A Linguistically-driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages

Conference Paper · May 2015

DOI: 10.1145/2740908.2741722

4 authors, including:

Stefano Cresci
Italian National Research Council
17 PUBLICATIONS **37** CITATIONS
SEE PROFILE

Maurizio Tesconi
Italian National Research Council
51 PUBLICATIONS **280** CITATIONS
SEE PROFILE

Felice Dell'Orletta
Italian National Research Council
45 PUBLICATIONS **163** CITATIONS
SEE PROFILE

A Linguistically-driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages

Stefano Cresci, Maurizio Tesconi
Institute for Informatics and Telematics (IIT)
National Research Council (CNR), Pisa, Italy
[name].[surname]@iit.cnr.it

Andrea Cimino, Felice Dell'Orletta
Institute for Computational Linguistics (ILC)
National Research Council (CNR), Pisa, Italy
[name].[surname]@ilc.cnr.it

ABSTRACT

This work focuses on the analysis of Italian social media messages for disaster management and aims at the detection of messages carrying critical information for the damage assessment task. A main novelty of this study consists in the focus on *out-domain* and *cross-event* damage detection, and on the investigation of the most relevant tweet-derived features for these tasks. We devised different experiments by resorting to a wide set of linguistic features qualifying the lexical and grammatical structure of a text as well as ad-hoc features specifically implemented for this task. We investigated the most effective features that allow to achieve the best results. A further result of this study is the construction of the first manually annotated Italian corpus of social media messages for damage assessment.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*; K.4.1 [**Computers and Society**]: Public Policy Issues—*Human safety*

Keywords

Damage assessment, feature selection, social sensing, social media mining, emergency management, crisis informatics

1. INTRODUCTION AND MOTIVATION

Nowadays, social media platforms such as Twitter, Weibo and Facebook, convey an unprecedented amount of information about the activities, interests and opinions of their users [22]. These platforms have become primary hubs for public expression and interaction because of their large user base, interactive nature and ease of use. In the last few years researchers have therefore turned their attention to the analysis of the information streams of social media services, thus enabling a new wave of experimentation feeding on the *digital traces* left by people's every day life interactions.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741722.

The amount of information shared on social media is even bigger in the aftermath of natural disasters and emergencies. Indeed in such situations, people usually share their experiences on these media which become rapidly overwhelmed by messages regarding the unfolding scenario. Thus, an emerging line of research is that of social media analysis for disaster response and management [10]. Preliminary results in this field have mainly focused on qualitative analyses and visualizations aimed at increasing the overall situational awareness during disasters. Situational awareness is generally achieved by displaying the content of disaster-related messages, for example via word-clouds [4]. Other approaches focus instead on high-level aggregate characteristics of the messages shared, such as the overall number and frequency of those messages [21]. However, a shift from qualitative situational awareness to in-depth quantitative damage assessments can provide even more valuable insights. In the aftermath of natural disasters one of the most urgent needs among decision makers is the acquisition of actionable information. Eyewitness messages reporting damages to the infrastructures or consequences on the population are arguably one of the most decisive sources of actionable information. Having access to a selected corpus of actionable messages would open new possibilities for emergency response, such as the real-time estimation of disaster impact (e.g. earthquake intensity estimation). Therefore, automatic systems mining the sheer amount of messages shared during and after disasters and capable of selecting the ones carrying damage information, could have a huge impact on emergency management procedures. However, acquiring a specific corpus of messages for every disaster to monitor is obviously time-expensive and may even result practically infeasible. Therefore, a fundamental challenge along this line of research is that of *cross-event* performance. The additional challenge is posed by the development of systems trained on a single corpus and able to achieve a good performance over a broad range of different events.

Our work aims to address these issues by building an annotated corpus of messages that is exploited to train a system for the detection of those messages carrying damage information. Although retaining a language-independent approach, we built the first corpus of social media messages for damage assessment in Italian. This corpus has been manually annotated into 3 different classes of messages: those carrying damage information, those without any damage information but still relevant to the disaster, and those that are not rele-

vant. Furthermore, to the best of our knowledge, this is the first work investigating *cross-event* damage assessment.

