Cross-Lingual Classification of Crisis Data

Prashant Khare[1](✉), Grégoire Burel[1](✉), Diana Maynard[2](✉), and Harith Alani[1](✉)

[1] Knowledge Media Institute, The Open University, Milton Keynes, UK
{prashant.khare,g.burel,h.alani}@open.ac.uk
[2] Department of Computer Science, University of Sheffield, Sheffield, UK
d.maynard@sheffield.ac.uk

Abstract. Many citizens nowadays flock to social media during crises to share or acquire the latest information about the event. Due to the sheer volume of data typically circulated during such events, it is necessary to be able to efficiently filter out irrelevant posts, thus focusing attention on the posts that are truly relevant to the crisis. Current methods for classifying the relevance of posts to a crisis or set of crises typically struggle to deal with posts in different languages, and it is not viable during rapidly evolving crisis situations to train new models for each language. In this paper we test statistical and semantic classification approaches on cross-lingual datasets from 30 crisis events, consisting of posts written mainly in English, Spanish, and Italian. We experiment with scenarios where the model is trained on one language and tested on another, and where the data is translated to a single language. We show that the addition of semantic features extracted from external knowledge bases improve accuracy over a purely statistical model.

Keywords: Semantics · Cross-lingual · Multilingual
Crisis informatics · Tweet classification

1 Introduction

Social media platforms have become prime sources of information during crises, particularly concerning rescue and relief requests. During Hurricane Harvey, over 7 million tweets were posted about the disaster in just over a month[1], while over 20 million tweets with the words #sandy and #hurricane were posted in just a few days during the Hurricane Sandy disaster.[2] Sharing such vital information on social media creates real opportunities for increasing citizens' situational awareness of the crisis, and for authorities and relief agencies to target their efforts more efficiently [23]. However, with such opportunities come real challenges, such as the handling of such large and rapid volumes of posts, which

[1] https://digital.library.unt.edu/ark:/67531/metadc993940/.

[2] Mashable: Sandy Sparks 20 Million Tweets http://mashable.com/2012/11/02/hurricane-sandy-twitter.

D. Vrandečić et al. (Eds.): ISWC 2018, LNCS 11136, pp. 617–633, 2018.
https://doi.org/10.1007/978-3-030-00671-6_36

renders manual processing highly inadequate [7]. The problem is exacerbated by the findings that many of these posts bear little relevance to the crisis, even those that use the dedicated hashtags [11].

Because of these challenges, there is an increasingly desperate need for tools capable of automatically assessing crisis information relevancy, to filter out irrelevant posts quickly during a crisis, and thus reducing the load to only those posts that matter. Recent research explored various classification methods of crisis data from social media platforms, which aimed at automatically categorising them into *crisis-related* or *not related* using supervised [10,13,21,25] and unsupervised [18] machine learning approaches. Most of these methods use statistical features, such as n-grams, text length, POS, and hashtags.

One of the problems with such approaches is their bias towards the data on which they are trained. This means that classification accuracy drops considerably when the data changes, for example when the crisis is of a different type, or when the posts are in a different language, in comparison to the crisis type and language the model was trained on. Training the models for all possible crisis types and languages is infeasible due to time and expense.

In our previous work, we showed that adding semantic features increases the classification accuracy when training the model on one type of crisis (e.g. floods), and applying it to another (e.g. bushfires) [11]. In this paper, we tackle the problem of language, where the model is trained on one language (e.g. English), but the incoming posts are in another (e.g. Spanish). We explore the role of adding semantics in increasing the multilingual fitness of supervised models for classifying the relevancy of crisis information.

The main contributions of this paper can be summarised as follow:

1. We build a statistical-semantic classification model with semantics extracted from BabelNet and DBpedia.
2. We experiment with classifying relevancy of tweets from 30 crisis events in 3 languages (English, Spanish, and Italian).
3. We run relevancy classifiers with datasets translated into a single language, as well as with cross-lingual datasets.
4. We show that adding semantics increases cross-lingual classification accuracy by 8.26%–9.07% in average F_1 in comparison to traditional statistical models.
5. We show that when datasets are translated into the same language, only the model that uses BabelNet semantics outperforms the statistical model, by 3.75%.

The paper is structured as follows: Sect. 2 summarises related work. Sections 3 and 4 describe our approach and experiments on classifying cross-lingual crisis data using different semantic features. Results are reported in Sects. 4.2 and 4.3. Discussion and conclusions are in Sects. 5 and 6.

2 Related Work

Classification of social media messages about crises and disasters in terms of their relevancy has been addressed already by a number of researchers [2,3,9–12,22, 25]. The classification types can differ, however. Some classify simply as relevant (related) or not; some include a partly relevant category; while others include the notion of informativeness (where informative is taken to mean providing useful information about the event). For example, Olteanu et al. [16] use the categories *related and informative*, *related but not informative*, and *not related*. Others treat relevance and informativeness as two separate tasks [5].

Methods for this kind of classification use a variety of supervised machine learning approaches, usually relying on linguistic and statistical features such as POS tags, user mentions, post length, and hashtags [8–10,19,21]. Approaches range from traditional classification methods such as Support Vector Machines (SVM), Naive Bayes, and Conditional Random Fields [9,17,21] to the more recent use of deep learning and word embeddings [3].

