The Role of Unstructured Data in Real-Time Disaster-related Social Media Monitoring

Francesco Tarasconi, Michela Farina, Antonio Mazzei, Alessio Bosca
CELI Language Technology
Turin, Italy
Email: tarasconi@celi.it, farina@celi.it, mazzei@celi.it, bosca@celi.it

***Abstract*—Social media can be an important, constantly updated, source of information concerning natural disasters. User-generated, free text messages contain useful elements for the three main phases of disaster management: awareness/early warning, response, post-disaster assessments. However, most of the previous research focus on studying contents collected in relation to specific events. More work can be done in extending Information Extraction tasks to continuous streams of documents (potentially) hazard-related, regardless of time or location. We describe a Natural Language Processing architecture, employed in our study, to collect and monitor keyword-based streams, associated to different languages and event types. Starting from existing work, we review the definitions of disaster-related Information Types and Informativeness to better capture relevant and interesting items in the newly defined streams. To act as both a guideline in this procedure and a gold standard in automatic classification we created and annotated a multi-language, multi-hazard corpus of more than 10,000 tweets, sampled from our collected data-streams. We conclude by discussing the methodology behind and the results achieved by rule-based classifiers that we developed using domain and linguistic knowledge. Our approach is found to be viable in performing Information Extraction on generic, hazard-related (but noisy), social media data streams.**

***Keywords*-social media monitoring; natural language processing; information extraction; corpus annotation; text classification:**

I. Introduction and Related Work

Social media platforms are a popular tool to share information about everything that is happening around us. In the emergency domain such information can become a powerful resource for assessing the development of an hazard, its impact and how it is perceived by the affected population. Hence, the goals of a social media monitoring module include, in this context: retrieving contents related to different types of hazard, extract specific types of information that could be useful to citizens, first responders, and decision makers.

In the present work, we assess the value that can be provided by social media unstructured text in emergency management: without focus on a specific event, but on the broader scale of **Event Types**, that can be continuously monitored. Extracted information can be useful in different phases of emergency management: early warning/awareness, response, post-disaster.

In the work done by Olteanu et al. [9], the authors present the result of a manual labeling campaign to describe what to expect from social media data across a variety of emergencies (natural disasters, terrorist attacks, explosions, etc.) in terms of volume, informative level, type and source.

Klein et al. [5] propose a Natural Language Processing approach coupled with a clustering algorithm to tag tweets as related to an emergency event or not.

Several works have been done concerning the classification of online data into information classes or topics. Caragea et al. [1] evaluates Bag-of-Words approaches to classify text messages written during the Haiti earthquake and gathered by the Ushahidi platform[1] into different information classes.

A review of the use of big data generated by social media platforms during emergencies is provided in [2].

A good reference for Information Extraction tasks on free text is contained in [8].

Starting from previous work, that forms the basis for our research (Olteanu et al. [9] in particular), we consider two main categories of information to be extracted and that constitute the key features of the *classification model* we employ:

- **Information Types**, a taxonomy of common tweet content produced during emergencies, based on the specific kind of information contained in the tweet. The definition of Information Types is mostly taken from previous work. The main difference is moving from an exclusive Information Type classification to a multi-class model. Therefore, a post can have multiple Information Types associated to it.
- **Informativeness** class aims to capture useful posts which can actually help in the emergency management. The concept of Informativeness had to be reworked from previous work and generalized to involve the three emergency management phases.

Additional features can be extracted, such as geographical information implicit in the post content and the presence of panic/anxiety. The first one relates to Named Entities of Location type and is instrumental in the proposed automatic classification of informative content, as we will

[1] https://www.ushahidi.com/

see in Section IV. Detecting panic can instead be seen as a kind of Sentiment Analysis: we will not explore this task in the present work. The monitoring procedure and the classification model we employ are described in Section II. We have conducted two campaigns of text annotations of Twitter documents, or *tweets*, with the aim of exploring the robustness and subjectivity of our classification model (measured as inter-annotator agreement) and build automatic classifiers to classify new contents according to this schema. During the first campaign, we considered only Italian tweets concerning Floods, producing an *IT-Floods corpus*. Later, we considered tweets collected in relation to different hazards, therefore belonging to different Event Types, and in multiple languages. During the second campaign, we annotated as much as eight Event Types for three different languages: Italian, English, Spanish. We created a *multi-language, multi-hazard corpus*. The main feature of these corpora is the way the tweets were extracted i.e. by *listening* to generic social media streams, without a focus on specific events. They contain informative content concerning all phases of disaster management. The creation of corpora, the manual annotation process and the measurement of inter-annotator agreement are described in detail in Section III. We then developed and tested automatic classifiers with the goal of continuously classifying relevant contents from event-related social media streams. The first corpus was used in the development phase of rule-based classifiers of Information Types and Informativeness for Italian posts, later translated into other languages: English and Spanish. A key step in this phase, detailed in Section IV, was understanding the relation between Information Types, Informativeness and tweets containing a textual reference to a specific location (not necessarily geo-tagged). Rules, originally queries on text content in Italian, were translated and adapted where necessary using professional linguistic knowledge. They were then tested on the second corpus. Performances of rule-based classifiers were measured on both datasets. Two types of performances are measured on the test set: on hazard-related data only (per manual annotation, theoretical performance) and all the dataset (to measure effectiveness in real-world applications, such as the one we propose). Evaluation of performances is also described in Section IV. The approach, after accepting a certain degree of subjectivity and error in the classification procedure, seems viable for the Information Extraction task, without specific knowledge about events taking place and their current progress.
Section V summarizes the conclusion of our work.