The remainder of this paper is organized as follows: Section 2 describes the related work and Section 3 discloses the details of our dataset. Section 4 describes our approach to the damage assessment task. Section 5 presents the experiments carried out and the results of our study while Section 6 draws the conclusions and defines future works.

2. RELATED WORK

Works such as [7] [9] highlighted how disasters and emergencies cause changes in human behaviors that are clearly distinguishable from social media activity. Therefore, sudden changes in social media activity can be exploited for detecting and monitoring such disasters. These early studies showed the potential for the exploitation of such information and paved the way for the for the body of work described in the remainder of this section. Along this line of research current efforts are mainly focused on the event detection and situational awareness tasks. Social media-based systems for earthquake detection are discussed in [19] [8] [2, 3]. In these works the detection is triggered by an exceptional growth in the frequency of messages containing earthquake-related keywords. Authors of [3] also proposed some first solutions towards the automatic assessment of the consequences of earthquakes by mining the content of social media messages. Similar studies have been performed in [4] where authors describe the ESA system for emergency events detection and monitoring. The goal of ESA is to increase situational awareness during disasters by displaying word-clouds of relevant messages after the detection of an emergency. Other systems for increasing situational awareness are described in [18] and [12]. The former presents the Twitris system, which is capable of collecting, aggregating and analyzing data to give deeper insights and facilitate coordination and action during emergency events. The latter describes the AIDR system, which exploits a combination of automatic analyses and online human annotations to classify disaster-related messages in a real-time fashion.

Recently [11] presents a survey on computational techniques for social media data processing in emergencies and can be considered for additional references about works in the field of social media disaster management. Among these works, the most relevant ones for our approach to the damage assessment task are [20, 13]. In [20] is shown how Natural Language Processing techniques contribute to the detection of Twitter messages (henceforth *tweets*) carrying information that is relevant for situational awareness during mass emergencies. The work in [13] focuses on extracting "information nuggets" from tweets, i.e. self-contained information items relevant to disaster response. Similarly to these works we rely on Natural Language Processing techniques but we employ features taken from more sophisticated levels of automatic linguistic annotation. Besides, to our knowledge this is the first time that features typically used in sentiment analysis and features extracted from similarity lexicons automatically created using word embeddings have been used and proven effective for this task. A peculiarity of our work consists in the focus on *out-domain* and *cross-event* damage detection. These tasks are inherently harder than *in-domain* detections and consist in training a system on a natural disaster and testing it on different disasters. Therefore, we investigate which tweet-derived features are most relevant for these two tasks. In addition, even if our approach is language independent, to our knowledge this is the first study on damage assessment carried out for the Italian language.

3. DATASET

The datasets exploited for our experiments are composed of Italian tweets, collected in the aftermath of several natural disasters. We exploited the Twitter Streaming API[1] for Twitter data acquisition about recent disasters, and we bought data from Twitter resellers[2] for past disasters. Both Twitter's Streaming API and resellers' Historical APIs, give access to a global set of tweets, optionally filtered by search keywords. We exploited a different set of search keywords for every different disaster in order to collect the most relevant tweets about it. Whenever possible, we exploited *#hashtags* specifically created to share reports of a particular disaster, such as the *#allertameteoSAR* hashtag for the severe flood that struck Sardegna regional district in November 2013. In this way we were able to select only tweets actually related to that disaster. For historical disasters however, we couldn't rely on specific hashtags and had to exploit more generic search keywords such as "terremoto" (*earthquake*) and "scossa" (*tremor*) for the historical earthquake of L'Aquila in 2009. For our experiments we considered two different kinds of disasters: earthquakes and floods, both recent and historical. To investigate a broad range of situations we also picked disasters having variable degrees of severity: some caused only moderate damages, while other produced widespread damages and tens of deaths. Table 1 shows the characteristics of the natural disasters considered in our study as well as the size of the related datasets.