One of the drawbacks to these approaches is their lack of adaptability to new kinds of data. [9] took early steps in this area by training a model on messages about the Joplin 2011 tornado and applying it to messages about Hurricane Sandy, although the two events are still quite similar. [12] took this further by using semantic information to adapt a relevance classifier to new crisis events, using 26 different events of varying types, and showed that the addition of semantics increases the adaptability of the classifier to new, unseen, types of crisis events. In this paper, we develop that approach further by examining whether semantic information can help not just with new events, but also with events in different languages.

In general, adapting classification tools to different languages is a problem for many NLP tasks., since it is often difficult to acquire sufficient data to train separate models on each language. This is especially true for tasks such as sentiment analysis, where leveraging information from data in different languages is required. In that field, two main solutions have been explored: either translating the data into a single language (normally English) and using this single dataset for training and/or testing [1]; or training a model using weakly-labelled data without supervision [6]. Severyn et al. [20] improved performance of sentiment classification using distant pre-training of a CNN, consisting of inferring weak labels (emoticons) from a large set of multilingual tweets, followed by additional supervised training on a smaller set of manually annotated labels. In the other direction, annotation resources (such as sentiment lexicons) can be transferred from English into the target language to augment the training resources available [14]. A number of other approaches rely on having a set of correspondences between English and the target language(s), such as those which build distributed representations of words in multiple languages, e.g. using Wikipedia [24].

We test two similar approaches in this paper for the classification of information relevancy in crisis situations: (a) translate all datasets into a single language; (b) make use of high-quality feature information in English (and other languages) to supplement the training data of our target language(s).

As far as we know, while these kinds of language adaptation methods have been frequently applied to sentiment analysis, they have not been applied to crisis classification methods. Our work extends mainly on the previous work using hierarchical semantics from knowledge graphs to perform crisis-information classification through a supervised machine learning approach [11,12], by generating statistical and semantic features for all relevant languages and then using this to train the models, regardless of which language is required.

3 Experiment Setup

Our aim is to train and validate a binary classifier that can automatically differentiate between *crisis-related* and *not related* tweets in cross-lingual scenarios. We generate the statistical and semantic features of tweets from different languages and then train the machine learning models accordingly. In the next sections we detail: (i) the datasets used in our experiments; (ii) the statistical and semantic sets of features used; and (iii) the classifier selection process.

3.1 Datasets

For this study, we chose datasets from multiple sources. From the CrisisLex platform[3] we selected 3 datasets: CrisisLexT26, ChileEarthquakeT1, and SOSItalyT4. CrisisLexT26 is an annotated dataset of 26 different crisis events that occurred between 2012 and 2013. Each event has 1000 labeled tweets, with the labels *'Related and Informative', 'Related but not Informative', 'Not Related' and 'Not Applicable'*. These events occurred around the world and hence covered a range of languages. ChileEarthquakeT1 is a dataset of 2000 tweets in Spanish (from the Chilean earthquake of 2010), where all the tweets are labeled by relatedness (relevant or not relevant). The SOSItalyT4 set is a collection of tweets spanning 4 different natural disasters which occurred in Italy between 2009 and 2014, with almost 5.6k tweets labeled by the type of information they convey ("damage", "no damage", or "not relevant"). Based on the guidelines of the labeling, both *"damage"* and *"no damage"* indicate relevance.

We chose all the labeled tweets from these 3 collections. Next, we converged some of the labels, since we aim to generate a binary classifier. From CrisisLexT26, we merged *'Related and Informative'* and *'Related but not Informative'* into the *Related* category, and merged *Not Related* abd *Not Applicable* into the *Not Related* category. For SOSItalyT4 we add the tweets labeled as "damage" and "no damage" to the "Related" category, and the "not relevant" to the "Not Related" category.

Finally, we removed all duplicate instances from the individual datasets to reduce content redundancy, by comparing the tweets in pairs after removing the special characters, URLs, and user-handles (i.e., '@' mentions). This resulted in 21,378 *Related* and 2965 *Not Related* documents in the CrisisLexT26 set, 924 *Related* and 1238 *Not Related* in the Chile Earthquake set, and 4372 *Related* and 878 *Not Related* in the SOSItalyT4 set.

[3] crisislex.org/.

Next, we applied 3 different language detection APIs: detectlanguage[4], langdetect[5], and TextBlob[6]. We labeled the language of the tweet where there was agreement by at least 2 of the APIs. The entire data constituted more than 30 languages, where English (en), Spanish (es), and Italian (it) comprised almost 92% of the collection (29,141 out of 31755). Considering this distribution, we focused our study on these 3 languages. To this end, we first created an unbalanced set (in terms of language) for training the classifier (see Table 1-*unbalanced*). In order to reduce the imbalance between *Related* and *Not Related* tweets, we thus only selected 8,146 tweets in total out of the 29,141 tweets. Next, we create a *balanced* version of the corpus where we split the data into a training and test set for each language, with equal distribution throughout, to remove any bias (Table 1- *balanced*).