II. Mining Twitter

The social media platform that was monitored in our research is Twitter. It is widely used in the study of natural disasters (see: [5], [9]–[11], [13]). Because the basic form of communication on Twitter (the limited length *tweet*) is essentially a broadcast, the platform is especially suitable for the emergency management use case: relevant keywords and *hashtags* (characteristic user-generated metadata tags) can be monitored and related content from most users can be extracted and analyzed in near real-time.
In this section we describe the monitoring system used in the current study, specifying how we access and analyze Twitter data, what can kind of information we extract in the process i.e. the adopted classification model.

A. Data Streams

The monitoring module was implemented in Java and Scala as a job within the Spark Streaming architecture[2], which is an open source infrastructure designed to deal with real-time data analysis, transformations and operations. To retrieve social media contents, the monitoring modules relies on the Streaming API provided by Twitter[3]: these APIs are designed to allow monitoring specific topics (or users) enabling low latency access to Twitter global stream of data. The streaming clients will be pushed messages without any overhead associated with polling a REST endpoint. However, the public, cost free, Streaming APIs are characterized by an overall volume limitation of 1% (randomly subsampled) of the total Twitter stream[4]. Therefore, in order to maximize the volume of retrieved relevant contents, it is important to limit the off-topic contents by configuring the access to the global Twitter stream with one of the different filtering parameters exposed by the Streaming APIs:

- *language*: the content language,
- *locations*: one or more geographical regions,
- *follow*: a list of author IDs to follow,
- *track*: a set of keywords, key-phrases (including one or more terms separated by spaces) and/or hashtags that should be present in the content. At least one of the track items must be found for a tweet to be included in the collected stream.

Among these filtering parameters, our social media monitoring module uses the language and the track items. The follow parameter is not pertinent in our use case, since the module aims to retrieve all the contents related to a topic regardless of the author. Locations parameter instead is too restrictive, as it filters out all the non geo-localized data from Twitter stream and currently only a minority of such data is tagged with location (less than 2% of the posts, see: [6]). The monitoring module is thus configured with a set of *track queries*: one for each combination of the supported Event Types and languages. Track queries are textual queries that constitute a first, coarse-grained level of classification for the contents, as their goal consists in restricting the full-stream contents to a specific hazard type without losing too

[2]https://spark.apache.org/streaming/
[3]https://dev.twitter.com/streaming/overview
[4]https://dev.twitter.com/rest/public/rate-limiting

much relevant data.
Track words were selected following standard practices commonly used for this type of data collection, and typically include hashtags and common names of the disaster. Proper names (e.g. Hurricane Irma) or affected locations (e.g. China) were not inserted to keep the analysis independent from specific events and better generalize to future ones. We report below the list of monitored Event Types and an extract of the track words we used for the English language. Similar lists were compiled and used for Italian and Spanish languages.

- **Floods**: floods, flood, flooding, #floodsafety, #floodaware, #floodprep, #floodsmart, #floodready
- **Wildfires**: #fire, wildfire, wildfires, #fires
- **Storms**: #storm, stormwarning, stormalert, #stormy, stormageddon, #StormWatch, lightningstorm, thunderstorm, #toomuchrain, #somuchrain, #stormyweather, windstorm, typhoon, tornado, #rain, tornadowarning, typhoonwarning, hurricane, hurricanewarning, typhoonalert, tornadoalert, hurricanealert, typhoons, tornados, hurricanes, storms
- **Extreme Weather Conditions**: #frost, "heat wave", heatwave, "hot weather", #hotweather, #extremeweather, #climatechange, "climate change", "cold winter", "ice winter", "frosty winter", "frozen winter", "cold weather", "ice weather", "frosty weather", "frozen weather", "hot summer", "hottest summer", "hot weather", "hottest weather", "hot sun"
- **Earthquakes**: #earthquake, #earthquakes, #seismicwave, #seismicwaves, magnitude, epicentre, seismic
- **Landslides**: landslide, landslides
- **Drought**: drought, droughts, "water scarcity", dryness
- **Snow**: #snow, "snow day", snowing, snowdays, snowfall, snowalert, avalancherisk, snowmageddon, snowapocalypse, avalanchedanger, snowpanic, bigsnow, snowpocalypse, snowstorm, snowday, blizzard, blizzardalert, blizzardwarning, avalanche, avalanchewarning, avalanchealert

