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.

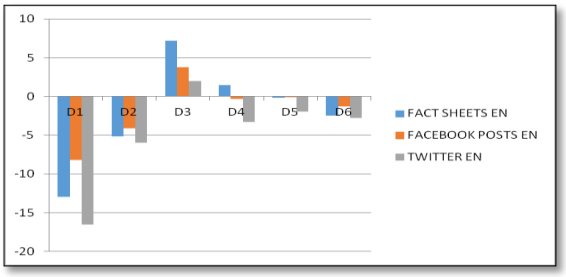

Table 5: Results of multidimensional analysis for English and Italian fact sheets, Facebook posts and tweets.

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.*