Table 1. Data size for English(*en*), Spanish(*es*), and Italian(*it*)

Language	Unbalanced		Balanced			
			Train		Test	
	Not related	Related	Not related	Related	Not related	Related
English (*en*)	2060	2298	612	612	201	200
Italian (*it*)	813	812	612	612	201	200
Spanish (*es*)	1039	1124	612	612	201	200
Total	3912	4234	1836	1836	603	600

We also provide, in Table 2, a breakdown of all the original datasets to give an overview of the language distribution within each crisis event set.

3.2 Feature Engineering

We define two types of feature sets: *statistical* and *semantic*. Statistical features are widely used in various text classification problems [8–10,13,21,25] and so we consider these as our baseline approach. These capture the quantifiable statistical properties and the linguistic features of a textual post, whereas semantic features determine the named entities and associated hierarchical semantic information.

Statistical Features were extracted for each post in the dataset, following previous work, as follows:

- *Number of nouns:* nouns refer to entities occurring in the posts, such as people, locations, or resources involved in the crisis event [8,9,21].
- *Number of verbs:* these indicate actions occurring in a crisis event [8,9,21].

[4] https://detectlanguage.com.
[5] https://pypi.python.org/pypi/langdetect.
[6] http://textblob.readthedocs.io/en/dev/.

Table 2. Language distribution (in %) in crisis events data

Event	Language (%)				Event	Language (%)			
	en	*it*	*es*	Other		*en*	*it*	*es*	Other
Colorado Wildfire	99.30	0	0.09	0.61	CostaRica Quake	45.67	1.96	44.03	8.33
Guatemala Quake	23.84	1.20	69.56	5.40	Italy Quake	18.53	71.10	9.70	0.77
Philippines Flood	91.31	0	0.98	7.71	Typhoon Pablo	81.22	0.22	4.40	14.17
Venezuela Refinery	8.93	0.22	89.8	1.06	Alberta Flood	99.48	0	0	0.52
Australia Bushfire	98.94	0.0	0.10	0.97	Bohol E'quake	86.5	0.12	0.12	13.25
Boston Bombing	93.22	0.21	2.12	4.34	Brazil Club Fire	31.6	0	1.79	66.61
Colorado Floods	99.67	0	0.11	0.22	Glasgow Helicopter	99.86	0	0.11	0.03
LA Airport Shoot	97.07	0.11	1.30	1.52	LacMegantic Train	52.57	0.21	1.16	46.06
Manila Flood	72.40	0.22	0.22	27.16	NY Train Crash	99.86	0.14	0	0
Queensland Flood	99.56	0.09	0	0.35	Russia Meteor	87.56	0.64	2.56	9.24
Sardinia Flood	10.93	88.49	0.12	0.46	Savar Building	86.90	0.82	5.19	7.09
Singapore Haze	97.47	0.0	0	2.53	Spain Train Crash	43.13	0	54.67	2.20
Typhoon Yolanda	91.59	0.11	1.83	6.47	Texas Explosion	94.99	0	3.00	2.01
L'Aquila Quake	4.89	88.58	1.43	5.10	Emilia Quake	1.02	87.99	0.34	10.65
Genova Flood	2.09	95.12	0	2.79	Chile Quake	10.82	0.19	82.00	6.99

- *Number of pronouns:* similar to nouns, pronouns include entities such as people, locations, or resources.
- *Tweet Length:* total number of characters in a post. The length of a post could indicate the amount of information [8,9,19].
- *Number of words:* similar to Tweet Length, number of words may also be an indicator of the amount of information [9,10].
- *Number of Hashtags:* these reflect the themes of a post, and are manually generated by the authors of the posts [8–10].
- *Unigrams:* the entire data (text of each post) is tokenised and represented as unigrams [8–10,13,21,25].

The spaCy library[7] is used to extract the Part Of Speech (POS) features (e.g., *nouns*, *verbs*, *pronouns*). Unigrams for the data are extracted with the regexp tokenizer provided in NLTK.[8] We removed stop words using a dedicated list.[9] Finally, we applied the TF-IDF vector normalisatiton on the unigrams in order to weight the importance of tokens in the documents according to their relative importance within the dataset, and represent the entire data as a set of vectors. This results in a vocabulary size in unigrams (for each language in the balanced data, combining test and train data) of *en*-7495, *es*-7121, and *it*-4882.

Semantic Features are curated to generalise the information representation of the crisis situations across various languages. Semantic features are designed to be broader in context and less crisis-specific, in comparison to the actual text of the posts, thereby helping to resolve the problem of data sparsity. To this

[7] SpaCy Library, https://spacy.io.

[8] http://www.nltk.org/_modules/nltk/tokenize/regexp.html.