B. Event Classification

The contents in the filtered stream are then processed through a Language Analysis pipeline (involving lemmatization, key phrases detection, named entity recognition, classification and sentiment analysis) that enriches them with additional linguistic and semantic meta-data. Within the analysis pipeline, a more refined classification (compared to the original extraction) is applied to the contents in order to filter out unrelated items (e.g. landslide victory, flood of votes). These classification rules are based on language and semantic features (i.e. lemmatization, proximity expressions, exclusions) and are manually composed by domain experts (mother tongue and professional linguists).
The resulting related contents are stored in a database to be further classified at a more granular level (see Information Type and Informativeness in Subsection II-C).
CELI[5] proprietary resources were used in the Language Analysis pipeline. More details on the used pipeline can be found in [12]. Storage was performed on PostgreSQL, a well known open source database[6].
In our current system, three languages are supported by the linguistic pipeline of analysis: English, Spanish, Italian. Event Types that are classified are the same as reported in II-A: Floods, Wildfires, Storms, Wildfires, Extreme Weather Conditions, Earthquakes, Landslides, Drought, Storm. Each Event Type has a refinement/classification rule associated to it. Each combination of language and Event Type defines a sub-stream (compared to the full Twitter stream) of unstructured documents that is fed to subsequent Information Extraction modules. We report some of the rules used at the level of Event classification: as we can see they employ a more powerful syntax, compared to track queries. Rules are of Lucene Query kind; they follow Lucene Query Syntax[7], and work with logical operators (AND, OR and NOT). When a document matches a rule, it will be inserted into the corresponding class. More details on the classification process are reported in Section IV and in particular in Subsection IV-A.

- **Extreme Weather Conditions**: frost "heat wave" heatwave heatwaves hotweather extremeweather ((climatechange "climate change")NOT(dup party parties president deny* denier deniers)) ((cold ice frosty frozen)AND(winter weather)) ((hot*)AND(summer sun weather))
- **Landslides**: (landslide*NOT(votes party parties president))
- **Drought**: ((drought*)AND(water flood* rain emergency aid famine poverty poor* refugee* emigr* immigr* migr* farmer wine* climat* alarm* prevent dry* weather nutrients rainwater surviv* starv*)) "drought in" "water scarcity" dryness

C. Information Types and Informativeness

In this work we check the possibility of classifying emergency content using only unstructured information inside the message. We create a classification system based on natural language processing rules. That will make the system able to identify Informativeness and Information Type from language knowledge and Event Type knowledge, without any specific information about the specific disaster event. This allows us to work a priori on any event considered in the classification model. Section IV deals in greater detail with an actual classifier of these elements; in this Section we outline, instead, the classification model.

[5]https://www.celi.it/
[6]https://www.postgresql.org/
[7]https://lucene.apache.org/

Previous literature provides different types of classifications, in terms of Informativeness and Information Type, that can potentially generalize to multiple events (see [9]). Original Information Types proposed in literature are:

- **Affected Individuals**: information about casualties, killed, missing or displaced people, including personal updates about oneself, family, or others. Medical emergencies were also considered relevant.
- **Infrastructure and Utilities**: information about affected buildings, roads closed, utilities/services that are damaged, interrupted, restored or operational.
- **Donations and Volunteering**: donations of money, goods or services; financial relief, charity, fund raising; practical relief; volunteer information and coordination; food collection and distribution.
- **Caution and Advice**: information about warnings issued or lifted, guidance and tips.
- **Sympathy and Emotional support**: thoughts, prayers, gratitude, sadness, etc.
- **Other Info**: specific information but NOT covered by any of the above categories.

The main change from previous researches was to move from mutually exclusive Information Types to a multi-class data model where a single tweet can be classified, for example, as related to Affected Individuals *and* Infrastructure and Utilities. This allows more flexibility in classifying tweets and the classification model can more easily accommodate changes in time to the classification structure (previously assigned classes are maintained when new classes are added to the model). During the course of our research, we moved from the "Donations and Volunteering" single class originally suggested in literature to two separate classes, one for Donations and one for Volunteering. We believe that distinguishing between these classes can provide more precise information during different emergency phases: Volunteering is thought to be of greater use in the response phase, whereas Donation mainly impact the post-disaster management.
As described in Section III, after a campaign of text annotations, we discarded the "Other Info" class, because of too low inter-annotator agreement. Instead, we plan on adding Information Types and features with clearer semantics, in a manner similar to what we did with geographical (implicit in text content as opposed to explicit geolocation) information (see section II-D). We hypothesize that the reason why we found the class "Other info" of scarce practical use in our classification model is because we employ a multi-class model. In a single-class Information Type model it acts as a "miscellaneous" class. Its semantics in a multi-class model appear much less clear. Therefore the Information Type model we adopted is:

- **Affected Individuals**;
- **Infrastructure and Utilities**;
- **Donations**: donations of money, goods or services; financial relief, charity, fund raising.
- **Volunteering**: practical relief; volunteer information and coordination; food collection and distribution.
- **Caution and Advice**;
- **Sympathy and Emotional Support**.

A document, related to any hazard, can belong to a single, multiple or none of these classes.
Informativeness notion, originally tied to awareness and usefulness *during* an event (see [9]), had instead to be revised to operate in the more generic context we propose. We define Informative content in Section III. A strong relation with Information Types and geographical information contained in the text content was found during the automatic classification process, described in Section IV, where we will also explore in more details the task of correctly classifying tweets. Ground truth for this task is established in Section III with the manual annotation of corpora.

