Statistical Semantic Classification of Crisis Information

Conference or Workshop Item

Statistical Semantic Classification of Crisis Information

Prashant Khare, Miriam Fernandez, and Harith Alani

Knowledge Media Institute, Open University, UK,
{prashant.khare,miriam.fernandez,h.alani}@open.ac.uk

Abstract. The rise of social media as an information channel during crisis has become key to community response. However, existing crisis awareness applications, often struggle to identify relevant information among the high volume of data that is generated over social platforms. A wide range of statistical features and machine learning methods have been researched in recent years to automatically classify this information. In this paper we aim to complement previous studies by exploring the use of semantics as additional features to identify relevant crisis information. Our assumption is that entities and concepts tend to have a more consistent correlation with relevant and irrelevant information, and therefore can enhance the discrimination power of classifiers. Our results, so far, show that some classification improvements can be obtained when using semantic features, reaching +2.51% when the classifier is applied to a new crisis event (i.e., not in training set).

Keywords: semantics, crisis informatics, tweet classification

1 Introduction

As per the 2016 World Humanitarian Data and Trends report by UNOCHA,[1] there were around 102 million people, from 114 countries, affected by natural disasters in the year of 2015 alone, causing an estimated damage of $90 billion. During such disasters there is normally a surge of real time content across multiple social media platforms. For example, during the 2011 Japan earthquake, there was a 500% increase in the number of tweets.[2] All these messages constitute a critical source of information for relief and search teams, communities and individuals.

However, it is almost impossible to manually absorb and process the sheer volume of social media reports generated during a crisis, and to efficiently filter any relevant and actionable information [5]. Tools to automatically identify relevant information are largely unavailable, and the characteristics of social media messages (short length, use of colloquialisms, ill-formed words and syntactic structures) increases the challenges of automatically processing and understanding of such messages.

[1] https://data.humdata.org/dataset/world-humanitarian-data-and-trends
[2] https://blog.twitter.com/official/en_us/a/2011/global-pulse.html

Much research explored methods for the classification of social media data into crisis-related or unrelated, based on supervised [10,8,16,20] and unsupervised [14] machine learning (ML) methods. These methods tend to identify relevant data based on n-grams and statistical features (message length, URLs, Hashtags, etc.). This paper aims to complement previous works by investigating the impact of semantic features to identify relevant information from Twitter data during crisis situations. The semantic features explored in our work include entities (e.g., "London", "Colorado", "Fire") extracted from tweets, as well as their hypernyms from BabelNet, which is an external knowledge base[11]. Our hypothesis is that entities and concepts may have a more consistent correlation with relevant and irrelevant crisis information, and therefore can be used to better interpret the content of the tweets and to enhance the discrimination power of classifiers.

We explore the effectiveness of semantic features by creating and testing classifiers to identify relevant crisis information, as well as by testing these classifiers with previously unseen information from different crisis events. The dataset used in our research is a small subset of CrisisLexT26;[3] a library of 205K annotated tweets posted during 26 real crisis events in 2012 and 2013. Our subset consists of a balanced related-unrelated set of 3.2K tweets on 9 crisis events (detailed in Section 3.1). Our results show that using semantic information can indeed help to enhance classification results, but only by a small margin. When the classifier is applied to a new crisis event, results show that the use of semantic annotations of concepts and entities in itself is effective, and the use of semantically expanded concepts (i.e., entities and their hypernyms) further improves over it slightly. However, the use of hypernyms also sometimes introduces generic concepts, such as "person", that appear in both, crisis related and non-crisis related posts, and thus effects the discrimination power of semantic features.

The contributions of this work can be summarised as follows:

- Demonstrating the impact of using a variety of semantic features for identifying crisis-related information from social media posts.
- Showing that adding semantic features is especially useful when classifying new crisis events that were not seen during the model training phase.
- Testing using annotated data from CrisisLexT26 of 9 real crisis events.
- Discussing and reflecting on the potential use of semantics to identify crisis-relevant information.