[9] https://raw.githubusercontent.com/6/stopwords-json/master/stopwords-all.json.

end, we use the Named Entity Recognition (NER) service Babelfy[10], and two different knowledge bases for creating these features: BabelNet[11] and DBpedia[12]. Note that the semantics extracted by these tools are in English, and hence they bring the multilingual datasets a bit closer linguistically. The following semantic information is extracted:

- *Babelfy Entities*: Babelfy extracts the entities from each post in different languages (e.g., *news*, *sadness*, *terremoto*), and disambiguates them with respect to the BabelNet [15] knowledge base.
- *BabelNet Senses (English)*: for each entity extracted from Babelfy, the English labels associated with the entities are extracted (e.g. *news→news*, *sadness→sadness*, *terremoto→earthquake*).
- *BabelNet Hypernyms (English)*: for each entity, the direct hypernyms (at distance-1) are extracted from BabelNet and the main sense of each hypernym is retrieved in English. From our original entities, we now get *broadcasting*, *communication*, and *emotion*).
- *DBpedia Properties*: for each annotated entity we also get a DBpedia URI from Babelfy. The following properties associated with the DBpedia URIs are queried via 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`).

The inclusion of semantic features such as hypernyms has been shown to enhance the semantic and contextual representation of a document by correlating different entities, from different languages, with a similar context [12]. For example, the entities *policeman*, *policía* (Spanish for police), *fireman*, and *MP (Military Police)* all have a common hypernym (English): *defender*. By generalising the semantics in one language, English, we avoid the sparsity that often results from having various morphological forms of entities across different languages (see Table 3 for an example). Similarly, the English words *floods* and *earthquake* both have *natural disaster* as a hypernym, as does *inondazione* in Italian, ensuring that we know the Italian word is also crisis relevant. Adding the semantic information, through *BabelNet Semantics*, results in a vocabulary size in unigrams of: *en*-12604, *es*-11791, and *it*-8544.

Finally, we extract DBpedia properties of the entities (see Table 3) in the form of subject, label, and location-specific properties. This semantic expansion of the dataset forms the *DBpedia Semantics* component, and results in a vocabulary size in unigrams of: *en*-21905, *es*-15388, *it*-10674. The two types of semantic features (*BabelNet* and *DBpedia*) are used both individually and also in combination, to develop the binary classifier.

[10] http://babelfy.org.

[11] http://babelnet.org.

[12] http://dbpedia.org.

Table 3. Semantic expansion with BabelNet and DBpedia semantics

Feature	Post A	Post B
	'#WorldNews! 15 feared dead and 100 people could be missing in #Guatemala after quake http://t.co/uHNST8Dz'	*'Van 48 muertos por terremoto en Guatemala http://t.co/nAGG3SUi via @ejeCentral'*
Babelfy entities	*feared, dead, people, missing, quake*	*muertos, terremoto*
BabelNet sense (English)	*fear, dead, citizenry, earthquake*	*slain, earthquake*
BabelNet hypernyms (English)	*geological_phenomenon, natural disaster, group*	*geological_phenomenon, natural disaster, dead*
DBpedia properties	`dbr:Death`, `dbc:Communication`, `dbr:News`, `dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`	`dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`, `dbr:Death`
Google translation	To *es-' #Noticias del mundo! 15 muertos temidos y 100 personas podrían estar desaparecidas en #Guatemala después terremoto http://t.co/uHNST8Dz'*	To *en-'48 people killed by earthquake in Guatemala http://t.co/nAGG3SUi via @ejeCentral'*

3.3 Classifier Selection

In order to address the binary classification problem, the high dimensionality resulting from unigrams of tweets and semantic features, and the need to avoid overfitting, were taken into consideration. The training data instances (which varied between 1200–4500 under different experimental setups) were much smaller in size than the large dimensionality of the features (ranging between 9000–20000). We therefore opted for a Support Vector Machine (SVM) with a Linear Kernel [4] as the classification model. As discussed in [3,11], SVM performs better than other common approaches such as classification and regression trees (CART) and Naive Bayes in similar classification problems. The work by [3] also shows almost identical performance (in terms of accuracy) of SVM and CNN models in classification of the tweets. In [11], we showed the appropriateness of SVM Linear kernel over RBF kernel, Polynomial kernel, and Logistic Regression in such a classification scenario.

4 Cross-Lingual Classification of Crisis-Information

We demonstrate and validate our classification models through multiple experiments designed to test various criteria and models. We experiment on the models created with the following combinations of statistical and semantic features, thereby enabling us to assess the impact of each classification approach:

- *SF*: uses only the statistical features; this model is our baseline.
- *SF+SemBN*: combines statistical features with semantic features from *BabelNet* (entity sense, and their hypernyms in English, as explained in Sect. 3.2).
- *SF+SemDB*: combines statistical features with semantic features from *DBpedia* (label in English, type, and other DBpedia properties).
- *SF+SemBNDB*: combines statistical features with semantic features from *BabelNet* and *DBpedia*.

We apply and validate the models above in the following three experiments:

Monolingual Classification with Monolingual Models: In this experiment, we train the model on one language and test it on data in the same language. This tests the value of adding semantics to the classifier over the baseline when the language is the same.

Cross-lingual Classification with Monolingual Models: Here we evaluate the classifiers on crisis information in languages that were not observed in the training data. For example, we evaluate the classifier on Italian when the classifier was trained on English or Spanish.

Cross-lingual Classification with Machine Translation: In the third experiment, we evaluate the classifier when the model is trained on data in a certain language (e.g. Spanish), and used to classify information that has been automatically translated from other languages (e.g. Italian and English) into the language of the training data. The translation is performed using the Google Translate API.[13] To perform this experiment, we first translate the data from each of our three languages in turn into the other two languages.