D. Other Information Classes

As described in Section II-A, by relying on a pre-existing NLP pipeline, we can extract additional elements, such as the Panic feature (a type of Sentiment Analysis, see for example [7], [12]) and the presence of geographical Named Entities (locations). A good overview of the Named Entity Recognition task on Twitter is contained in [4].
Panic can be important in social media monitoring, for example to provide real-time panic maps or indexes of panic levels: we won't go into further details in this work.
Locations, implicit in the tweet text content, can be important as well to generate maps of specific phenomena; in the present work, we will use their presence or absence as an additional feature in the Informativeness classification task (see Section IV).
The usage of automatically extracted key-phrases from classified posts to gather bottom-up additional information about their content is explored in Section III-B.

III. The Corpora

We empirically found that track queries and Event Type queries (respectively shown in Subsection II-A and Subsection II-B) accomplish their job of extracting (potentially uninteresting or uninformative) hazard-related tweets on several streams, without discarding too much relevant content. We didn't conduct a full study of performances of these components (which is especially challenging in the case of measuring Recall over the whole Twitter stream). We estimate the amount of actually emergency-related tweets to be about 80% of tweets stored in the database. Our main interest lies in assessing the information quality and classification performances that can be achieved on Informativeness and Information Type recognition on multiple streams (by language and event-type).
The key requirements in these Information Extraction tasks are: we want to classify all tweets corresponding to certain

queries, potentially related to several catastrophic events happening in the same time-frame and in different stages of development; we want to capture a notion of Informativeness that generalizes to the awareness/early-warning and post-disaster phases.
Therefore, we constructed a multi-language and multi-hazard text resource, had it annotated according to Information elements of interest by linguist professionals. A first corpus (single Event and single Language) was used in the development phase of automatic, rule-based, classifiers. The second, wider corpus was used as test of classification rules on English, Spanish and Italian languages, on all the hazards we monitor (Section IV). Both corpora were used to measure inter-annotator agreement and assess the robustness of the chosen information classes.
Corpora will eventually be made available for further researches.

A. Data Collection

In previous related work, tweets were mostly collected ex post with the help of crisis-specific hashtags. Instead, we used tweets collected by generic keywords before, during and after emergencies.
Download from Twitter produces hundreds of thousands of (potentially event-related) documents per day; we selected randomly the ones to be manually labeled. All tweets were sampled by the streams collected with the configuration described in Section II.
The first corpus is composed by 1,186 tweets, collected in November 2016 from the Italian language and Flood Event Type stream. During that period, flooding struck, among the others, the Liguria and Piedmont regions, therefore tweets might be related to different locations and crises. The corpus also contains tweets collected in the preparedness phase and in the post disaster one.
Composition of the second corpus (containing 9,695 tweets in total) is reported in Table I. We collected Italian tweets about Earthquakes and Snow between October 2016 and January 2017. Rest of the tweet collection took place during Summer 2017: we were therefore able to cover all hazards for Italian language (leaving out Floods that is tackled by the first corpus). For Spanish and English languages, some hazards were not deemed of interest in the considered period: we excluded Snow from Spanish and English, and Floods from Spanish.

B. Dataset Contents Overview

In this Subsection we present a qualitative representation of the contents in the datasets, obtained by visualizing, per dataset and language, the most frequent *emerging key-phrases* (not to be confused with the key-phrases used in the track queries). This representation also serves as a first validation of the downloaded tweets that were included in the emergency-related corpora.

Table I
LANGUAGE AND EVENT COMPOSITION OF MULTI-LANGUAGE AND MULTI-HAZARD KEYWORD BASED CORPUS

Event type	**EN**	**ES**	**IT**	**Total**
Drought	500	195	500	1195
Earthquakes	500	500	500	1500
Extreme weather conditions	500	500	500	1500
Floods	500	-	-	500
Landslides	500	500	500	1500
Snow	-	-	500	500
Storms	500	500	500	1500
Wildfires	500	500	500	1500
Total	3500	2695	3500	9695
Between Oct. 2016 and Jan. 2017	-	-	1000	1000
Between June 2017 and July 2017	3500	2695	2500	8695

These key-phrases are extracted within the language analysis pipeline by identifying specific patterns of terms with desired linguistic features (e.g. a noun followed by a preposition and another name, an adjective followed by a noun). *Word clouds* are then computed by selecting the most frequent key-phrases within a subset of tweets: for example, by restricting to a certain period, hazard, or, in this case, to the considered corpora. Figure 1 represents a word cloud of the most frequent key-phrases extracted from the IT-Flood corpus (mainly related to flood alerts and heavy rains in Piedmont and Liguria areas), while Figures 2, 3 and 4 from the multi-language, multi-hazard one (different clouds for different languages).
It can be observed that the main topics emerging from the multi-hazard word clouds are related to June 2017 alerts:

- **English**: tweets are about crises from all over the world: the most important are "China landslide", "earthquake off Turkish and Greek coasts", but also domestic alarms about Storm Cindy: "severe thunderstorm", "flood advisory", "tornado warning".
- **Spanish**: the news are about both domestic problems ("ola de calor") and foreign news: "incendio en Londres" (London skyscraper fire), "deslizamiento de tierra" (China landslide) and "sismo de magnitud" (Greece earthquake)
- **Italian**: most tweets are about domestic problems: "ondata di calore", "ondata di caldo" and similar expressions are about extremely hot temperatures in Italy during Summer 2017; "emergenza siccità" and "siccità a Roma" are about drought in Rome.