The rest of the paper is structured as follows. Section 2 summarises the related work on processing social media data for identifying crisis related content. Section 3 describes our approach, including the selected semantic features and how they are used to created various types of classifiers. The experiments are results are reported in Section 4. Section 5 discusses the lessons learned from this work, as well as its limitations and the future lines of work. Conclusions are reported in Section 6.

[3] `crisislex.org`

2 Related Work

During a crisis, a very large number of messages are often posted on various social media platforms. Processing all such messages requires substantial time and effort to ensure that crisis related messages are efficiently spotted and handled, since a good percentage of messages posted about a crisis tends to be irrelevant and unrelated. Olteanu and colleagues observed that crisis reports could be classified into three main categories: related and informative, related but not informative, and not related [12]. In this work, we focus primarily on the automatic identification of crisis related information. The identification of informativeness in crisis scenarios is a complex task that requires a deeper reflection and investigation of the meaning of informativeness and its dimensions (freshness, novelty, location, scope). It is therefore an important part of our future work.

To identify crisis related messages from social media data, several works have proposed the use of supervised [10,8,16,20] and unsupervised [14] ML classification methods. Supervised methods tend to make use of n-grams as well as of linguistic and statistical features such as part of speech (POS), number of hashtags, mentions, or message length. They also highlight the use of location as an important indicator, since people tend to create and retweet messages with locally actionable information [9]. These works make use of various supervised classification methods, from traditional classification algorithms such as Naive Bayes, Support Vector Machines or Conditional Random Field [13,16,6] to more novel techniques such as deep learning [3]. Unsupervised methods, on the other hand, are mainly based on keyword processing and clustering [14]. Our work aims to complement these studies by investigating the use of semantics, and particularly the use of entities extracted from tweets, and their hypernyms, as additional features to boost classification. As previously done by [8] we not only aim to generate classifiers able to identify crisis-related information, but we also aim to test the generated classifiers on crisis events that the classifiers have not previously seen.

While semantic models have been developed and used to represent and capture the information that emerge from crisis events (e.g., MOAC - Management of a Crisis [4], or HXL - Humanitarian eXchange Language[5]), few works in the literature have explored the use of semantics to identify and filter crisis-related information. In [2], Abel and colleagues presented Twitcident, a system that uses semantic information to facilitate filtering and search of crisis related information. The system extracts semantic information from social media data in the form of entities using Name Entity Recognisers (NER) and external knowledge bases. However, as opposed to our work, they do not explore the use of entities as features for classification. Instead, they develop similarity models in which the crisis event and the posts are profiled based on this semantic enrichment, and the Jaccard similarity coefficient[6] is used to compute whether the content of the posts is similar or not to the event.

[4] `http://www.observedchange.com/moac/ns.`
[5] `http://hxlstandard.org/`
[6] `https://en.wikipedia.org/wiki/Jaccard_index`

3 Classification: Identifying Crisis Related Information

Our approach for identifying crisis related information among tweets functions by generating binary classifiers to differentiate crisis-related from non-related posts. In this section, we explain (i) the dataset used in our experiments, (ii) the two set of features (statistical and semantic) that we use to build the classifiers, and (iii) our classifier selection process.

3.1 Data Selection

To conduct our study we selected the CrisisLexT26[7] dataset [12], an annotated dataset of 205K tweets posted during 26 crisis events occurring between 2012 and 2013. The search keywords used to construct CrisesLexT26 were selected following the standard practices of hashtags and/or terms often paired with canonical forms of a disaster name and impacted location (e.g., Queensland floods) or a meteorological term (e.g., Hurricane Sandy). For each of the 26 crisis events, around 1,000 tweets are annotated (Related and Informative, Related but not Informative, Not Related, or Not Applicable). Given our focus on English tweets, we selected 9 events for which the content was predominantly in English: *West Texas Explosion*(WTE), *Colorado WildFire*(CWF), *Colorado Flood*(CFL), *Australia Bushfire*(ABF), *Boston Bombing*(BB), *LA Shooting*(LAS), *Queensland Flood*(QFL), *Savar Building Collapse* (SBC), and *Singapore Haze*(SGH).