All experiments are performed on both (i) the *unbalanced* dataset, to adhere to the natural distribution of these languages; and (ii) the *balanced* dataset, to remove bias towards any particular language which is caused by the uneven distribution of these languages in our datasets. By default, we refer to results from the balanced dataset unless we specifically mention the unbalanced one. Results are reported in terms of *Precision* (P), *Recall* (R), F_1 score (F_1), and ΔF_1 (% change over baseline $\frac{(semantic\ model\ F_1 - SF\ F_1)*100}{SF\ F_1}$, where $SF\ F_1$ is the F_1 score in SF model).

[13] https://cloud.google.com/translate/.

4.1 Results: Monolingual Classification with Monolingual Models

For the monolingual classification, a 5-fold cross validation approach was adopted and applied to individual datasets of *English*, *Italian*, and *Spanish*. Results in Table 4 show that adding semantics has no impact compared with the baseline (SF model) when the language of training and testing is the same.

Table 4. Monolingual Classification Models – 5-fold cross-validation (best F_1 score is highlighted for each model). *en*, *it*, and *es* refer to English, Italian, and Spanish respectively.

Unbalanced Data (from Table 1-*unbalanced*)																
		SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Test	Size	P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en	4358	0.833	0.856	0.844	0.84	0.858	0.849	0.59	0.826	0.844	0.835	-1.07	0.829	0.845	0.836	-0.95
it	1625	0.703	0.721	0.711	0.712	0.714	0.713	0.28	0.696	0.706	0.701	-1.4	0.702	0.715	0.708	-0.42
es	2163	0.801	0.808	0.804	0.812	0.809	0.810	0.75	0.799	0.795	0.797	-0.87	0.798	0.798	0.798	-0.75
Avg.				0.786			0.791	0.54			0.778	-1.1			0.781	-0.71

Balanced Data (from Table 1-*balanced*)																
		SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Test	Size	P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en	1224	0.832	0.830	0.831	0.835	0.805	0.820	-1.32	0.835	0.799	0.816	-1.80	0.829	0.808	0.818	-1.56
it	1224	0.690	0.729	0.709	0.703	0.722	0.712	0.42	0.689	0.716	0.702	-0.99	0.708	0.718	0.712	0.42
es	1224	0.798	0.765	0.781	0.794	0.783	0.789	1.02	0.779	0.754	0.766	-1.92	0.780	0.773	0.776	-0.64
Avg.				0.774			0.774	0.04			0.761	-1.57			0.769	-0.59

4.2 Results: Cross-Lingual Classification with Monolingual Models

This experiment involves training on data in one language and testing on another. Results, shown in Table 5, indicate that when using the statistical features alone (SF - the baseline), average F_1 is 0.557. When semantics are included in the classifier, average classification performance improvement (ΔF_1) is by 8.26%–9.07%, with a standard deviation (SDV) between 10.9%–13.86% across all three semantic models, for all the test cases. Similarly, when applied to *unbalanced* datasets, performance increases by 7.44%–9.78%.

While the highest gains are observed in SF+SemBNDB, the SF+SemBN seems to exhibit a consistent performance by improving over the SF baseline in 5 out of 6 cross-lingual classification tests, while SF+SemDB and SF+SemBNDB each show improvement in 4 out of 6 tests.

4.3 Results: Cross-Lingual Crisis Classification with Machine Translation

The results from cross-lingual classification after language translation are presented in Table 6. For each training dataset, we translate the test data into the

Table 5. Cross-Lingual Classification Models (best F_1 score is highlighted for each model).

Unbalanced Data (from Table 1- *unbalanced*)																	
		Size	SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Train	Test		P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en		4358															
	it	1625	0.576	0.522	0.417	0.598	0.562	0.518	24.2	0.595	0.576	0.553	32.6	0.609	0.588	0.568	36.2
	es	2163	0.674	0.633	0.604	0.663	0.654	0.645	6.79	0.653	0.649	0.643	6.46	0.649	0.641	0.633	4.8
it		1625															
	en	4358	0.469	0.474	0.449	0.547	0.545	0.538	19.82	0.508	0.508	0.504	12.25	0.516	0.516	0.516	14.9
	es	2163	0.635	0.610	0.586	0.643	0.627	0.612	4.43	0.601	0.60	0.596	1.70	0.625	0.620	0.614	4.78
es		2163															
	en	4358	0.633	0.62	0.604	0.60	0.572	0.532	-11.9	0.623	0.618	0.610	0.99	0.606	0.592	0.571	-5.46
	it	1625	0.536	0.533	0.521	0.529	0.529	0.528	1.34	0.526	0.526	0.526	0.96	0.539	0.539	0.539	9.78
Avg.					0.530			0.562	7.44			0.572	9.16			0.573	9.78
SDV					0.082			0.053	13.08			0.053	12.3			0.044	14.47
Balanced Data (from Table 1-*balanced*)																	
		Size	SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Train	Test		P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en		1224															
	it	401	0.539	0.515	0.429	0.588	0.571	0.549	28	0.569	0.568	0.568	32.4	0.578	0.576	0.572	33.3
	es	401	0.689	0.688	0.688	0.669	0.668	0.668	-2.9	0.647	0.644	0.641	-6.8	0.666	0.661	0.659	-4.2
it		1224															
	en	401	0.521	0.521	0.521	0.581	0.581	0.580	11.3	0.558	0.552	0.539	3.5	0.550	0.546	0.538	3.3
	es	401	0.655	0.646	0.640	0.672	0.655	0.647	1.1	0.638	0.636	0.635	-0.78	0.637	0.633	0.631	-1.4
es		1224															
	en	401	0.609	0.593	0.578	0.657	0.620	0.597	3.3	0.667	0.666	0.665	15	0.660	0.653	0.650	12.4
	it	401	0.529	0.522	0.489	0.534	0.534	0.532	8.8	0.551	0.546	0.533	9	0.555	0.551	0.543	11
Avg.					0.557			0.596	8.26			0.597	8.71			0.599	9.07
SDV					0.096			0.053	10.94			0.057	13.86			0.054	13.6