		Tweets$^\sharp$			
Dataset	**Type**	*dmg*	*no dmg*	*not rel*	*TOT*
Sardegna	Flood	717	194	65	976
L'Aquila	Earthquake	312	480	270	1,062
Emilia	Earthquake	507	2,141	522	3,170
Genova	Flood	187	201	46	434

Table 1: Characteristics of our datasets. ($^\sharp$) *dmg*: tweets of the *damage* class; *no dmg*: tweets of the *no damage* class; *not rel*: tweets of the *not relevant* class; *TOT*: total number of tweets.

All the tweets contained in the datasets of Table 1 have been exploited for the training and testing of our system. Tweets have been manually annotated by three human annotators who employed a web-based tweet annotation tool. With regards to the damage assessment task, tweets have been annotated as in the following, according to the kind of information they convey: (i) tweets related to the disaster and carrying information about damages to the infrastructures or on the population (*damage*); (ii) tweets related to the disaster but not carrying relevant information for the assessment of damages (*no damage*); (iii) tweets not related to the disaster (*not relevant*). The inclusion of a class for tweets that are not related to a disaster (*not relevant*) is necessary because the automatic data collection strategy we adopted does not guarantee that all the tweets collected are actually

[1] https://dev.twitter.com/docs/api/streaming
[2] http://gnip.com/sources/twitter/historical

related to the disaster under investigation. This is especially true for the datasets collected with generic search keywords and represents an added challenge for the classification task. The annotated dataset will be made freely available at the following website: *http://socialsensing.eu/datasets*.

4. SYSTEM DESCRIPTION

Our approach to damage assessment was implemented in a software prototype, i.e. a classifier operating on morpho-syntactically tagged and dependency parsed texts, which assigns to each tweet a score expressing its probability of belonging to a given class: *damage*, *no damage*, *not relevant*. The highest score represents the most probable class. Given a set of features and a training corpus, the classifier creates a statistical model using the feature statistics extracted from the training corpus. This model is used in the classification of unseen tweets. The set of features and the machine learning algorithm can be parameterized through a configuration file. For this work, we used linear Support Vector Machines (SVM) using LIBSVM as the machine learning algorithm. Since our approach relies on multi–level linguistic analysis, both training and test data were automatically morpho-syntactically tagged by the POS tagger described in [6] and dependency-parsed by the DeSR parser using Multi-Layer Perceptron as the learning algorithm [1], a state-of-the-art linear-time Shift-Reduce dependency parser.

In addition, we developed sentiment polarity and similarity lexicons to improve the overall accuracy of our system. These lexicons were exploited in the lexical expansion and sentiment polarity features used by the classifier. The lexical expansion features were used to overcome the problem of the lexical sparsity in tweets, due to their short length in terms of words. The sentiment polarity features are commonly used to infer the polarity of a piece of text, in this work we aim to verify whether these can be useful for the damage assessment task.

4.1 Lexicons

Sentiment Polarity Lexicons. Sentiment polarity lexicons provide mappings between a word and its sentiment polarity (positive, negative, neutral). For our experiments, we used a publicly available lexicon. In addition, we adopted an unsupervised method to automatically create a lexicon specific for the Italian twitter language.

Existing Sentiment Polarity Lexicons. We used the Italian sentiment polarity lexicon [15] developed within the OpeNER European project[3]. This is a freely available lexicon for the Italian language[4] and includes 24,000 Italian word entries. It was automatically created using a propagation algorithm and manually reviewed for the most frequent words.

Automatically created Sentiment Polarity Lexicons. We built a corpus of positive and negative tweets following the [17] approach adopted in the Semeval 2013 sentiment polarity detection task. For this purpose we queried the Twitter API with a set of hashtag seeds that indicate positive and negative sentiment polarity. We selected 200 positive word seeds (e.g. "vincere" *to win*, "splendido" *splendid*, "affascinante" *fascinating*), and 200 negative word seeds (e.g. "tradire" *to betray*, "morire" *to die*). These terms were chosen from the OpeNER lexicon. The resulting corpus is made up of 683,811 tweets extracted with positive seeds and 1,079,070 tweets extracted with negative seeds [5]. The main purpose of this procedure was to assign a polarity score to each n-gram occurring in the corpus. For each n-gram (we considered up to five n-grams) we calculated the corresponding sentiment polarity score with the following scoring function: $score(ng) = PMI(ng, pos) - PMI(ng, neg)$, where PMI stands for pointwise mutual information. A positive or negative score indicates that the n-gram is relevant for the identification of positive or negative tweets.