We merged those tweets labelled as *Not Related* and *Not Applicable* under the class *Not Related*, obtaining a total of 1539 non crisis-related tweets. We also merged those tweets labelled as *Related and Informative* and *Related but not Informative* under the class *Related*, obtaining a total of 7461 crisis related tweets. In line with common practice, we balanced the dataset to remove classification bias towards the bigger class *Related*, by randomly selected 1667 crisis related tweets. This gives us a balanced and annotated dataset of 3206 of *Related* and *Not Related* tweets.

3.2 Feature Engineering

To generate classifiers able to identify crisis-related posts, we explore two distinct feature sets, statistical and semantic features. Statistical features have been widely studied in the literature [10,8,16,20] and are used as the baseline for our experiments. They capture the linguistic and quantifiable attributes of posts. Semantic features, on the other hand, capture the different named entities that emerge from tweets, as well as their hierarchical information which we extract from an external knowledge source.

3.2.1 Statistical Features (SA) For each social media post, we extract the following statistical features:

– Number of nouns: nouns generally refer to locations, resources, or actors involved in the crisis event.

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

– Number of verbs: verbs are an indication of the different actions that are occurring during the crisis event.
– Number of pronouns: as with nouns, pronouns may be an indication of the actors, locations, or resources that are named during the crisis event.
– Tweet Length: number of characters contained in the posts. The longer the post is, the higher the amount of information it may contain.
– Number of words: number of words may be another indication of the amount of information the post may have.
– Number of Hashtags: hashtags indicate the themes of the post and are manually generated by the posts' authors.
– Readability: Gunning fog index using average sentence length (ASL) and the percentage of complex words (PCW): 0.4 * (ASL + PWC). This feature gauges how hard the post is to parse by humans.[8]
– Unigrams: unigrams provide a keyword-based representation of the content of the posts

To extract the unigrams from social media posts we make use of the Weka data mining software[9], and specifically its StringToWord functionality, including lower case conversion for all tokens, stemming (using Lovins' algorithm)[10], stopword removal, and tf*idf transformation. The total number of unigrams, or vocabulary size, for the complete dataset is 10655. To extract the Part of Speech (POS) tags and the statistical features listed above (top five), we make use of the Stanford Core NLP software.[11] Hashtags are identified by the use of the # character, and readability is computed using the Gunning fog index.

3.2.2 Semantic Features (SemF) The semantic feature extraction process is summarised in Figure 3.2.2 and consists of three main steps: (i) *semantic annotation*, (ii) *semantic expansion*, and (iii) *semantic filtering*. Each of these three steps generates a different set of semantic features that we explore, individually and in combination, when generating binary classifiers to distinguish crisis-related posts from unrelated ones.

Semantic Annotation Features (SemAF): In the initial step (semantic annotation) semantic entities are extracted from the posts by using Babelfy.[12] This Name Entity Recogniser (NER) identifies the different entities that appear in the text, disambiguates them, and links them to the BabelNet[11] knowledge base, providing a unique identifier (SynsetID) for each of the identified entities. For example (Figure 1), for the post "*A 15-year-old High River boy is missing due to the flood. Call police if you see Eric St. Denis #abflood*" Babelfly identifies entities such as High River, Boy, Flood, etc. The annotation of the entire dataset (see Section 3.1) resulted in 12,006 unique concepts.

[8] `https://en.wikipedia.org/wiki/Gunningfogindex`
[9] `http://www.cs.waikato.ac.nz/ml/weka/`
[10] `http://www.mt-archive.info/MT-1968-Lovins.pdf`
[11] `https://stanfordnlp.github.io/CoreNLP/`
[12] http://babelfy.org