From this overview we also note that key-phrases can be an useful instrument in order to assess the presence of meteorological events, like droughts, as well as the affected locations.

Figure 1. Word cloud of frequent emerging key-phrases (IT - Flood)

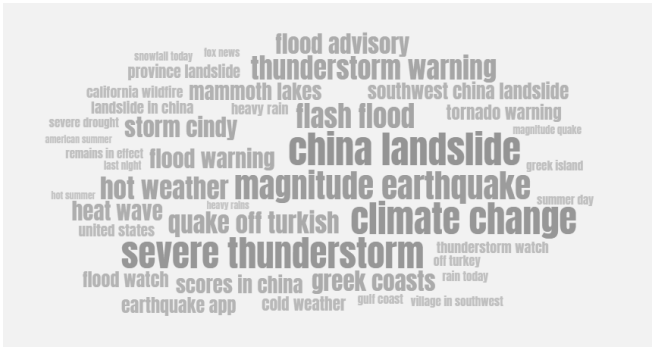

Figure 2. Word cloud of frequent emerging key-phrases (EN - Multi-hazard)

C. Annotation Process

We employed native language speakers to perform manual annotation of tweets, during the course of September 2017 and in two separate campaigns. We refer to Subsection III-A for specific compositions of the corpora.

Annotators were all professional linguists or translators. We used native speaker annotators to make sure they understood every nuance of the language (e.g. irony, panic, hurry), in addition to cultural and geographical references. They

Figure 3. Word cloud of frequent emerging key-phrases (ES - Multi-hazard)

Figure 4. Word cloud of frequent emerging key-phrases (IT - Multi-hazard)

were not the same linguists involved in the downloading and classification query writing (Section II and IV), to prevent their experience from influencing the annotation task. Annotators received an Excel file with the tweet text (including any links, which they were invited, but not forced, to follow), author, timestamps, and the kind of natural hazard it was downloaded in relation with.

In the annotation file, there was a column for every type of label: annotators should mark as (t) all the category columns corresponding to the categorization they choose and mark as (f) all the categories the tweet does not fall into. A tweet could be associated to more than one Information Type, or to none of the Information Types.

The first task was the the **Relatedness task**: annotators had to classify tweets as related or not related to hazards. Downloading from Twitter was keyword based, and a tweet may contain relevant keywords but be unrelated to the hazard. Some keywords may be quite general, and refer to any number of topics other than the disaster situation. Some keywords could be proper nouns or brand names, or convey spam. In this sense, we could use the results of this first phase to clean our tweet databases and track queries. If annotators labeled a tweet as not related (f), they had to skip the Informativeness, Information Type and Panic detection tasks for that tweet, but Geographical location task was compulsory anyway (see below).

The second task was the **Informativeness task**: annotators should classify tweets as informative or not informative in relation to natural hazards. This task was the one that raised more questions and more perplexity among annotators, because of its subjectivity. Moreover, we avoided giving the annotators indications that were too specific in order to leave them a margin of reasoning.

Informativeness definition: *Annotators should classify tweets as Informative (t) if they contain useful information that could help people understand the hazard and the situation. All phases of the emergency must be considered of potential interest: before, to get prepared; during, to respond effectively; and after, to better manage consequences. Tweets are considered Not Informative (f) if they do not contain*

useful information that could help people understand the situation.

Then annotators had to label every tweet with the **Information Type(s)** they convey, writing true or false in each corresponding column. Information Types have already been described in Section II and the report of a similar annotation task can be found in [9].

Last two tasks were Panic annotation and Geographical Entities annotation.

In the **Panic annotation** manual tagging annotators were asked to classify tweets as "panic detected (t)" or "panic not detected (f)". In the first case, the text of the tweet conveys a sense of panic or worry. In the second one, it does not.

In the **Geographical Entities** annotation task annotators were asked to mark tweets containing a geographical location. The tweets should contain a textual reference to a specific, well-delimited, real world place/location.

In the annotation phase, the Informativeness and Information Type tasks were given unrelated to each other: each annotator was free to annotate a tweet as Informative without marking any Information Type as (t).

We collected labels from 3 different annotators per tweet and task, and determined the final label of each tweet by simple majority. A total of 12 different annotators were employed in the process. Being in direct contact with annotators allowed us to clarify any formal mistake or missing annotation. All tweets in both corpora were fully annotated.

D. Measure of Agreement

We measured inter-annotator agreement on both corpora. In Tables II and III we report the final frequency of agreed (t) annotations ("Freq." column), and the related agreement, as average percentage agreement between annotator pair over the whole corpus ("Mean agr." column) and average Cohen's K ("Mean agr. K" column). Most annotations score what we deemed a fair agreement, above 0.30 in terms of K (see [3]). We report additional observations below.