Word Similarity Lexicons. We trained two predict models using the word2vec[5] toolkit [16]. As recommended in [16], we used the CBOW model that learns to predict the word in the middle of a symmetric window based on the sum of the vector representations of the words in the window (we considered a context window of 5 words). These models learn lower-dimensional word embeddings. Embeddings are represented by a set of latent (hidden) variables, and each word is a multidimensional vector that represent a specific instantiation of these variables. We built the word similarity lexicons by applying the cosine similarity function between the embedded words. Starting from two corpora, we developed two different similarity lexicons. The first lexicon was built using the lemmatized version of the PAISÀ[6] corpus [14]. PAISÀ is a freely available large corpus of authentic contemporary Italian texts from the web, and contains approximately 388,000 documents for a total of about 250 million tokens. The second lexicon was built from a lemmatized corpus of tweets. This corpus was collected starting from 30 generic seed keywords used to query Twitter APIs. The resulting corpus is made up of 1,200,000 tweets. These tweets were automatically morpho-syntactically tagged and lemmatized by the POS tagger described in [6].

4.2 Features

We focused on a wide set of features ranging across different levels of linguistic description. The whole set of features is organized into five main categories: *raw and lexical text features*, *morpho-syntactic features*, *syntactic features*, *lexical expansion features* and *sentiment analysis features*. This proposed partition closely follows the different levels of linguistic analysis automatically carried out on the text being evaluated, (i.e. tokenization, lemmatization, morpho-syntactic tagging and dependency parsing) and the use of external lexical resources.

Raw and Lexical Text Features. *Number of tokens*: number of blocks consisting of 5 tokens occurring in the analyzed tweet. *Character n-grams*: presence or absence of contiguous sequences of characters. *Word n-grams*: presence or absence of contiguous sequences of tokens. *Lemma n-grams*: presence or absence of contiguous sequences of lemma occurring in the tweet. *Repetition of n-grams chars*: presence or absence of contiguous repetition of characters. *@ number*: number of @ occurring in the tweet. *Hashtags number*: number of hashtags. *Finishes with punctuation*: checks whether the tweet finishes with one of the following punctuation characters: "?", "!".

Morpho-syntactic Features. *Coarse grained part-of-speech n-grams*: presence or absence of contiguous sequences

[3]http://www.opener-project.eu/
[4]https://github.com/opener-project/public-sentiment-lexicons
[5]http://code.google.com/p/word2vec/
[6]http://www.corpusitaliano.it/

