Classifying Crises-Information Relevancy with Semantics

Conference or Workshop Item

Classifying Crises-Information Relevancy with Semantics

Prashant Khare, Grégoire Burel, and Harith Alani

Knowledge Media Institute, The Open University, United Kingdom
{prashant.khare,g.burel,h.alani}@open.ac.uk

Abstract. Social media platforms have become key portals for sharing and consuming information during crisis situations. However, humanitarian organisations and effected communities often struggle to sieve through the large volumes of data that are typically shared on such platforms during crises to determine which posts are truly relevant to the crisis, and which are not. Previous work on automatically classifying crisis information was mostly focused on using statistical features. However, such approaches tend to be inappropriate when processing data on a type of crisis that the model was not trained on, such as processing information about a *train crash*, whereas the classifier was trained on *floods, earthquakes, and typhoons*. In such cases, the model will need to be retrained, which is costly and time-consuming. In this paper, we explore the impact of semantics in classifying Twitter posts across same, and different, types of crises. We experiment with 26 crisis events, using a hybrid system that combines statistical features with various semantic features extracted from external knowledge bases. We show that adding semantic features has no noticeable benefit over statistical features when classifying same-type crises, whereas it enhances the classifier performance by up to 7.2% when classifying information about a new type of crisis.

Keywords: semantics, crisis informatics, tweet classification

1 Introduction

The 2017 World Humanitarian Data and Trends report by UNOCHA[1] indicated that in 2016 alone, there were 324 natural disaster, affecting 204 million people, from 105 countries, causing an overall damage cost of $147 billion. During the course of natural disasters, large amounts of content are typically published in real time on various social media outlets. For instance, over 20 million tweets with the words #sandy and #hurricane were posted in just a few days during the Hurricane Sandy disaster.[2]

Although these messages act as critical information sources for various communities and relief teams, the sheer volume of data generated on social media platforms during crises makes it extremely difficult to manually process such

[1] UNOCHA, https://data.humdata.org/dataset/world-humanitarian-data-and-trends.
[2] Mashable: Sandy Sparks 20 Million Tweets, http://mashable.com/2012/11/02/hurricane-sandy-twitter.

streams in order to filter relevant pieces of information quickly[7]. Automatically identifying crisis-information relevancy is not trivial, especially given the characteristics of social media posts such as colloquialisms, short post length, non-standard acronyms, and syntactic variations in the text. Furthermore, many posts that carry the crisis hashtag/s can be irrelevant, hence hashtags are inadequate filters of relevancy.

Various works explored classification methods of crisis-data from social media platforms, to automatically categorise them into *crisis-related* or *not related*. These classification methods include both supervised[14,11,20,25] and unsupervised[18] machine learning approaches. Most of these methods are based on statistical features of the text, such as n-grams, text length, POS, and Hashtags. Although statistical models have shown to be efficient in classifying relevancy of crisis-information, their accuracy naturally drops when applied to information that were not included in the training sets. The typical approach to remedy this problem, is to retrain the model on new datasets or apply complex domain adaptation techniques, which are costly and time consuming, and thus are inadequate for crisis situations which typically require immediate reaction.

This work aims to bridge this gap by adding semantic features for the identification of crisis-related tweets on seen and unseen crises types. We hypothesise that adding concepts and properties (e.g., *type, label, category*) improves the identification of crisis information content across crisis domains, by creating a non-specific crisis contextual semantic abstraction of crisis-related content. The main contributions of this paper can be summarised as follow:

1. Build a statistical-semantic classification model with semantics extracted from BableNet and DBpedia.
2. Experiment with classifying relevancy of tweets from 26 crisis events of various types and in multiple languages.
3. Run relevancy classifiers with multiple feature combinations and when crisis types are included/excluded from training data.
4. Show that adding semantics increase of classification accuracy on unseen crisis types by +7.2% in F1 in comparison to non-semantic models.

The paper is structured as follows: Section 2 summarises related work. Sections 3 and 4 describe our approach and experiments on classifying relevancy while using different semantic features and crisis datasets. Results are reported in Sections 4.2 and 4.3. Discussion and conclusions are in Section 5 and 6.

2 Related Work

Large volumes of messages are typically posted across different social media platforms during crisis situations. However, a considerable number of these messages are potentially not related and irrelevant. Olteanu et al. [16] made an observation about the broad categories that crisis reports from social media can be categorised into: *related and informative*, *related but not informative*, and *not related*.

Identifying crisis related content from social media is not a new research area. Most supervised machine learning approaches used in this domain rely on linguistic and other statistical attributes of the post such as part of speech (POS), user mentions, length of the post, and number of hashtags. Supervised machine learning approaches range from traditional classification methods such as Support Vector Machines (SVM), Naive Bayes, Conditional Random Fields [20,17,8] to recent trends of deep learning [3]. In [3,4], word embeddings are applied and semantics are added in the form of extracted entities and their types, but adaptability of the model to unseen types of crisis data is not evaluated.

Complex domain adaptation methods has found its application in the areas of text classification and sentiment analysis [6], but have not been applied to crisis situations. In crisis classification, a closely related work [8] took a step towards domain adaptation by considering crisis data from two disasters, Joplin 2011 tornado and Hurricane Sandy. They trained the model on a part of Joplin tornado, and tested it on Hurricane Sandy and remaining part of Joplin data. However, their work was limited to only two crises; one hurricane and one tornado, which often cast similar types of impact on human life and infrastructure. Additionally, the semantic aspect of the crisis was not taken into consideration, which could have potentially highlighted the applicability of the method in multiple crisis scenario.