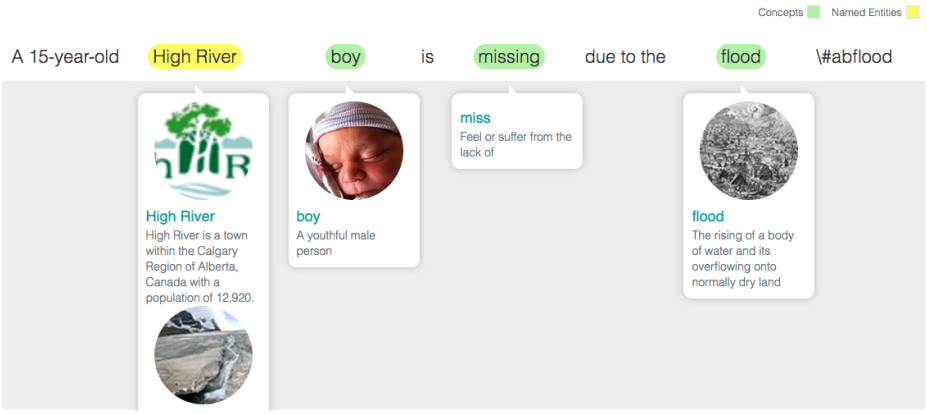

Fig. 1. Example of a semantically annotated post, generated with Babelfly.

Semantic Expansion Features (SemEF): In the second step (semantic expansion) the BabelNet knowledge base is used to extract every direct hypernym (distance-1) of these entities. Our hypothesis for considering hypernyms is that, by introducing upper level concepts, we might be able to better encapsulate the semantics of crisis-related tweets. For example, if the entities *fireman* and *policeman* appear often in crisis related posts. These entities have a common hypernym, *defender*. As a result, a post with the entity *MP (Military Police)*, is more likely to also be crisis-related since this entity also has the hypernym *defender*. The semantic expansion process resulted in an additional 7032 unique concepts.

Semantic Filtering Features (SemFF): When semantically expanding the initially extracted entities, we could sometimes introduce very generic concepts with low discrimination power. For example, the hypernym *Person* appears in both crisis and non-crisis related posts, and thus does not help the classifiers to identify crisis-related information. Our filtering process aims to discard such semantic annotations that might be too generic and hence are likely to reduce the discrimination power of semantics. Our proposed filtering process is based on the computation of the depth of a concept in the hierarchy of BabelNet. To determine the depth of concepts, we query iteratively through the hierarchy of BabelNet. Abstract concepts, i.e., concepts with a lower depth are therefore removed. To determine the shortest depth of a concept in the hierarchy of BabelNet, we used nearly 4 million relations extracted by iteratively querying for hypernyms and generated a Directed Graph. The node with highest *betweenness centrality* (SynSetID 'bn:00031027n', which relates to the main sense 'Entity') was determined to be the most abstract concept. The NetworkX[13] graph library for Python was used for this task. We then computed the *Shortest path* between the node 'Entity' and all the extracted hypernyms. The maximum depth found was 21, where level 0 is assigned to the concept 'Entity'. By performing an em-

[13] `https://networkx.github.io/`

pirical analysis of the concepts using Information Gain, we observed that the most informative concepts are those whose depth is between 3 and 7. Those are therefore the ones selected as features for classification. This filtering process resulted in 574 concepts filtered out from the semantics across 9 events.

