work is being carried out to achieve further improvement.

6. Acknowledgements

The research leading to these results has received funding from the European community's Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL).

7. Bibliographical References

Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., and Zhang, Z. (2013). What's buzzing in the blizzard of buzz? automotive component isolation in social media postings. *Decision Support Systems*, 55(4):871 – 882. 1. Social Media Research and Applications 2. Theory and Applications of Social Networks.

Ahmad, K. and Gillam, L., (2005). *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005, Agia Napa, Cyprus, October 31 - November 4, 2005, Proceedings Part II*, chapter Automatic Ontology Extraction from Unstructured Texts, pages 1330–1346. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ahmad, K., Davies, A., Fulford, H., and Rogers, M. (1994). What is a term? the semi-automatic extraction of terms from text. *Translation studies: an interdiscipline*, 267:278.

Ahmad, K., Cheng, D., and Almas, Y. (2006). Multi-lingual sentiment analysis of financial news streams. In *Proc. of the 1st Intl. Conf. on Grid in Finance*.

Ahmad, K. (2001). The role of specialist terminology in artificial intelligence and knowledge acquisition. *Handbook of Terminology Management: Application-oriented terminology management*, 2:809.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer.

Bolea, S. C. (2015). Vocabulary, synonyms and sentiments of hazard-related posts on social networks. In *Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on*, pages 1–6. IEEE.

Elhadad, N., Zhang, S., Driscoll, P., and Brody, S. (2014). Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 516. American Medical Informatics Association.

Grishman, R. and Kittredge, R. (2014). *Analyzing language in restricted domains: sublanguage description and processing*. Psychology Press.

Harris, Z. (1991). Theory of language and information: a mathematical approach.

Manek, A. S., Shenoy, P. D., Mohan, M. C., and Venugopal, K. (2016). Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier. *World Wide Web*, pages 1–20.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

8. Language Resource References

Davies, Mark. (2008-). *The Corpus of Contemporary American English: 520 million words, 1990-present*. Availalbe online at `http://corpus.byu.edu/coca/`.