Unsupervised methods were also explored, often based on clustering [18] and keyword based processing. Our work in this paper complements and extends the aforementioned studies by investigating the use of semantics, derived from knowledge graphs, such as entities occurring in the tweets, and expanding them to their hypernyms and extended information through DBpedia properties.

Previously, we used hierarchical semantics from knowledge graphs to perform crisis-information classification through a supervised machine learning approach [12]. However, the study was limited to 9 crisis events, and confined to training and testing on the same type of crisis-events (i.e., no cross-crisis evaluation).

Some systems were developed that use semantics extracted with Named Entity Recognition tools on DBpedia and WordNet, to support searching of crisis-related information (e.g., Twitcident [2], Armatweet [23]). These system are focused on search, and do not include machine learning classifiers.

As opposed to previous work, we focus on applying these classifiers to two particular cases. First, when the classification model was trained on the data that contained crisis-event type, and secondly, when the crisis event type was not included in the training set. These two cases are aimed to help us better understand if, and when, adding semantics outperforms purely statistical approaches.

3 Semantic Classification of Crisis-related Content

The automatic identification of crisis-related content on social media requires the training and validation of a binary text classifier that is able to distinguish between *crisis-related* and *not related* crisis content. In this paper, we focus on generating statistical and semantic features of tweets and then training different

machine learning models. In the following sections, we present (i) the dataset used for training our classifiers, (ii) the statistical and semantic set of features used for building the classifiers, and (iii) the classifier selection process.

3.1 Dataset and Data Selection

In this study, we use the CrisisLexT26[3] dataset [16]. It contains annotated datasets of 26 different crisis events, which occurred between 2012 and 2013, with 1000 labeled tweets (*'Related and Informative', 'Related but not Informative', 'Not Related' and 'Not Applicable'*) for each event. The search keywords used to collect the original data used hashtags and/or terms that are often paired with the canonical forms of a disaster name and the impacted location (e.g., Queensland floods) or meteorological terms (e.g., Hurricane Sandy). We selected all 26 events, and for each event we combined the *Related and Informative* and *Related but not Informative* into the *Related* class, and combined the *Not Related* and *Not Applicable* into the *Not Related* class. These two classes are then used for distinguishing crisis-related content from unrelated content for creating binary text classifiers.

To reduce content redundancy in the data, we removed replicated instances from the collection of individual events by comparing tweets pairs after removing user-handles (i.e., '@' mentions), URL's, and special characters. This resulted in 21378 documents annotated with the *Related* label and 2965 annotated with the *Not Related* label. For avoiding classification bias towards the majority class, we balanced the data from each event by matching the number of *Related* documents with the *Not Related* ones. This was achieved by randomly selecting the same number of *Related* and *Not Related* tweets in any given event. This resulted in a final overall size of 5931 tweets (2966 *Related* and 2965 *Not Related* documents). Table 1 shows the distribution of selected tweets for each event.

Table 1. Crisis events data, balanced between related and not-related classes

			Category						Category		
Nb.	Id	Event	Related	Not-Related	Total	Nb.	Id	Event	Related	Not-Related	Total
1	CWF	Colorado Wildfire	242	242	484	2	COS	Costa Rica Earthquake	470	470	940
3	GAU	Guatemala Earthquake	103	103	206	4	ITL	Italy Earthquake	56	56	112
5	PHF	Philippines Flood	70	70	140	6	TYP	Typhoon Pablo	88	88	176
7	VNZ	Venezuela Refinery	60	60	120	8	ALB	Alberta Flood	16	16	32
9	ABF	Australia Bushfire	183	183	366	10	BOL	Bohol Earthquake	31	31	62
11	BOB	Boston Bombing	69	69	138	12	BRZ	Brazil Nightclub Fire	44	44	88
13	CFL	Colorado Floods	61	61	122	14	GLW	Glasgow Helicopter Crash	110	110	220
15	LAX	LA Airport Shoot	112	112	224	16	LAM	Lac Megantic Train Crash	34	34	68
17	MNL	Manila Flood	74	74	148	18	NYT	NY Train Crash	2	1	3
19	QFL	Queensland Flood	278	278	556	20	RUS	Russia Meteor	241	241	482
21	SAR	Sardinia Flood	67	67	134	22	SVR	Savar Building	305	305	610
23	SGR	Singapore Haze	54	54	108	24	SPT	Spain Train Crash	8	8	16
25	TPY	Typhoon Yolanda	107	107	214	26	WTX	West Texas Explosion	81	81	162

[3] CrisisLexT26 `http://crisislex.org/data-collections.html#CrisisLexT26`.

3.2 Features Engineering

In order to assess the advantage of using semantic features compared to more traditional statistical features, we distinguish two different feature sets; (1) *statistical* features, and; (2) *semantic* features. *Statistical* features have widely been used in the literature [8,9,14,11,20,25] and are posed as the baseline approach for our work. They capture quantifiable linguistic features and other statistical properties of a given post. On the other hand, *semantic* features capture more contextual information of documents, such as the named entities emerging in a given text, as well as their hierarchical semantic information extracted from external knowledge graphs.