		no damage			damage			not relevant		
Model	Accuracy	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
Sardegna *(cross–event)*										
MorphoSyntax	41.80	42.18	**13.91**	**20.93**	94.25	45.74	61.59	9.39	81.53	16.85
Syntax	39.34	**44.89**	11.34	18.10	93.29	42.67	58.56	9.34	86.15	16.86
LexicalExpansion	46.51	39.47	7.73	12.93	**94.76**	52.99	**67.97**	**10.98**	**90.76**	**19.60**
Global	**46.61**	39.53	8.76	14.34	93.38	**53.13**	67.73	10.85	87.69	19.32
L'Aquila *(in–domain)*										
MorphoSyntax	80.79	78.35	86.48	82.17	89.67	83.42	86.21	74.68	67.01	70.38
Syntax	80.51	78.18	86.30	81.99	88.80	83.29	85.66	75.07	66.54	70.07
LexicalExpansion	82.67	80.43	87.22	83.66	**91.82**	86.27	88.80	**76.42**	69.83	72.78
Global	**82.86**	**80.83**	**87.25**	**83.85**	91.69	**87.14**	**89.14**	76.41	**70.20**	**73.09**
Emilia *(out–domain)*										
MorphoSyntax	73.97	75.06	**94.68**	83.73	87.75	60.56	71.66	27.08	7.38	11.60
Syntax	73.46	74.91	94.37	83.52	86.45	58.45	69.74	25.49	7.38	11.45
LexicalExpansion	**74.89**	76.77	93.92	**84.48**	89.79	61.97	73.33	**34.24**	14.20	20.08
Global	74.59	**76.81**	93.16	84.20	**91.75**	**62.67**	**74.47**	32.09	**14.77**	**20.23**
Genova *(cross–event)*										
MorphoSyntax	22.44	75.86	11	19.21	90.90	10.30	18.51	14.95	93.84	25.79
Syntax	21.35	**76.19**	8	14.48	95	9.79	17.75	15.07	**96.92**	26.08
LexicalExpansion	28.10	68.57	12	20.42	93.33	21.64	35.14	**16.62**	**96.92**	**28.37**
Global	**28.32**	70.27	**13**	**21.94**	**93.47**	**22.16**	**35.83**	16.22	93.84	27.66

Table 2: Experiments performed using the L'Aquila dataset as training set.

of coarse-grained Part-of-speech. *Fine grained part-of-speech n-grams*: presence or absence of contiguous sequences of fine-grained PoS, which represent subdivisions of the coarse-grained tags (e.g. nouns are subdivided into proper vs common nouns, verbs into main verbs, gerund forms, past particles, etc.). *Coarse grained part-of-speech distribution*: the distribution of nouns, adjectives, adverbs, numbers. **Syntactic Features.** *Dependency types n-grams*: presence or absence of sequences of dependency types in the tweet. The dependencies are calculated with respect to the surface linear ordering of words. *Lexical dependency n-grams*: presence or absence of sequences of lemmas calculated with respect to the hierarchical parse tree. *Coarse grained part-of-speech dependency n-grams*: presence or absence of sequences of coarse-grained part–of–speech calculated with respect to the hierarchical parse tree.

Lexical expansion Features. *Lexical expansion*: for each lemma of the tweet, the feature increases the tweet lexicon with the first 15 similar lemmas occurring in the similarity lexicons.

Sentiment analysis features. *Emoticons*: presence or absence of positive or negative emoticons in the tweet[7]. *Lemma sentiment polarity n-grams*: for each lemma n-grams in the tweet, the feature checks the polarity of each component lemma in sentiment polarity lexicon. Lemmas that are not present are marked with the *ABSENT* tag (e.g. the trigram "tutto molto bello" (*all very nice*) is marked as "*ABSENT-POS-POS*" because *molto* (*very*) and *bello* (*nice*) are marked as positive in the considered polarity lexicon and *tutto* is absent). *Polarity modifier*: for each lemma in the tweet occurring in the existing sentiment polarity lexicons, the feature checks the presence of adjectives or adverbs in a left context window of size 2. If this is the case, the polarity of the lemma is assigned to the modifier (e.g. in the bigram "non interessante" (*not interesting*), "interessante" is a positive word, and "non" is an adverb. Accordingly, the feature "non_POS" is created). *PMI score*: for each set of unigrams, bigrams, trigrams, four-grams and five-grams that occur in the tweet, the feature computes the score given by $\sum_{\text{i–}gram \in tweet} score(\text{i–}gram)$ and returns the minimum and the maximum values of the five values (approximated to the nearest integer). *Distribution of sentiment polarity*: this feature computes the percentage of positive, negative and neutral lemmas that occur in the tweet (the percentages are rounded to the nearest multiple of 5). The feature is computed for each existing lexicon. *Most frequent sentiment polarity*: the feature returns the most frequent sentiment polarity of the lemmas occurring in the tweet. *Sentiment polarity through lexical expansion*: for each lemma of the tweet, the feature extracts the first 15 similar words occurring in the similarity lexicons. For each similar lemma, the feature checks the presence of negative or positive polarity. In addition, the feature calculates the most frequent polarity. Since we have two different similarity lexicons and one sentiment lexicon, the feature is computed 2 times. *Sentiment polarity in tweet sections*: the feature first splits the tweet in three equal sections. For each section the most frequent polarity is computed using the available sentiment polarity lexicons. The purpose of this feature is aimed at identifying change of polarity within the same tweet.