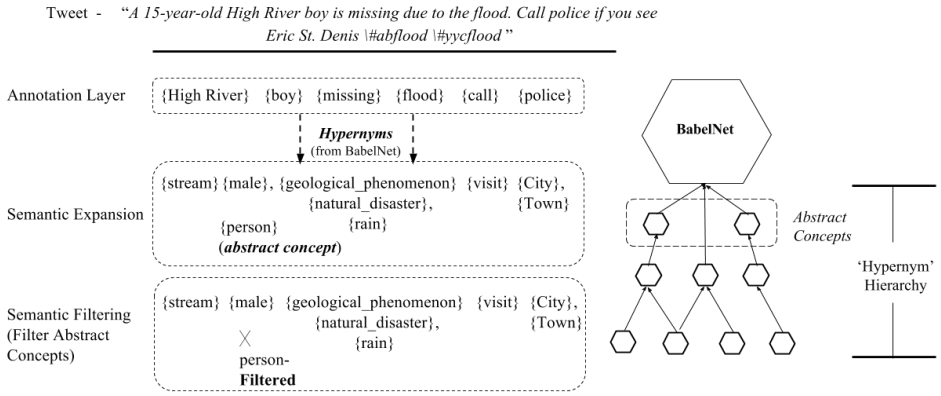

Fig. 2. Semantic Features: Annotation, Expansion, & Filtering

3.3 Classifier Selection

When selecting a classification method for the problem at hand we considered the high dimensionality of features, particularly the high number of unigrams and semantic features, the limited set of labelled data (3,206 posts) and the importance of avoiding over-fitting. Given the large set of features in comparison with the number of training examples, we opted for selecting the Support Vector Machine (SVM) classification model [4] with Linear Kernel. SVM has proven effective for problems with these characteristics.[14]

4 Experiments

In this section we describe our experimental set up, and particularly the design of our model selection and testing experiments. We report on the obtained results and later discuss how semantic features can help enhancing the performance of classifiers based on statistical features, and especially when the classifier is applied to cross-crisis scenarios.

4.1 Experimental Setup

We designed two main experiments where we train and test our classification models on (i) all 9 crisis events, and (ii) on 8 events, and retest on the 9th event, i.e., cross-crisis testing.

[14] `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf` Radial Basis Function (rbf) kernel or a Polynomial Kernel may cause an over-fitting problem,hence we opted for a linearly separable hyperplane.

– Crisis Classification Model: In our first experiment we compare the performance of classifiers generated with statistical features vs. classifiers enhanced with semantic features and analyse if the use of semantics does indeed help boosting the performance of binary classifiers when identifying crisis-related posts. We compare the performance of four different classifiers generated using the complete dataset (see Section 3.1), and tested using 10-fold cross validation. We use the WEKA software (v.3.8)[15] to generate the classifiers.
  - SF: A classifier generated with statistical features; our baseline.
  - SF+SemAF: A classifier generated with statistical features, and semantic annotation features.
  - SF+SemAF+SemEF: A classifier generated with statistical features, semantic annotations, and their hypernyms, i.e., the Semantic Expanded Features.
  - SF+SemFF: A classifier generated with statistical features, and filtered semantic annotations, along with their hypernyms, i.e., the Semantic Filtered Features
– Cross-crisis Classification: In our second experiment we retest the classifiers above by applying them to a new crisis data, i.e., on data from a new crisis event that was not part of the training set. For this experiment, we generate the same four classifiers described in the previous task. However, rather than using the complete dataset to generate the model, we use 8 out of the 9 crisis events to generate the model, and then apply the models to the remaining event for validation. We therefore generate 36 different classification models for this experiment.

4.2 Results: Crisis Classification

The results of our first experiment (each model statistically evaluated with 10 iterations of 10-fold cross validation) are presented in Table 1. The table presents F-measure(F) value (from 10-fold cross validation), mean of F-measure (F_{mean}) of 100 results from 10 iterations, standard deviation in F-measure (σ),and the increment of F_{mean} over the baseline $\Delta F/F$. Precision and Recall values where equal to F in this experiment, and hence were omitted from table. As we can see in this table, the use of semantic features helps to enhance classification results in all cases, but almost negligibly (less than 0.6%). However, the use of annotations alone (SF+SemAF) produces slightly better results than the use of annotations and hypernyms (SF+SemAF+SemEF).

To better understand the impact of semantics in this context, we manually analyse some of the tweets that were *misclassified* by the statistical baseline model, but were correctly classified when using semantics (see Table 2)In addition, we perform feature selection using Information Gain (IG) over the generated classifiers to determine which are the most discriminative statistical and semantic features when identifying crisis related posts.

[15] `http://www.cs.waikato.ac.nz/ml/weka/`

Features	F	F_{mean}	Std. Dev. σ	$\Delta F/F$
SF (Baseline)	0.865	0.872	0.017	-
SF+SemAF	0.870	0.877	0.017	0.0057
SF+SemAF+SemEF	0.868	0.873	0.017	0.0011
SF+SemFF	0.864	0.873	0.018	0.0011

Table 1. 10 iterations of 10-fold Cross Validation $\Delta F/F$, showing performance of our statistical semantics classifiers against the statistical baseline classifier.