Statistical Features: For every tweet in the dataset, the following *statistical* features are extracted:

- *Number of nouns*: nouns generally refer to different entities involved in the crisis event such as locations, actors, or resources involved in the crisis event [8,9,20].
- *Number of verbs*: verbs indicate actions that occur in a crisis event [8,9,20].
- *Number of pronouns*: as with nouns, pronouns may indicate involvement of the actors, locations, or resources.
- *Tweet Length*: number of characters in a post. The length of a post may determine the amount of information contained [8,9,19].
- *Number of words*: number of words may be another indicator of the amount of information contained within a post [8,11] .
- *Number of Hashtags*: hashtags reflect the themes of the post and are manually generated by the posts' authors [8,9,11].
- *Unigrams*: unigrams provide a keyword-based representation of the content of the posts [8,9,11,14,25,20]

The Part Of Speech (POS) features (e.g., *nouns*, *verbs*, *pronouns*) are extracted using the spaCy library.[4] Unigrams are extracted with the regexp tokenizer provided in NLTK.[5] Stop-words are removed using a stop-words list,[6] Stemming is also performed using the Porter Stemmer. Finally, TF-IDF vector normalisatiton is also applied in order to weight the importance of words (tokens) in the documents according to their relative importance within the dataset. This resulted in a total number of 10757 unigrams (i.e., vocabulary size) for the entire balanced dataset.

Semantic Features: Semantic features are designed to generalise information representation across crises. They are designed to be less crisis specific compared to *statistical* features. We use the Name Entity Recogniser (NER) service Babelfy,[7]

[4] SpaCy Library, `https://spacy.io`.
[5] Regexp Tokenizer (NLTK), `http://www.nltk.org/_modules/nltk/tokenize/regexp.html`.
[6] Stop Words List, `https://raw.githubusercontent.com/6/stopwords-json/master/stopwords-all.json`.
[7] Babelfy, `http://babelfy.org`.

and two different knowledge bases for creating these features: (1) BabelNet,[8] and; (2) DBpedia:[9]

- *Babelfy Entities*: the entities extracted by the BabelNet NER tool (e.g., *news*, *sadness*, *terremoto*). Babelfy extracts and disambiguates entities linked to the BabelNet[15] knowledge base.
- *BabelNet Senses (English)*: the English labels associated with the entities returned by Babelfy (e.g., *news→news*, *sadness→sadness*, *terremoto→earthquake*).
- *BabelNet Hypernyms (English)*: the direct English hypernyms (at distance-1) of each entities extracted from BableNet. Hypernyms can broaden the context of an entity, and can enhance the semantics of a document [12] (e.g., *broadcasting*, *communiucation*, *emotion*).
- *DBpedia Properties*: a list of properties associated with the DBpedia URI returned by Babelfy. The following propertiesare queried using SPARQL: `dct:subject`, `rdfs:label` (only in English), `rdf:type` (only of the type http://schema.org and http://dbpedia.org/ontology), `dbo:city`, `dbp:state`, `dbo:state`, `dbp:country` and `dbo:country` (the location properties fluctuate between `dbp` and `dbo`) (e.g., `dbc:Grief`, `dbc:Emotions`, `dbr:Sadness`).

Using hypernyms shown to enhance the semantics of a document [12], and can assist the context representation of documents by correlating different entities with a similar context. For instance, the following four entities *fireman*, *policeman*, *MP (Military Police)*, and *garda* (an Irish word for police) share a common English hypernym: *defender*. To generalise the semantics for tweets in different languages, we formulate the semantics in English. As a result, we prevent the sparsity that results from the varying morphological forms of concepts across languages (see Table 2 to see an example). The senses and hypernyms are both derived from the BabelNet, and together form the *BabelNet Semantics*. The semantic expansion of the data-set through *BabelNet Semantics* expands the vocabulary (in comparison to the case with statistical features) by 3057.

Besides the *BabelNet Semantics*, we also use DBpedia properties to obtain more information about the entity (see Table 2) in the form of subject, label, and location specific properties. Semantic expansion of the dataset through *DBpedia Semantics* increases the vocabulary (in comparison to the vocabulary from statistical features) by 1733.

We use both of these semantic features, *BabelNet & DBpedia Semantics*, individually and also in combination with each other, while developing the binary classifiers to identify crisis-related posts from unrelated ones. When both BabelNet Semantics and DBpedia semantics are used, the vocabulary (in comparison to the vocabulary as determined in statistical features) is increased by 3824. Our experiments will determine whether or not such vocabulary extensions can be regarded as enhancements.

[8] BabelNet, `http://babelnet.org`.
[9] DBpedia, `http://dbpedia.org`.

Table 2. Semantic expansion with BabelNet and DBpedia semantics.

	Post A	Post B
Feature	*'Sad news to report from #Guatemala -at least 8 confirmed dead, possibly more, by this morning's major earthquake.'*	*'Terremoto 7,4 Richter Guatemala deja 15 fallecidos,casas en el suelo, 100 desaperecidos, 100MIL personas sin luz FO'*
Babelfy Entities	*news, sadness, dead, describe, earthquake*	*terremoto, casas, suelo, luz, fallecidos*
BabelNet Sense (English)	*news, sadness, dead, describe, earthquake*	*earthquake, house, soil, light, dead*
BabelNet Hypernyms (English)	*broadcasting, communication, emotion, feeling, people, deceased, inform, natural disaster, geologica phenomenon*	*natural disaster, geological phenomenon, building, Structure, residential_building granular material, people, deceased*
DBpedia Properties	`dbc:Grief`, `dbc:Emotions`, `dbr:Sadness`, `dbc:Demography`, `dbr:Death`, `dbc:Communication`, `dbr:News`, `dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`	`dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`, `dbc:Home`, `dbc:Structural_system`, `dbc:Light`, `dbr:Death`, `dbc:Demography`