language of the training data. For example, when the training data is in English (*en*), the Italian data is translated to English, and is represented in the table as *it2en*. We aim to analyse two aspects here: (i) how semantics impacts the classifier on the translated content; and (ii) how the classifiers perform over the translated data in comparison to cross-lingual classifiers, as seen in Sect. 4.2.

From the results in Table 6, we see that based on average % change ΔF_1 of all translated test cases (*en2it*,*es2it*, etc.), SF+SemBN outperforms the statistical classifier (SF) by 3.75% (balanced data) with a standard deviation (SDV) of 4.57%. However, the other two semantic feature models (SF+SemDB and SF+SemBNDB) do not improve over the statistical features when the test and training data are both in the same language (after translation). The SF+SemBN shows improvement in 4 out of 6 translated test cases, except when trained on Spanish (*es*).

Comparing the best performing model from translated data, i.e. SF+SemBN, and overall baseline (SF model from cross-lingual classification Table 5-*balanced*), the SF+SemBN (translation) has an average F_1 gain (ΔF) across each translated test case over the baseline of 15.23% (with a SDV 12.6%). For example, compare

Table 6. Cross-Lingual Crisis Classification with Machine Translation (best F_1 score is highlighted for each event).

Unbalanced Data (from Table 1- *unbalanced*)

		Size	SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Train	Test		P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en		4358															
	it2en	1625	0.644	0.613	0.591	0.635	0.611	0.593	0.34	0.582	0.568	0.548	-7.27	0.597	0.580	0.561	-5.0
	es2en	2163	0.681	0.681	0.681	0.667	0.667	0.667	-2.0	0.669	0.661	0.659	-3.2	0.664	0.661	0.660	-3.1
it		1625															
	en2it	4358	0.609	0.601	0.588	0.636	0.618	0.597	1.53	0.570	0.570	0.569	-3.2	0.575	0.574	0.571	-2.9
	es2it	2163	0.647	0.629	0.612	0.675	0.636	0.607	-0.81	0.609	0.595	0.578	-5.5	0.620	0.603	0.583	-4.7
es		2163															
	en2es	4358	0.643	0.626	0.609	0.661	0.634	0.610	0.16	0.654	0.654	0.653	7.2	0.649	0.648	0.646	6.07
	it2es	1625	0.585	0.584	0.583	0.590	0.590	0.589	1.03	0.581	0.580	0.580	-0.51	0.586	0.585	0.584	0.17
Avg.					0.611			0.611	0.03			0.598	-2.1			0.60	-1.6
SDV					0.036			0.029	1.3			0.046	5.1			0.04	4.2

Balanced Data (from Table 1-*balanced*)

		Size	SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Train	Test		P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en		1224															
	it2en	401	0.624	0.583	0.546	0.622	0.598	0.577	5.7	0.561	0.558	0.554	1.46	0.594	0.588	0.581	6.4
	es2en	401	0.675	0.671	0.669	0.704	0.696	0.693	3.6	0.701	0.671	0.658	-1.6	0.695	0.674	0.664	-0.74
it		1224															
	en2it	401	0.583	0.578	0.572	0.639	0.631	0.625	9.3	0.547	0.546	0.545	-4.7	0.551	0.551	0.551	-3.6
	es2it	401	0.638	0.621	0.609	0.703	0.668	0.653	7.2	0.619	0.603	0.590	-3.1	0.610	0.596	0.582	-4.4
es		1224															
	en2es	401	0.686	0.678	0.675	0.691	0.670	0.661	-2.0	0.691	0.691	0.691	2.3	0.683	0.683	0.683	1.2
	it2es	401	0.594	0.594	0.593	0.586	0.586	0.586	-1.2	0.580	0.576	0.570	-3.9	0.579	0.576	0.571	-3.7
Avg.					0.610			0.633	3.75			0.601	-1.59			0.605	-0.83
SDV					0.052			0.045	4.57			0.059	2.9			0.054	4.14

ΔF between *it*-*en2it* in SF+SemBN in the translated model and *it*-*en* in SF in the cross-lingual model, similarly for the other 5 test cases. Based on an average of ΔF across all the test cases, the SF+SemBN (from translation) and SF+SemBN (from cross-lingual models), both perform well over the baseline (SF from cross-lingual model), by 8.26% and 15.23% respectively.