1) Agreement on Other Info is low, and we decided eventually to discard this class from the chosen classification model. We provide a possible explanation of this fact in Subsection II-C.
2) In terms of commonly found Information Types, we observe that Affected Individuals and Infrastructures and Utilities score a much higher agreement compared to Caution and Advice.
3) Low agreement on Sympathy and Emotional Support and Panic annotations was expected, given their subjective nature.
4) As reported in Subsection II-C, after the annotation of the first corpus, "Donations and Volunteering" class was split into "Donations" and "Volunteering" separate classes. Agreement on "Donations" is much higher than agreement on "Volunteering".

Table II
CLASS COMPOSITION AND AGREEMENT OF IT-FLOOD KEYWORD BASED CORPUS (1,186 TWEETS)

Class	**Freq.**	**Mean agr.**	**Mean agr.** K
Related to hazard	700	73%	0.48
Affected individuals	102	96%	0.78
Caution and advice	273	73%	0.37
Donations and volunteering	22	97%	0.31
Infrastructures and utilities	191	91%	0.68
Sympathy and emo. support	9	97%	0.20
Other info	213	63%	0.09
Informative	640	70%	0.35
Panic	19	96%	0.27
Contains location	707	89%	0.79

Table III
CLASS COMPOSITION AND AGREEMENT OF MULTI-LANGUAGE AND MULTI-HAZARD KEYWORD BASED CORPUS (9,695 TWEETS)

Class	**Freq.**	**Mean agr.**	**Mean agr.** K
Related to hazard	7626	88%	0.64
Affected individuals	1380	96%	0.82
Caution and advice	927	79%	0.31
Donations	39	99%	0.46
Infrastructures and utilities	967	92%	0.59
Sympathy and emo. support	155	97%	0.43
Volunteering	69	98%	0.22
Other info	2005	69%	0.14
Informative	5682	71%	0.43
Panic	167	85%	0.11
Contains location	5155	85%	0.80

5) A fair agreement was reached on the Informative annotation, although the task is clearly influenced by subjective judgment (percentage agreement of about 70%).

The corpora can overall serve as a resource for the tasks of Information Extraction we described on hazard-related data streams. Specific annotation agreements can serve as a measure of subjectivity, and prompt further researches in taxonomies for classifying emergency-related tweets.

IV. AUTOMATIC TEXT CLASSIFICATION

Document classification is the task that consists into assigning a document to one or more classes or categories, depending on the words it contains (*content-based* classification). Every document can be classified into one class, more than one class or no class at all.

In this Section we describe an automatic classification system implemented within the monitoring system we described in Section II. The system is a rule-based multi-classifier that operate efficiently on multiple language-hazard streams, effectively analyzing hundreds of thousands of documents in a given day. We describe the methodology behind the

classifiers and report performances they achieve on the two annotated corpora (introduced in Section III) in the Informativeness and Information Type classification.
We aim to provide a comparison point for other rule-based and machine learning automatic classifiers, at the same time showing the Information Extraction tasks we discussed are viable on generic hazard-related data streams, albeit with different levels of subjectivity and expected performances.
The methodology used to classify texts is that of *NLP-enhanced* queries. This approach allows not only keywords to be used in queries, but also other information extracted from natural language processing analysis.
The process of developing rule-based classifiers involved manually analyzing the annotated corpus and utilizing domain knowledge. This in turn gave us insights on the thought process followed by annotators in classifying Informative content.
Such classifiers will detect Information Types and Informative tweets, regardless of the hazard they are related to: however, knowledge of the Event Type taxonomy was used during development. Therefore, they should be tested on new Event Types, as performances are likely to drop in this case.

A. Implementation

Documents are processed through the linguistic analysis pipeline described in Section II. Sophia Analytics, CELI proprietary text mining software that allows to analyze unstructured text in several languages, was employed to classify according to Informativeness and Information Types. Within the tool, it's possible to search keywords as *Lemma* or as *Exact Match*. If you search as Lemma, the tool will consider the word you put in the rule as whole lemma, and will match all the correspondent flex forms; if you search as Exact Match, system will match the exact portion of text as it was put into the rule. Moreover, is possible to take into account the *lexical categories* and *parts of speech*.
Classifier development began with an empirical study on downloaded documents from Twitter until November 2016, and in particular on the IT-Flood corpus described in Section III. During this phase, experienced linguists extracted relevant keywords for the classification of Information Types. Manual and a priori classification was compared with manual annotation results of the IT-Flood corpus.
The relations between Informativeness and different information classes (Information Types and other results of NLP analysis, such as Panic and the presence of Geographical Entities) were explored.
It emerged that it is possible to classify the "Informative" tweets, with high Precision and Recall (see Subsection IV-C), as the aggregation of:

1) "Affected Individuals", "Infrastructures and Utilities" and "Caution and Advice" Information Types plus
2) tweets containing at least one Geographical Entity.

Queries resulting from this phase were then tuned for other languages, as most of the work up this point had been done on Italian.
Performances of classification were finally measured on the multi-language, multi-hazard manually annotated corpus.

B. Examples of Classifiers

We report examples of text queries to classify Information Types (English language).
Terms enclosed between [ ] match the lemmatized form of the textual contents. Terms (or expressions) preceded by a plus sign MUST be present in the retrieved contents. Terms (or expressions) preceded by a minus sign represent term that should NOT be present in the retrieved contents. Expressions followed by a tilde and a number N are proximity expressions that identify documents containing all the terms in the expression, each one within a maximum distance of N terms from the others.