3.3 Classifier Selection

For our binary classification problem, we took into consideration the high dimensionality generated from unigrams and semantic features, and the need to avoid over fitting. In comparison to the large dimensionality of the features, which is in the range of 10-15k under different feature combinations, the training examples are smaller in size (around 6000). This encouraged us to opt for Support Vector Machine(SVM) with a Linear Kernel as the classification model, since this model has been found effective for such kind of problems.[10] Additionally, we validated the appropriateness of SVM Linear Kernel against RBF kernel, Polynomial kernel, and Logistic Regression. Based on 20 runs of 5 fold cross-validation of different feature combinations, SVM Linear Kernel was found to be more statistically significant, and had a better mean F_1 value of 0.8118 and a p-value of < 0.00001 when compared to other classifiers (by performing a t-test followed by calculating p-value).

4 Crisis-related Content Classification Across Crises

In this section, we detail the experimental set up and create the models based on various criteria. Further, we report the results and discuss how including

[10] A Practical Guide to Support Vector Classification, `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

the expanded semantic features impacted the performance of our classifiers, particularly in the cases when it is applied to cross-crisis scenarios.

4.1 Experimental Setting

The experiments are designed to train and evaluate the classification models on (i) the entire dataset, i.e., on all 26 crisis events, (ii) a selection of train/test crisis event data, based on certain criteria for cross-crisis evaluation.

Crisis Classification Models: For the first experiment, we create different classifiers to compute and compare the performance of various feature combinations. Here, we aim to see when all the 26 events (section 3.1) are merged, whether the inclusion of semantics boosts the binary classification. We create multiple classifiers and evaluate them using 5-fold cross validation. To this end, we used scikit-learn library.[11] The different classifiers are trained based on different features combinations:

- *SF*: A classifier generated with the statistical features only; our baseline.
- *SF+SemEF_BN*: A classifier generated with the statistical features and the semantic features from *BabelNet Semantics* (entity sense, and their hypernyms).
- *SF+SemEF_DB*: A classifier generated with the statistical features, and the semantic features from *DBpedia Semantics* (label, type, and other DBpedia properties).
- *SF+SemEF_BNDB*: A classifier generated with the statistical features, and the combination of semantic features from *BabelNet and DBpedia Semantics*.

Cross-Crisis Classification: For the second experiment, we aim at evaluating models on event types that are not observed during training (e.g., evaluate models on earthquake data, whereas it was trained on flood events). The models are trained on different combination of features and various types of crisis events. We generate the classifiers for the feature combinations as described in the previous experiment (see above). However, in this case, we divide the data into training and test sets based on 2 different criteria as described below:

1. Identify posts from a crisis event, when the *type* of event is already included in the training data (e.g., process tweets from a new flood incident when tweets from other flood crisis are in the training data).
2. Identify posts from a crisis event, when the *type* of the event is not included in the training data.

Since the criteria are defined on the *types* of the events, we hereby distribute the 26 events broadly in 11 *types* as given in Table 3. This categorisation is based on personal understandings of the nature of different types of crisis events, and how related or discrete they might be based on their effects. For instance, we have assumed the *type* of Flood and Typhoon as highly similar, considering that flood are typical direct outcomes of Typhoons (more about this in Section 5).

[11] Scikit-learn, `http://scikit-learn.org`.
[12] NYT has only 3 tweets in total.

Table 3. Types of events in the dataset.

Event Type (Nb.)	Event Instances	Event Type (Nb.)	Event Instances
Wildfire/Bushfire (2)	CWF, ABF	Haze (1)	SGR
Earthquake (4)	COS, ITL, BOL, GAU	Helicopter crash (1)	GLW
Flood/Typhoon (8)	TPY, TYP, CFL, QFL, ALB, PHF, SAR, MNL	Building collapse (1)	SVR
Terror Shooting/Bombing (2)	LAX, BOB	Location Fire (2)	BRZ, VNZ
Train crash (2)	SPT, LAM[12]	Explosion (1)	WTX
Meteor (1)	RUS		

4.2 Results: Crisis Classification

In this section, we present the results from the first experiment, where the entire data (spread across 26 events and all our 11 event types) is merged. The models are trained using 20 iterations of 5-fold cross validation. The results are presented in Table 4. We report the mean of *Precision* (P_{mean}), *Recall* (R_{mean}), and F_1 score (F_{mean}) from 20 iterations, standard deviation in F_1 score (σ), and percentage change of F_1 score compared to the baseline ($\Delta F/F$).

Table 4. Crisis-related content classification results using 20 iterations of 5-fold cross validation, $\Delta F/F(\%)$ showing percentage gain/loss of the statistical semantics classifiers against the statistical baseline classifier.