4.4 Cross-Lingual Ranked Feature Correlation Analysis

To understand the impact of the semantics and the translation on the discriminatory nature of the cross-lingual data from different languages, we analysed the correlation between ranked features of each dataset under different models. For this, we considered the *balanced* datasets across each language and took the entire data by merging the training and test data for each language. Next, we calculated Information Gain over each dataset (English (*en*), Spanish, (*es*), and Italian (*it*)), across all 4 models (SF, SF+SemBN, SF+SemDB, SF+SemBNDB). We also calculated Information Gain over the translated datasets (*en2it*, *en2es*, *es2it*, *es2en*, *it2en*, and *it2es*). This provides the ranked list of features, in terms of their discriminatory powers in the classifiers, in each selected dataset.

Table 7. Spearman's Rank Order Correlation between ranked informative features (based on IG) across models and languages

Data/Model	SF	SF+SemBN	SF+SemDB	SF+SemBNDB	Translation	
$en - es$	0.573	0.385	0.349	0.373	0.515(en-es2en)	0.449(es-en2es)
$en - it$	−0.179	0.402	0.111	0.315	0.266(en-it2en)	0.594(it-en2it)
$es - it$	0.418	0.222	0.503	0.430	0.678(es-it2es)	0.612(it-es2it)

For each pair of datasets, such as English (*en*) - Spanish (*es*), we consider the common ranked features with $IGscore > 0$, and calculate the Spearman's Rank Order Correlation (ranges between $[-1, 1]$) across the two ranked lists. For the translated data, we analysed pairs where one dataset is translated to the language of another dataset, such as *en-it2en* and *it-en2it*.

Table 7 shows how the correlation varies across the data. These variations can be attributed to a number of aspects. The overlap of crisis events while sampling the data is a crucial parameter, as the data was sampled based on language, and the discreteness of the source events (Table 2) was not taken into consideration. This can particularly be observed in the *en-es* correlation, where the highest correlation is without the semantics. This also explains the better performance of the SF model over the *semantic* models when trained on *en* and evaluated on *es* (Table 5-*balanced*). The correlation between *en-it* is ˜–0.179, which indicates nearly 'no correlation'. The increase in discriminative-feature correlations between datasets once semantics are added is in part due to the extraction of semantics in English (see Sect. 3.2), thus bringing the terminologies closer semantically as well as linguistically.

Translating the data to the same language shows an increment in the correlation. This is expected for multiple reasons. Firstly, having the data in the same language enables the identification of more similar features such as verbs and adjectives across the datasets. Secondly, given the similarity in the different types of events covered under the three languages, such as *floods* and *earthquakes*, the nature of the information is likely to have a high contextual overlap.

5 Discussion and Future Work

Our aim is to create hybrid models, by mixing semantic features with the statistical features, to produce a crisis data classification model that is largely language-agnostic. The work was limited to English, Spanish, and Italian, due to the lack of sufficient data annotations in other languages. We are currently designing a CrowdFlower annotation task to expand our annotations to several other languages.

We ran our experiments on both *balanced* and *unbalanced* datasets. However, performance over the balanced dataset provides a fairer comparison, since biases towards the dominant languages are removed. We also experimented with classifying data in their original languages, as well as automatically translating the data into the language of the training data. Results show that with balanced datasets, translation improves the performance of all classifiers, and reduces the benefits of using semantics in comparison to the statistical classifier (SF; the baseline). One could conclude that if the data is to be translated into the same language that the model was trained on, then the statistical model (SF) might be sufficient, whereas if translation is not viable (e.g., data arriving in unpredicted languages, or where translations are too inaccurate or untrustworthy) then the model that mixes statistical and semantics features is recommended, since it produces higher classification accuracies.

In this work, the classifiers were trained and tested on data from various types of crisis events. It is natural for some nouns to be identical across various languages, such as names of crises (e.g. Typhoon Yolanda), places, and people. In future work we will measure the level of terminological overlap between the datasets of different languages.

We augmented all datasets with semantics in English (Sect. 3.2). This is mainly because BabelNet (version 3.7) is heavily biased towards English[14]. Most existing entity extractors are skewed heavily towards the English language, and hence as a byproduct of adding their identified semantics, more terms (concepts) in a single language (English) will be added to the datasets. As a consequence, this will bring the datasets of different languages closer together linguistically, thus giving an advantage to semantic models over purely statistical ones in the context of cross-lingual analysis. We performed a comparison of vocabulary similarity between the language datasets, before and after the addition of semantics, to also comprehend the overlapping of the vocabulary. For instance, the cosine similarity between (without semantics) *en-it* is 0.311, *en-es* is 0.536, and *it-es* is 0.32. Adding semantics increased the cosine similarity across all the datasets. In current experiments, we had 6 test cases in each classification model; despite the consistency observed across 6 cross-lingual test cases, we would need more observations to establish that the gain achieved by the semantic models over the baseline models is statistically significant. Repeating these experiments over more languages should help in this; alternatively, creating multiple train and test splits for each test case could also complement such analysis, which was not feasible in this study due to insufficient data to create multiple splits for each dataset. However, we did perform a 10 iteration of 5-fold cross validation over the entire dataset across all the feature sets and found that SF+SemBN(*BabelNet Semantics*) model outperformed all others (particularly baseline with a statistically significant value of $p = 0.0192$, on a two-tailed t-test).