[7]The lexicon of emoticons was extracted from http://it.wikipedia.org/wiki/Emoticon and manually classified.

5. EXPERIMENTS AND RESULTS

Three different sets of experiments were devised to test the performance of our system. In the first experiment, we trained and tested the classifier on the same class of natural disaster in the same locality (*in–domain* experiment). In the second experiment, we trained the classifier on a natural disaster and we tested it on the same natural disaster type but occurred in a different place (*out–domain* experiment). In the last experiment, we trained the classifier on a natural disaster (e.g. flood) and tested on a different disaster (e.g. earthquake) (*cross–event* experiment).

In order to evaluate the performance of our classifier with respect to the features defined in section 4.2, we ran our experiments using 4 different classification models: the **Morpho-syntax model** using raw, lexical and morpho–syntatic features; the **Syntax model** using raw, lexical, morpho–syntatic and syntax features; the **Lexical expansion model** using

		no damage			damage			not relevant		
Model	Accuracy	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
Sardegna *(in–domain)*										
MorphoSyntax	78.47	53.09	45.79	48.02	85.11	93.08	88.83	**35.00**	**15.33**	**20.50**
Syntax	**78.68**	**53.48**	**46.92**	**48.67**	85.55	**93.26**	**89.14**	32.33	13.66	18.47
LexicalExpansion	77.96	50.46	45.82	47.07	85.84	92.35	88.91	30.5	14.7	19.15
Global	77.86	49.27	45.69	46.64	**86.19**	92.34	89.10	28.16	13.66	17.34
L'Aquila *(cross–event)*										
MorphoSyntax	39.17	58.63	26.87	36.85	34.08	91.66	49.69	33.33	0.37	0.73
Syntax	38.13	54.10	23.33	32.60	34.11	**93.26**	49.95	10.00	0.74	1.47
LexicalExpansion	42.37	**56.35**	**34.16**	**42.54**	36.91	90.35	52.41	57.14	1.48	2.88
Global	**42.75**	56.33	33.33	41.88	**37.45**	91.34	**53.12**	**52.94**	**3.33**	**6.27**
Emilia *(cross–event)*										
MorphoSyntax	28.58	61.78	26.29	36.88	15.23	74.64	25.29	0	0	0
Syntax	28.99	62.15	27.20	37.84	15.11	73.23	25.06	0	0	0
LexicalExpansion	33.06	64.11	**33.13**	43.68	17.11	76.05	27.94	40	1.13	2.21
Global	**34.01**	**65.75**	32.97	**43.92**	**17.42**	**78.16**	**28.49**	**44.44**	**2.27**	**4.32**
Genova *(out–domain)*										
MorphoSyntax	50.54	54.05	50	51.94	49.25	68.04	57.14	0	0	0
Syntax	49.25	**92.34**	50.79	**65.53**	8.51	59.18	14.89	0	0	0
LexicalExpansion	52.50	60	**51**	55.13	49.28	71.13	58.22	**11.11**	**1.53**	**2.70**
Global	**53.81**	62.19	**51**	56.04	**50.70**	**74.22**	**60.25**	9.09	**1.53**	2.63

Table 3: Experiments performed using the Sardegna dataset as training set.

raw, lexical, morpho–syntatic, syntax and lexical expansion features; the **Global model**, including all the features of the previous model combined with the sentiment features.

We balanced the training and test sets by randomly selecting 976 tweets from the Emilia dataset (see table 1). We obtained a dataset containing about the same number of tweets for each natural disaster. Moreover, due to it's small size, the Genova tweet collection was used only as a test set.