Model	P_{mean}	R_{mean}	F_{mean}	Std. Dev. (σ)	$\Delta F/F$ (%)	Sig. (p-value)
SF (Baseline)	0.8145	0.8093	0.8118	0.0101	-	-
SF+SemEF_BN	0.8233	0.8231	0.8231	0.0111	1.3919	<0.000 01
SF+SemEF_DB	0.8148	0.8146	0.8145	0.0113	0.3326	0.018 78
SF+SemEF_BNDB	0.8169	0.8167	0.8167	0.0106	0.6036	0.000 011

In general, we observe that there is a very small change against the baseline classifier and that both classifiers are able to achieve $F_{mean} > 81\%$. The most noticeable improvement compared to the baseline can be observed for SF+SemEF_BN (1.39%) and SF+SemEF_BNDB (0.6%), which are both statistically significant ($p < 0.05$) based on a 2-tailed one-sample t-test, where the F_{mean} of SF is treated as the null-hypothesis.

To better understand the impact of semantics on the classifier, we perform feature selection using Information Gain (IG) to determine the most informative features and how they vary across the classifiers. In SF model, we observe very event-specific features such as *collapse, terremoto, fire, earthquake, #earthquake, flood, typhoon, injured, quake* (Table 5). Within the top features, we also see 7 hashtags among the top 50 features, which reflects how event specific vocabulary plays a role in our classifier and how it may be an issue when dealing with new crisis types.

For SF+SemEF_BN and SF+SemEF_DB models, we observed concepts such as *natural_hazard, structural_integrity_and_failure, conflagration, geolog-*

ical phenomenon, perception, dbo:location, dbo:place, dbc:building_defect, dbc:solid_mechanics among the top 50 crisis-relatedness predictors (Table 5).

Looking more into the results, we can observe that *Structural_integrity_and_failure* is the annotated entity for terms like *collapse, building collapse* which are frequently occurring terms in the earthquake events, floods events, and Savar Building collapse. This is expected considering the significant number of earthquakes and floods events in the data. The *natural_disaster* hypernym is linked to several crisis events terms in the data such as *flood, landslide, earthquake*. Similarly, SF+SemEF_BNDB reflected a combination of both BabelNet and DBpedia semantics among informative features. These results show that semantics may help when dealing with new crisis types.

Although semantic models do not appear to be highly beneficial compared to purely statistical models when dealing with already seen event types, we observed the potential limitations of statistical features when dealing with new event types. Statistical features appear to be overly tied to event instances whereas semantic features seems to better generalise crisis-related concepts.

Table 5. IG-Score ranks of features for: SF, SF+SemEF_BN and SF+SemEF_DB.

	SF		SF+SemEF_BN		SF+SemEF_DB	
R.	IG	Feature	IG	Feature	IG	Feature
1	0.106	No.OfHashTag	0.106	No.OfHashTag	0.106	No.OfHashTag
2	0.046	costa	0.056	costa	0.044	No.OfNouns
3	0.044	No.ofNoun	0.044	No.OfNouns	0.036	costa_rica
4	0.044	rica	0.044	rica	0.035	dbc:countries_in_central_americ
5	0.035	collapse	0.036	costa_rica	0.035	collapse
6	0.033	terremoto	0.035	central_american_country	0.031	terremoto
7	0.026	TweetLength	0.032	collapse	0.027	dbo:place
8	0.025	7	0.031	terremoto	0.026	TweetLength
9	0.024	#earthquake	0.026	TweetLength	0.024	#earthquake
10	0.023	bangladesh	0.026	fire	0.024	dbo:location
11	0.022	No.OfVerb	0.024	#earthquake	0.023	dbo:populatedplace
12	0.022	#redoctober	0.023	structural_integrity_and_failur	0.023	dbc:safes
13	0.021	No.OfWords	0.023	coastal	0.022	structural_integrity_and_failure
14	0.018	tsunami	0.022	information	0.022	dbc:building_defect
15	0.017	fire	0.022	financial_condition	0.022	dbc:solid_mechanics
16	0.016	building	0.022	No.OfVerbs	0.022	dbc:engineering_failure
17	0.016	rt	0.022	#redoctober	0.022	bangladesh
18	0.015	factory	0.021	No.OfWords	0.022	dbc:flood
19	0.014	toll	0.020	shore	0.022	dbr:wealth
20	0.014	flood	0.020	building	0.022	No.OfVerbs
21	0.013	#bangladesh	0.019	anatomical_structure	0.021	No.OfWords
22	0.013	#colorad	0.019	phenomenon	0.02	dbc:coastal_geography
23	0.012	alert	0.018	natural_disaster	0.019	dbc:article_containing_video_clip
24	0.012	hit	0.018	failure	0.018	dbc:natural_hazard
25	0.012	typhoon	0.017	conflagration	0.017	fire

4.3 Results: Cross-Crisis Classification

We now evaluate the ability of the classifiers and feature to deal with event types that are not present in training data. We first evaluate the model on new instances of event types that have been already seen (Criteria 1) and then perform a similar task but omit event-types in the training dataset (Criteria 2).

Criteria 1 - Content relatedness classification of already seen event types. For the first sub-task, we evaluate our models on new event instances of event types already included when training the models (e.g., evaluate a new flood event on a model trained on data that include previous floods). We train the classifier on 25 crisis events, and use the 26th event as a test dataset.

As shown in Table 3, 26 crisis events have broadly been categorised under 11 *types*. In order to select the *type* of crisis events to test, we looked for such *types* which had a strong presence in the overall dataset. We opted for such crisis events which had at least 4 or more crisis events under the same type. As a result we consider two event types to evaluate: (1) *Flood/Typhoons* event types, and; (2) *Earthquake* event types.