In this work, we experimented with training the model on one language at a time. Another possibility is to train the model on multiple languages, thus increasing its ability to classify data in those languages. However, generating such a multilingual model is not always feasible, since it requires annotated data in all the languages it is intended to analyse. Furthermore, the need for models that can handle other languages is likely to remain, since the language of data shared on social media during crises tends to differ substantially, depending on

[14] Almost 17 million word senses in English, next highest is French, with 7 million senses http://live.babelnet.org/stats.

where these crises are taking place. Therefore, the ability of a model to classify data in a new language will always be a clear advantage. The curated data (with semantics) and code, in this work, is being made available for research purposes.[15]

6 Conclusion

Determining which tweets are relevant to a given crisis situation is important in order to achieve a more efficient use of social media, and to improve situational awareness. In this paper, we demonstrated the ability of various models to classify crisis related information from social media posts in multiple languages. We tested two approaches: (1) adding semantics (from *BabelNet* and *DBpedia*) to the datasets; and (2) automatically translating the datasets to the language that the model was trained on. Through multiple experiments, we showed that all our semantic models outperform statistical ones in the first approach, whereas only one semantic model (using BabelNet) shows an improvement over the statistical model in the second approach.

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

References

1. Araujo, M., Reis, J., Pereira, A., Benevenuto, F.: An evaluation of machine translation for multilingual sentence-level sentiment analysis. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 1140–1145. ACM (2016)
2. Burel, G., Saif, H., Alani, H.: Semantic wide and deep learning for detecting crisis-information categories on social media. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10587, pp. 138–155. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68288-4_9
3. Burel, G., Saif, H., Fernandez, M., Alani, H.: On semantics and deep learning for event detection in crisis situations. In: Workshop on Semantic Deep Learning (SemDeep) at ESWC (2017)
4. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, New York (2000)
5. Derczynski, L., Meesters, K., Bontcheva, K., Maynard, D.: Helping crisis responders find the informative needle in the tweet haystack. arXiv preprint arXiv:1801.09633 (2018)
6. Deriu, J., et al.: Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 1045–1052 (2017)

[15] https://github.com/pkhare/iswc_codebase.

7. Gao, H., Barbier, G., Goolsby, R.: Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intell. Syst. **26**(3), 10–14 (2011)
8. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.: Extracting information nuggets from disaster-related messages in social media. In: ISCRAM (2013)
9. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.: Practical extraction of disaster-relevant information from social media. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1021–1024. ACM (2013)
10. Karimi, S., Yin, J., Paris, C.: Classifying microblogs for disasters. In: Proceedings of the 18th Australasian Document Computing Symposium, pp. 26–33. ACM (2013)
11. Khare, P., Burel, G., Alani, H.: Classifying crises-information relevancy with semantics. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 367–383. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_24
12. Khare, P., Fernandez, M., Alani, H.: Statistical semantic classification of crisis information. In: Workshop on HSSUES at ISWC (2017)
13. Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.C.: TEDAS: a twitter-based event detection and analysis system. In: 2012 IEEE 28th International Conference on Data Engineering (ICDE), pp. 1273–1276. IEEE (2012)
14. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 976–983 (2007)
15. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
16. Olteanu, A., Vieweg, S., Castillo, C.: What to expect when the unexpected happens: social media communications across crises. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 994–1009. ACM (2015)
17. Power, R., Robinson, B., Colton, J., Cameron, M.: Emergency situation awareness: twitter case studies. In: Hanachi, C., Bénaben, F., Charoy, F. (eds.) ISCRAM-med 2014. LNBIP, vol. 196, pp. 218–231. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11818-5_19
18. Rogstadius, J., Vukovic, M., Teixeira, C., Kostakos, V., Karapanos, E., Laredo, J.A.: CrisisTracker: crowdsourced social media curation for disaster awareness. IBM J. Res. Dev. **57**(5), 4-1–4-13 (2013)
19. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
20. Severyn, A., Moschitti, A.: UNITN: training deep convolutional neural network for twitter sentiment classification. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 464–469 (2015)
21. Stowe, K., Paul, M.J., Palmer, M., Palen, L., Anderson, K.: Identifying and categorizing disaster-related tweets. In: Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, pp. 1–6 (2016)
22. Tonon, A., Cudré-Mauroux, P., Blarer, A., Lenders, V., Motik, B.: `ArmaTweet`: detecting events by semantic tweet analysis. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10250, pp. 138–153. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58451-5_10

23. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1079–1088. ACM (2010)
24. Wick, M., Kanani, P., Pocock, A.C.: Minimally-constrained multilingual embeddings via artificial code-switching. In: AAAI, pp. 2849–2855 (2016)
25. Zhang, S., Vucetic, S.: Semi-supervised discovery of informative tweets during the emerging disasters. arXiv preprint arXiv:1610.03750 (2016)