PostID	Text	Label
Post1	I GET 5078 REALL FOLLOWERS! http://t.co/qrF5dpD3 #BestRap,#boulderflood,#PutinsFlik,#Rem #in	Not Related
Post2	@Stana_Katic Can we get some loveballs in Colorado? We need it after all the flooding! Love you! Xo	Not Related
Post3	RT @LarimerCounty: #HighParkFire burn area map as of Monday night 10 p.m. http://t.co/1guBTcXX	Related
Post4	Colorado wildfires their worst in a decade http://t.co/RtfLmfds	Related
Post5	RT @RedCross: Thanks to generosity of volunteer blood donors there is currently enough blood on the shelves to meet demand. #BostonMarathon	Related

Table 2. Examples of posts that were misclassified by the statistical classifier, but classified correctly by the semantic classifiers.

When applying IG over the attributes of the baseline classifier, the *number of hashtags* was the most relevant feature. By manually checking some of the tweets, we observe that *Not Related* posts tend to either have no hashtags (see as example Post2) or contain many hashtags (see Post1). The number of nouns and pronouns is also a high discriminative feature. As we hypothesized, crisis related posts generally contain more nouns and pronouns mentioning persons, resources or locations relevant to the crisis event. When including semantics, we observe that the hypernyms and annotations are among the highly ranked features, based on IG. Apart from highly ranked statistical features, hypernyms such as '*Happening*' and '*Event*' (which, in BabelNet, are hypernyms of concepts such as 'Incident', 'Fire', 'Crisis','Disaster',and 'Death'), were among the top 10 attributes (out of almost 800 positive IG attributes).

Post3 was misclassified when using only statistical features. Although it contains the relevant term *burn*, it barely appears in the training data. However, the post is correctly classified by SF+SemAF, because the term *burn* returns the concept *Fire* as part of its semantic annotation. Post 4 was misclassified by SF+SemAF, but correctly classified when adding semantic expansion (SF+SemAF+SemEF).The original tweet was annotated with the concept *Wildfire*, which has the hypernym *Fire*; a feature with high IG and strongly associated with the class crisis-related. Therefore, in this case, the use of hypernyms helped to obtain the additional information needed to correctly categorise the post.

Post 5 was misclassified by SF+SemAF+SemEF but correctly classified by SF+SemFF. Annotations such as *Thanks* and *Meet* semantically expanded to hypernyms such as *Virtue*, and *Desire*, which have a very low discrimination

power, and hence weakens the classifier. We observe that removing such less informative abstract concepts results in increasing the discriminative power of the remaining, more informative, concepts, such as '*Volunteer*' and hypernym (of 'donor') '*Benefactor*'.

4.3 Results: Cross-crisis Classification

The results of this experiment are reported in Table 3. In this experiment, we compile 9 different datasets, where in each dataset 1 out of the 9 crisis events is entirely left out of the training sample used to train and test the classification model.[16] Each row is named after the the crisis event that was left out of the dataset during its creation (see Section 3.1). The data split for each dataset (train on 8 event/test on 9th event) is presented in the second column of the table. For each of these 9 datasets we created the four different classifiers described in Section 4.2. The results of each of these models for the 9 different datasets are reported in table along with their values of Precision (P), Recall (R), F1-measure (F) and the increment of F measure over the baseline, $\Delta F/F$.