For evaluating the models, we use following events as test data events: (1) For *Flood/Typhoons* we use *Typhoon Yolanda (TPY), Typhoon Pablo (TYP), Alberta Flood (ALB), Queensland Flood (QFL), Colorado Flood (CFL), Philippines Flood (PHF)* and *Sardinia Flood (SAR)* as evaluation data, and; (2) for *Earthquake*, we use *Guatemala Earthquake (GAU), Italy Earthquake (ITL), Bohol Earthquake (BOL)* and *Costa Rica Earthquake (COS)* as evaluation data. For example, when we evaluate the classifiers for TPY, we train our models on all the other 25 events and use the TPY data for the evaluation.

Table 6. Cross-crisis relatedness classification: criteria 1 (best F_1 score is highlighted for each event).

	Instances		SF			SF+SemEF_BN				SF+SemEF_DB				SF+SemEF_BNDB			
Test event	Train	Test	P	R	F_1	P	R	F_1	$\Delta F/F$ (in %)	P	R	F_1	$\Delta F/F$ (in %)	P	R	F	$\Delta F/F$ (in %)
TPY	5717	214	0.808	0.804	0.803	0.777	0.776	0.776	-3.44	0.772	0.771	0.771	-4.01	0.780	0.780	0.780	-2.83
TYP	5755	176	0.876	0.864	0.863	0.853	0.841	0.840	-2.66	0.831	0.83	0.829	-3.84	0.861	0.852	0.851	-1.29
ALB	5899	32	0.72	0.719	0.718	0.754	0.75	0.749	4.25	0.845	0.844	0.844	17.41	0.845	0.844	0.844	17.41
QFL	5375	556	0.791	0.784	0.783	0.80	0.793	0.792	1.18	0.780	0.772	0.77	-1.66	0.789	0.782	0.781	-0.22
CFL	5809	122	0.82	0.803	0.801	0.835	0.828	0.827	3.28	0.806	0.762	0.754	-5.88	0.796	0.77	0.765	-4.41
PHF	5791	140	0.764	0.764	0.764	0.769	0.764	0.763	-0.13	0.772	0.771	0.771	0.93	0.744	0.743	0.743	-2.83
SAR	5797	134	0.684	0.612	0.570	0.747	0.694	0.677	18.79	0.702	0.664	0.648	13.70	0.696	0.664	0.650	14.10
GAU	5725	206	0.788	0.782	0.780	0.739	0.728	0.725	-7.1	0.798	0.786	0.784	0.51	0.779	0.772	0.770	-1.30
ITL	5819	112	0.595	0.589	0.583	0.619	0.589	0.562	-3.58	0.667	0.634	0.615	5.49	0.659	0.616	0.588	0.98
BOL	5869	62	0.743	0.742	0.742	0.732	0.726	0.724	-2.38	0.758	0.758	0.758	2.20	0.684	0.677	0.674	-9.07
COS	4991	940	0.794	0.790	0.790	0.773	0.770	0.770	-2.56	0.740	0.739	0.739	-6.42	0.751	0.750	0.750	-5.08

From the results in Table 6 it can be seen that, when the event *type* is previously seen by the classifier in the training data, the improvement from adding semantic features is small and inconsistent over the test cases. SF+SemEF_BN shows improvement over the baseline in 4 out of 11 evaluation cases, while SF+SemEF_DB shows improvement in 6 out of 11 evaluation cases. The average percentage gain ($\Delta F/F$) varies between +0.52% (SF+SemEF_BN) and +1.67% (SF+SemEF_DB) with a standard deviation varying between 6.89% to 7.78%. It indicates that almost half of the test event cases do not show improvement over

the statistical features baseline's F_1 score.

Criteria 2 - Content relatedness classification of unseen crisis types. In criteria 1, we considered the classification of new event instances when similar events already appeared in the classifier training data. In criteria 2 we test the classifier on types of events that are not seen by the classifier in the training data *types*. We select the following events and event types: (1) train the classifiers on rest of the event *types* except *Terror Shooting/Bombing and Train Crash* and evaluate on *Los Angeles Airport Shooting (LAX), Lac Megantic Train Crash (LAM), Boston Bombing (BOB)*, and *Spain Train Crash (SPT)*; (2) train the classifiers on rest of the event *types* except *Flood/Typhoon* and evaluate on *TPY, TYP, ALB, QFL, CFL, PHF,* and *SAR*, and; (3) train the classifiers on rest of the event *types* except *Earthquake* and evaluate on *GAU, ITL, BOL,* and *COS*.

Table 7. Cross-crisis relatedness classification: criteria 2 (best F_1 score is highlighted for each event).