- **Affected Individuals**: "medical emergency" trapp* casualt* "people missing" "people found" "found alive" "found still alive" "seen alive" fatalit* injur* "i'm ok" kill* victim* rescue* lifesav* saved resident* witness* dead* "death toll" "several missing" "people missing" deaths "person missing" "hundreds of families" "took refuge" evacua* displac* died (+affected +(animals people)) evacuees (+forc* +("from homes" 3)) [witness] "flee homes" "ordered to flee" "people are feared" buried (+people +(buried buries bury missing)) "bodies found" [student] [tourist]
- **Caution and Advice**: [risk] (+predict* +weather) forecast* caution advice [tip] [alert] alarm* "is expected" "are expected" protocol* prepare* instruction* protect* "will impact" hit affect [prepare] [prevent] warn* emergenc* defence prevention "risk management" damage aware* floodaware [security] [resilience] watch "you need to know" advisor* "ordered to" "orders to" "notices ordered" "evacuation notices" "evacuation notice" breakingnews
- **Infrastructures and Utilities**: damag* [road] [street] [bridge] [building] [subway] [highway] [hospital] [clinic] (+water +(shortage sanitation)) (+traffic -light*) (+(fire burn*) +(home* acres)) blackout "black out" propertydamage "close for" "closed for" "until further notice" "close because" "closed because" (+home* +destroyed) "structure toll" "sweeps away" "sweep away" "swept away" "tourist sites" "tourist site" (+(buries buried bury covered) +(house* homes village*)) [hotel] [motel] [restaurant] [store] [freeway]
- **Volunteering**: [volunteer] relief* reliev* "brings care" "bring care" (+food +(collect* distrib*)) "food will be sent" (+assist* +victim*) "red cross" aid "distribution center" "how to help" "help those" salvationarmy redcross

As discussed in Subsection IV-A, the Informative query we used is different from the above as it does not employ specific keywords. Instead it consists in the set union of:

- Affected Individuals;
- Infrastructures and Utilities;
- Caution and Advice;
- tweets found containing at least one Geographical Entity, per Named Entity Recognition component of the pipeline (Section II).

C. Results

Performance on the manually annotated corpora are reported in Tables IV, V and VI. Each table row contains evaluation of the corresponding binary classification task; a single document can have none or multiple labels associated to it.

Table IV contains evaluation on the IT Flood corpus, used as a development set for the rule-based classifiers. As the theoretical goal of the examined classification tasks is to recognize informative tweets within generic hazard-related tweets, we measured development performances only on "related to hazard" tweets (numbering 700).

The relative frequency of each class (on the 700 related tweets, per manual annotation) is reported in the "Freq. rel." column. The following columns contain Precision, Recall and F1-score for each class.

Among the Information Types, "Affected individuals" achieves the best F1-score (0.868) whereas "Simpathy and emotional support" achieves the lowest (0.353). The "Informativeness" classifier achieves a F1-score of 0.935.

Tables V and VI contain evaluation on the multi-language, multi-hazard corpus, used as a test set (not examined by linguists during the development phase).

In Table V we report measurements only on "related to hazard" tweets (numbering 7,626), in the same fashion of the development set.

We found test results to be mostly in line with the development ones. The largest drop in performances was on the Infrastructure and Utilities class, which has an impact on Informativeness as well (Informativeness test F1-score 0.800). The list of relevant keywords and expressions could certainly be refined and benefit from further language and hazard specific tuning (also, proportions for these classes vary in a relevant manner when passing from development to test set, therefore from Floods to multiple hazards).

Performances on single languages are broken down and examined. In this case, the "Freq. rel" column contains relative frequency of Informative tweets within that language. Coherent with the fact that most of the development work was originally done on Italian tweets, we found that Precision is higher for English and Spanish, but Recall is lower (probable loss of colloquial expressions). F1-score for English is, however, found to be in line with the F1-score for Italian.

Table IV
CLASSIFICATION PERFORMANCE ON DEVELOPMENT SET: ITALIAN, FLOOD-RELATED TWEETS (700 ITEMS)

Class	Freq. rel.	Prec.	Rec.	F1
Affected individuals	0.146	0.836	0.902	0.868
Caution and advice	0.390	0.590	0.755	0.662
Donations and volunteering	0.031	0.875	0.318	0.467
Infrastructures and utilities	0.273	0.827	0.702	0.759
Sympathy and emo. support	0.013	0.375	0.333	0.353
Informativeness	0.914	0.943	0.927	0.935

Table V
CLASSIFICATION PERFORMANCE ON REDUCED TEST SET: ENGLISH/ITALIAN/SPANISH, HAZARD-RELATED TWEETS (7,626 ITEMS)