5.1 Evaluation methodology

The three experiments were evaluated in terms of i) overall Accuracy of the system and ii) Precision, Recall and F-score. Accuracy is a global score referring to the percentage of tweets correctly classified. Precision, Recall and F-score have been computed with respect to the defined classes. Precision is the ratio of the number of correctly classified tweets over the total number of tweets classified as belonging to a particular class; Recall has been computed as the ratio of the number of correctly classified tweets over the total number of tweets belonging to a particular class in the test sets. F-score is the harmonic mean of Precision and Recall. For each set of experiments, evaluation was carried out with respect to the four models of the classifier. To evaluate the *in-domain* performance of each model, we followed a 10-fold cross validation process: each dataset was randomly split in 10 different non overlapping training and test sets. The Accuracy, Precision, Recall and F-score were calculated as the average of the these values over all the 10 test sets.

5.2 Results

Tables 2, 3 and 4 report the accuracies achieved by the different classifier models. Each table reports the score achieved by different classifier models using the same natural disaster event as training set and tested on all the natural disasters corpora. The highest score for each feature model and for each evaluation metric is reported in bold font. Particularly interesting for this work are the scores obtained in the classification of the *damage* class.

Concerning the *in-domain* experiments, the best accuracies are achieved using a small set of basic features (i.e. *Morpho-syntax model*) (with the exception of the experiments conducted using the L'Aquila dataset as training set). This is in agreement with the experiments performed by [20], that demonstrated that a classifier based on low-level linguistic features performs well at identifying tweets that contribute to *situational awareness*. But this is not the case of the *out-domain* and *cross-event* experiments. For these experiments the *LexicalExpansion* and the *Global* models outperform the results obtained by the *Morpho-syntax* model by several percentage points, both in terms of overall Accuracy and F-scores, especially for the damage class. These results show the usefulness of our features, with emphasis on the effectiveness of the lexical expansion ones, for both *out-domain* and *cross-event* experiments. Finally, we checked whether using a training set created by merging the two different natural disaster events can positively affect the Accuracy of our system in the *out-domain* scenario. We created two new datasets: the first contains the *Emilia* and the *Sardegna* datasets, the second contains the *L'Aquila* and the *Sardegna* datasets. Since in the previous *out-domain* experiments the *Global* model outperformed the other ones, we considered only this feature model. The model obtained using the *Emilia+Sardegna* training set had an improvement on the model trained using only the *Emilia* training set of +2.5% as far as the *L'Aquila* test set is concerned. On the contrary, it didn't improve when tested on the *Genova* test set. We obtained similar results for the model obtained using the *L'Aquila+Sardegna* training set. This model improved over the model using only the *L'Aquila* training set when tested on the *Emilia* test set, but decreased the performance on the *Genova* test set.

6. CONCLUSIONS

In this work we discussed a system for the assessment of damages from social media messages, with a focus on the challenge of *cross-event* performance. To train and test our system we built the first Italian annotated corpus of messages for damage assessment. We demonstrated that the exploitation of advanced linguistic features, not employed in past works, yield better results especially in the *out-domain* and *cross-event* scenarios. Our results also show the con-