	Instances		SF			SF+SemEF_BN				SF+SemEF_DB				SF+SemEF_BNDB			
Test event	Train	Test	P	R	F_1	P	R	F_1	$\Delta F/F$ (in %)	P	R	F_1	$\Delta F/F$ (in %)	P	R	F	$\Delta F/F$ (in %)
LAX	5407	224	0.664	0.656	0.652	0.681	0.679	0.677	3.90	0.666	0.665	0.665	1.95	0.657	0.656	0.656	0.58
LAM	5844	68	0.655	0.632	0.618	0.642	0.632	0.626	1.2	0.619	0.618	0.616	-0.34	0.638	0.632	0.628	1.62
BOB	5407	138	0.669	0.630	0.608	0.663	0.645	0.635	4.40	0.613	0.609	0.605	-0.56	0.628	0.616	0.607	-0.19
SPT	5844	16	0.573	0.563	0.547	0.690	0.688	0.686	25.56	0.767	0.750	0.746	36.5	0.69	0.688	0.686	25.56
TPY	4409	214	0.714	0.664	0.642	0.715	0.640	0.606	-5.67	0.69	0.664	0.651	1.39	0.676	0.617	0.582	-9.45
TYP	4409	176	0.769	0.699	0.678	0.802	0.705	0.679	0.12	0.742	0.682	0.661	-2.54	0.733	0.642	0.603	-10.99
ALB	4409	32	0.727	0.719	0.716	0.771	0.719	0.705	-1.63	0.833	0.813	0.81	13.02	0.742	0.719	0.712	-0.63
QFL	4409	556	0.734	0.694	0.681	0.728	0.676	0.657	-3.51	0.733	0.707	0.698	2.58	0.741	0.707	0.696	2.23
CFL	4409	122	0.792	0.779	0.776	0.736	0.713	0.7060	-9.04	0.707	0.705	0.704	-9.27	0.755	0.754	0.754	-2.87
PHF	4409	140	0.589	0.564	0.532	0.672	0.607	0.566	6.52	0.662	0.643	0.632	18.9	0.617	0.586	0.556	4.67
SAR	4409	134	0.663	0.590	0.537	0.660	0.597	0.553	2.93	0.658	0.619	0.595	10.69	0.691	0.642	0.617	14.84
GAU	4611	206	0.610	0.553	0.487	0.584	0.549	0.495	1.62	0.692	0.650	0.630	29.39	0.667	0.621	0.593	21.79
ITL	4611	112	0.546	0.536	0.509	0.632	0.571	0.516	1.26	0.633	0.589	0.553	8.54	0.661	0.598	0.555	8.93
BOL	4611	62	0.732	0.726	0.724	0.656	0.645	0.639	-11.73	0.684	0.677	0.674	-6.86	0.606	0.597	0.588	-18.77
COS	4611	940	0.595	0.560	0.515	0.626	0.554	0.480	-6.71	0.618	0.578	0.538	4.56	0.645	0.580	0.527	2.33

From results in Table 7, we observe that the average best performing feature is the DBpedia semantics SF+SemEF_DB as it shows an average percentage gain in F_1 score ($\Delta F/F$) of +7.2% (with a Std. Dev. of 12.83%) and shows improvement over the baseline SF classifier in 10 out of 15 events.

Out of 5 events where it does not show improvement, in 2 events the percentage loss ($\Delta F/F$) is -0.34% and -0.56%. SF+SemEF_BNDB shows improvement over the baseline in 9 out of 15 events with an average percentage gain of +2.64% in F_1 score ($\Delta F/F$) over the SF classifier. When we compare this to criteria 1, it appears that semantic features (particularly from DBpedia) enhances the classification performance over statistical features alone when the *type* of event is not seen by the classifier during training. This result shows that although semantics may not improve relatedness classification when dealing with already

seen event types, semantics are useful when dealing with event types not found in training datasets. This makes semantic feature more robust than statistical features.

5 Discussion and Future Work

Our experiments explored the impact of mixing semantic features with statistical features, and created a hybrid model, to classify crisis *related* and *not related* posts. We noticed a significant impact of semantics in the scenario when the *type* of the crisis is new to the classifier.While both the *BabelNet* and *DBpedia semantics* performed better than the statistical features, *DBpedia semantics* was found to be more consistent in its performance while classifying a new *type* of crisis event. This is likely because of the better coverage and semantic depth that DBpedia provides.

To better understand the role of semantics in crisis-related content classification, we randomly picked some tweets that were misclassified by either the baseline classifier or the semantic classifiers in the criteria 1 and 2 evaluations. We observed that: (i) semantics can generalise event specific terms compared to statistical features and consequently adapt to new event types (e.g., `dbc:flood` and `dbc:natural_hazard`), (ii) semantic concept can be sometimes too general and not help the classification of the document (e.g., *desire* and *virtue* hypernyms), and (iii) general automatic semantic extraction tools can extract non-relevant entities and confuse the classifiers (e.g., entities about *Formula 1*).

Althought this analysis gives better insights concerning the behaviour of the classifiers, we plan to run a more in depth error analysis in the future by analysing additional misclassified documents. This will help improve our understanding of the scenarios and conditions under which each classification approach prevails, and thus would help us determine a more accurate merge between the two classification approaches.

In this work, we performed experiments across different *types* of crisis events. The event types present in the datasets are not uniformly distributed, where some types are more frequent than others, or have much bigger data than others. (See Table 3). In the view of developing automated classifiers that are able to learn about various crisis situations, such a skewed distribution could lead to learning bias. We designed the experiments in light of this distribution, but in order to create classifier models that are able to adapt to various domains of crisis, we would need to learn from more diverse set of crisis situations.

The type of each crisis in the data is the official type which is determined by official agencies (e.g., typhoon, earthquake, flood). We regarded each type to be different from the others, based solely on their *type* label. However, with regards to content, it is not necessarily the case that different type of crises would produce different type of content (e.g., typhoons and floods have a high overlap). To this end, while we do not add a certain *type* of crisis to the training data, we cannot ignore the possibility of having highly related content in the training data, that was the results of including similar or overlapping crises events. Hence

in future work, we will take into account not only the event *type*, but also their *content similarity*.

In this work, we dealt with data originating from different languages, but have not performed a cross-lingual analysis. As an immediate future work, we aim to analyse how the classifiers trained in a certain language can adapt to an entirely new language to detect crisis related content.