Class	Freq. rel.	Prec.	Rec.	F1
Affected individuals	0.181	0.838	0.889	0.863
Caution and advice	0.122	0.442	0.802	0.570
Donations	0.005	0.469	0.590	0.523
Infrastructures and utilities	0.127	0.609	0.507	0.553
Sympathy and emo. support	0.020	0.653	0.606	0.629
Volunteering	0.009	0.277	0.478	0.351
Informativeness	0.745	0.871	0.739	0.800
Informativeness (EN)	0.782	0.928	0.707	0.803
Informativeness (ES)	0.860	0.944	0.659	0.776
Informativeness (IT)	0.634	0.785	0.848	0.815
Informativeness (Earthquake)	0.766	0.902	0.724	0.803
Informativeness (Landslide)	0.945	0.968	0.869	0.915
Informativeness (Storms)	0.738	0.898	0.646	0.752
Informativeness (Wildfires)	0.801	0.873	0.752	0.808

We report also performances for Informativeness on single hazards, finding the classifier to generalize well on different hazards, with the major drop of performances within the Storm Event Type, which is also the Event Type presenting the lowest relative number of true Informative tweets ("Freq. rel" contains the proportion of Informative tweets within that Event Type).

In Table VI we report performances measured on the whole multi-language, multi-hazard corpus, to obtain a first estimate of what happens on "real-world" data-streams such as the ones we are monitoring. The classifiers can effectively filter out additional noise in the data: performances generally drops by about 0.02 points in terms of F1-score, in front of more than 2,000 unrelated tweets that were added (more than 25% of the reduced test set). F1-score on the Informativeness classification task is 0.776 on the full dataset .

V. CONCLUSION

The proposed approach looks viable for monitoring generic, emergency-related data streams from Twitter (and potentially other social media) and continuously extracting relevant information from them.

The annotated corpora can serve as an useful resource for

Table VI
CLASSIFICATION PERFORMANCE ON NOISY (FULL) TEST SET: ENGLISH/ITALIAN/SPANISH, TWEETS CONTAINING HAZARD-RELATED KEYWORDS (9,695 ITEMS)

Class	Freq. rel.	Prec.	Rec.	F1
Affected individuals	0.142	0.794	0.889	0.839
Caution and advice	0.096	0.421	0.802	0.552
Donations	0.004	0.451	0.590	0.511
Infrastructures and utilities	0.100	0.576	0.507	0.539
Sympathy and emo. support	0.016	0.595	0.606	0.601
Volunteering	0.007	0.258	0.478	0.335
Informativeness	0.586	0.816	0.739	0.776

this task, and will be made available for further research. Examining inter-annotator agreement gave us insights on which performances are realistic for automatic classification systems.

More work can be done in improving performances of Information Type classification and exploring the notion of Informativeness within hazard-related streams. We argue that the notion of Informativeness could be, in this generic context and for some applications, too subjective. If that is the case, other Information classes, such as Information Types and Geographical Information, implicit in the tweet content, could be of more and clearer help for end-users.

ACKNOWLEDGMENT

This work was partially funded by the EC through the I-REACT project (H2020-DRS-1-2015), grant agreement n.700256. We would like to thank: all the annotators that contributed to the creation of the corpora; Alessia Bianchini for her essential help in finding the right annotators; Andrea Bolioli for always prompting new research; Raffaella Ventaglio for her essential help in the preparation of this paper; all of our CELI colleagues for their daily contributions; our partners within the I-REACT project for the stimulating research we are conducting, and in particular Claudio Rossi for providing additional perspectives on employing social media data in emergency management.

REFERENCES

[1] C. Caragea et al., *Classifying Text Messages for the Haiti Earthquake*, Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011), 2011.

[2] C. Castillo, *Big crisis data: social media in disasters and time-critical situations*, Cambridge University Press, 2016.

[3] J. Cohen, *A coefficient of agreement for nominal scales*, Educational and psychological measurement, Volume 20, Number 1: Sage Publications Sage CA, 1960.

[4] L. Derczynski et al., *Analysis of named entity recognition and linking for tweets*, Information Processing & Management, Volume 51, Number 2: Elsevier, 2015.

[5] B. Klein et al., *Emergency Event Detection in Twitter Streams Based on Natural Language Processing*, Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction: 7th International Conference: Springer International Publishing, 2013.

[6] K. Leetaru et al., *Mapping the global Twitter heartbeat: The geography of Twitter*, First Monday, Volume 18, Number 5: 2013.

[7] B. Liu, *Sentiment analysis and opinion mining*, Synthesis lectures on human language technologies, Volume 5, Number 1: Morgan & Claypool Publishers, 2012.

[8] R.J. Mooney and R. Bunescu , *Mining knowledge from text using information extraction*, SCM SIGKDD explorations newsletter, Volume 7, Number 1: ACM, 2005.

[9] A. Olteanu et al., *What to Expect When the Unexpected Happens: Social Media Communications Across Crises*, Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing: ACM, 2015.

[10] T. Sakaki et al., *Earthquake shakes Twitter users: real-time event detection by social sensors*, Proceedings of the 19th international conference on World wide web: ACM, 2010.

[11] T. Simon et al., *Socializing in emergencies — A review of the use of social media in emergency situations*, International Journal of Information Management, Volume 35, Number 5: 2015.

[12] F. Tarasconi et al., *Geometric and statistical analysis of emotions and topics in corpora*, PIJCoL - Italian Journal of Computational Linguistics, Volume 1, Number 1: ACM, 2010.

[13] S. Vieweg et al., *Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: ccademia University Press, 2015.