		no damage			damage			not relevant		
Model	Accuracy	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
Sardegna *(cross–event)*										
MorphoSyntax	63.42	33.03	**37.62**	**35.18**	83.57	73.08	77.97	17.18	33.84	22.79
Syntax	61.27	29.06	30.41	29.72	83.88	71.13	76.98	17.57	44.61	25.21
LexicalExpansion	69.36	33.33	24.22	28.06	**85.44**	83.54	84.48	23.13	47.69	31.15
Global	**69.98**	**35.55**	24.74	29.17	85.14	**83.96**	**84.55**	**24.62**	**50.76**	**33.16**
L'Aquila *(out–domain)*										
MorphoSyntax	53.48	55.40	57.70	56.53	**71.56**	**70.19**	**70.87**	28.12	26.66	27.37
Syntax	54.23	57.35	56.87	57.11	68.86	**70.19**	69.52	31.34	31.11	31.22
LexicalExpansion	**55.93**	59.22	**60.20**	**59.71**	70.80	67.62	69.18	**34.05**	34.81	34.43
Global	55.74	**59.36**	58.12	58.73	70.45	69.55	70.00	33.80	**35.55**	**34.65**
Emilia *(in–domain)*										
MorphoSyntax	**81.25**	83.52	**90.72**	**86.90**	**90.22**	**85.17**	**87.33**	**57.16**	42.17	47.64
Syntax	81.04	83.65	90.39	86.83	89.52	83.31	85.81	56.28	43.16	48.46
LexicalExpansion	80.32	83.33	89.94	86.44	89.95	79.59	84.09	55.13	44.72	48.55
Global	80.32	**83.76**	89.46	86.43	88.68	79.63	83.50	54.56	**46.71**	**49.56**
Genova *(cross–event)*										
MorphoSyntax	**42.70**	46.10	**38.5**	**41.96**	87.25	45.87	60.13	15.78	46.15	23.52
Syntax	38.34	40.88	32.5	36.21	83.01	45.36	58.66	11.85	35.38	17.76
LexicalExpansion	42.48	**49.54**	27	34.95	87.15	**48.96**	62.70	**19.08**	**70.76**	**30.06**
Global	42.04	45.76	27	33.96	**90.38**	48.45	**63.08**	18.98	69.23	29.80

Table 4: Experiments performed using the Emilia dataset as training set.

tribution and strength of sentiment analysis and lexical expansion features. Our work, other than being interesting on its own, also opens new opportunities for emergency management. Results of our system can be further exploited by statistical models for the estimation of disaster impact.

7. REFERENCES

[1] G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. Accurate dependency parsing with a stacked multilayer perceptron. In *Evalita'09*.

[2] M. Avvenuti, S. Cresci, M. La Polla, A. Marchetti, and M. Tesconi. Earthquake emergency management by social sensing. In *PERCOM'14 Workshops*. IEEE.

[3] M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi. Ears (earthquake alert and report system): a real time decision support system for earthquake crisis management. In *KDD'14*. ACM.

[4] M. A. Cameron, R. Power, B. Robinson, and J. Yin. Emergency situation awareness from twitter for crisis management. In *WWW'12 companion*. ACM.

[5] A. Cimino, S. Cresci, F. Dell'Orletta, and M. Tesconi. Linguistically-motivated and lexicon features for sentiment analysis of italian tweets. In *Evalita'14*.

[6] F. Dell'Orletta. Ensemble system for part-of-speech tagging. In *Evalita'09*.

[7] P. Earle. Earthquake twitter. *Nature Geoscience*, 2010.

[8] P. S. Earle, D. C. Bowden, and M. Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 2012.

[9] L. Gao, C. Song, Z. Gao, A.-L. Barabási, J. P. Bagrow, and D. Wang. Quantifying information flow during emergencies. *Scientific reports*, 2014.

[10] A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *Int. Journal of Emergency Management*, 2009.

[11] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *arXiv preprint'14*.

[12] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg. Aidr: Artificial intelligence for disaster response. In *WWW'14 companion*. ACM.

[13] M. Imran, S. M. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Extracting information nuggets from disaster-related messages in social media. *ISCRAM'13*.

[14] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, and V. Pirrelli. The *PAISÀ* corpus of italian web texts. In *WAC'13*.

[15] I. Maks, R. Izquierdo, F. Frontini, R. Agerri, P. Vossen, and andoni Azpeitia. Generating polarity lexicons with wordnet propagation in 5 languages. In *LREC'14*. ELRA.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, 2013.

[17] S. M. Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval'13*.

[18] H. Purohit and A. P. Sheth. Twitris v3: From citizen sensing to analysis, coordination and action. In *ICWSM'13*.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW'10*. ACM.

[20] S. Verma, S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson. Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In *ICWSM'11*.

[21] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *CHI'10*. ACM.

[22] A. Zhou, W. Qian, and H. Ma. Social media data analysis for revealing collective behaviors. In *KDD'12*. ACM.