6 Conclusion

This work presents a hybrid approach by merging semantic and statistical features to develop classification models that detect crisis related information from social media posts. The main application of this approach is demonstrated in the case of identifying crisis-related content on *new types* of crisis events that have not been directly included in the data used for training the classifier. This proposes a way forward towards developing domain adaptive crisis classification models. Adding semantic features reflected an improvement over the statistical features in classification performance on an average of 7.2% when identifying crisis related content on new event types.

Acknowledgment: This work has received support from the European Unions Horizon 2020 research and innovation programme under grant agreement No 687847 (COMRADES).

References

1. Abel, F., Celik, I., Houben, G.J. and Siehndel, P. Leveraging the semantics of tweets for adaptive faceted search on twitter. Int. Semantic Web Conf. (ISWC), Bonn, Germany, 2011.
2. Abel, F., Hauff, C., Houben, G. J., Stronkman, R., and Tao, K. Semantics+ filtering+ search= twitcident. exploring information in social web streams. Conf. Hypertext and Social Media (Hypertext), WI., USA, 2012
3. Burel, G., Saif, H., Fernandez, M., and Alani, H. (2017). On Semantics and Deep Learning for Event Detection in Crisis Situations. Workshop on Semantic Deep Learning (SemDeep), at ESWC, Portoroz, Slovenia, 2017.
4. Burel, G., Saif, H., and Alani, H. (2017). Semantic Wide and Deep Learning for Detecting Crisis-Information Categories on Social Media. The Semantic Web ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, 2017.
5. Cristianini, N. and Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.
6. Dai, W., Xue, G.R., Yang, Q. and Yu, Y., 2007, July. Transferring naive bayes classifiers for text classification. In AAAI (Vol. 7, pp. 540-545).
7. Gao, H., Barbier, G., & Goolsby, R. Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intelligent Systems, 26(3), 10-14, 2011.
8. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. and Meier, P. Practical extraction of disaster-relevant information from social media. Int. World Wide Web Conf. (WWW), Rio de Janeiro, Brazil, 2013.
9. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. and Meier, P., 2013, May. Extracting information nuggets from disaster-Related messages in social media. In ISCRAM.

10. Jadhav, A.S., Purohit, H., Kapanipathi, P., Anantharam, P., Ranabahu, A.H., Nguyen, V., Mendes, P.N., Smith, A.G., Cooney, M. and Sheth, A.P. Twitris 2.0: Semantically empowered system for understanding perceptions from social data, http://knoesis.wright.edu/library/download/Twitris_ISWC_2010.pdf, 2010.
11. Karimi, S., Yin, J. and Paris, C. December. Classifying microblogs for disasters. Australasian Document Computing Symposium, Brisbane, QLD, Australia, 2013.
12. Khare, P., Fernandez, M. and Alani, H., 2017. Statistical Semantic Classification of Crisis Information. Workshop on HSSUES at ISWC, Vienna, Austria, 2017
13. Kogan, M., Palen, L. and Anderson, K.M. February. Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. Conf. on computer supported cooperative work & social computing (CSCW '15), Vancouver, Canada, 2015.
14. Li, R., Lei, K.H., Khadiwala, R. and Chang, K.C.C., 2012, April. Tedas: A twitter-based event detection and analysis system. IEEE 28th Int. Conf. on Data Engineering (ICDE), Washington, DC, USA, 2012.
15. Navigli, R. and Ponzetto, S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, pp.217-250, 2012.
16. Olteanu, A., Vieweg, S. and Castillo, C., 2015, February. What to expect when the unexpected happens: Social media communications across crises. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 994-1009). ACM.
17. Power, R., Robinson, B., Colton, J. and Cameron, M. Emergency situation awareness: Twitter case studies. Int. Conf. on Info. Systems for Crisis Response and Management in Mediterranean Countries (ISCRAM), Toulouse, France, 2014.
18. Rogstadius, J., Vukovic, M., Teixeira, C.A., Kostakos, V., Karapanos, E. and Laredo, J.A. CrisisTracker: Crowdsourced social media curation for disaster awareness. IBM Journal of Research and Development, 57(5), pp.4-1, 2013.
19. Sakaki, T., Okazaki, M. and Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. Int. Conf. World Wide Web (WWW), Raleigh, North Carolina USA, 2010.
20. Stowe, K., Paul, M., Palmer, M., Palen, L. and Anderson, K. Identifying and Categorizing Disaster-Related Tweets. Workshop on Natural Language Processing for Social Media, In EMNLP, Austin, Texas, USA, 2016.
21. Vieweg, S., Hughes, A.L., Starbird, K. and Palen, L.. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. Proc. SIGCHI Conf. on Human Factors in Computing Systems (CHI), GA, USA, 2010.
22. Vieweg, S.E., 2012. Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications (Doctoral dissertation, University of Colorado at Boulder), https://works.bepress.com/vieweg/15/
23. Tonon, A., Cudr-Mauroux, P., Blarer, A., Lenders, V. and Motik, B., 2017, May. ArmaTweet: Detecting Events by Semantic Tweet Analysis. In European Semantic Web Conference (pp. 138-153). Springer, Cham.
24. Yin, J., Lampert, A., Cameron, M., Robinson, B. and Power, R. Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 27(6), 2012.
25. Zhang, S. and Vucetic, S.. Semi-supervised Discovery of Informative Tweets During the Emerging Disasters. arXiv preprint arXiv:1610.03750, 2016.