	Class-1,0 Size				**SF**			**SF+SemAF**				**SF+SemAF+SemEF**				**SF+SemFF**	
Test	**Train**		**Test**		**P**	**R**	**F**	**P**	**R**	**F**	$\Delta F/F$	**P**	**R**	**F**	$\Delta F/F$	**F**	$\Delta F/F$
Event	**1**	**0**	**1**	**0**													
WTE	1556	1450	111	89	0.806	0.805	0.804	0.813	0.81	0.808	0.005	0.819	0.815	0.812	0.010	**0.823**	0.024
CWF	1420	1292	247	247	0.643	0.64	0.638	0.633	0.623	0.617	-0.033	0.716	0.715	**0.714**	0.119	0.71	0.113
CFL	1578	1464	89	75	0.784	0.774	0.774	0.796	0.793	0.793	0.025	0.79	0.787	0.787	0.017	**0.793**	0.025
ABF	1417	1289	250	250	0.776	0.774	0.774	0.782	0.778	0.777	0.004	0.811	0.8	**0.798**	0.031	0.788	0.018
BB	1588	1468	79	71	0.713	0.707	0.702	0.693	0.693	0.693	-0.013	0.734	0.733	0.732	0.043	**0.759**	0.081
LAS	1537	1419	130	120	0.811	0.808	**0.808**	0.777	0.776	0.776	-0.040	0.777	0.776	0.775	-0.041	0.787	-0.026
QFL	1347	1258	320	281	0.699	0.694	0.694	0.702	0.696	**0.695**	0.001	0.702	0.691	0.69	-0.006	0.691	-0.004
SBC	1306	1200	261	239	0.618	0.594	0.58	0.651	0.64	**0.636**	0.097	0.619	0.584	0.561	-0.033	0.565	-0.026
SGH	1587	1472	80	67	0.716	0.66	0.648	0.744	0.68	0.669	0.032	0.737	0.68	**0.67**	0.034	0.662	0.022
Avg.							0.714			0.718	0.009			0.727	0.0194	0.731	0.0251
%											0.9%				1.94%		**2.51**%

Table 3. Cross-Crisis Evaluation- SF, SemAF, SemEF, and SemFF feature sets (best set of features highlighted in bold)

As we can see, the use of semantics enhances classification results in all cases. We observe that SF+SemAF improves the classification over the baseline SF, in 6 out of 9 case, with an average of 0.9% increase in F-1 measure. As opposed to our previous experiment (10-fold cross-validation), however, the use of hypernyms makes the model more adaptable to unknown data in 6 out of 9 cases, with an average improvement of 1.94% over the baseline(SF). Semantic expansion (SemEF) improves over the annotation model (SemAF) in 5 out of 9 cases. Also, it is worth noting that filtering out the abstract concepts resulted in an improved performance of SF+SemFF over SF+SemAF+SemEF model (average of 0.6%), in 7 out of 9 cases. This validates the argument (Sec 3.2.2) that certain concepts tend to appear in both, crisis related and non-related tweets, and

[16] Each model was tested on the 8 event dataset it was trained on using 10 fold cross-validation to ensure its accuracy before applying it to the 9th event data. There accuracy drops around 17% on average when applied to new events.

therefore introduce noise rather than helping with the classification. Filtering out such concepts enhances the classification. SF+SemFF model improves over the baseline by an average of 2.51%.

5 Discussion and Future Work

Our findings show potential in mixing statistical and semantic features for classifying crisis-related and unrelated tweets. The highest, and more worthy, improvement is achieved when using this hybrid model to classify data of a new crisis event that the model was not trained on. This is due to the use of semantic knowledge graphs to expand the vocabulary into semantic concepts and hypernyms, and thus capturing the essence of the tweets and their terms. However, we showed that such a semantic expansion could introduce noise in the form of abstract concepts, which requires filtering to maximise benefit.

An issue we encountered was the unsymmetrical mappings of Hypernym-Hyponym relationship in BabelNet, which effected the hierarchical expansion of semantics and hierarchy generation. As a future work, we plan to refer to more symmetrically mapped resources, such as WordNet[17], and extend to the types and categories of semantics through external knowledge base such as DBpedia[18].

One of the limitations of this study is the small size of the dataset (3206 annotations) and type of crisis events (5 different types), which we plan to expand in future work. We also need to investigate whether the discriminative features differ across the various type of crisis, and languages. Additionally, we will investigate whether adding semantic features incorrectly classifies some tweets that are correctly classified by the statistical approach.

6 Conclusion

This work presents an approach to leverage semantic enrichment for classifying unseen crisis Twitter data. The two approaches of semantic enrichment; *annotation* and *semantic expansion*, exhibit an improvement in classification performance over the statistical features by 0.9%-2.51%. We have also demonstrated empirically that more abstract concepts are less discriminative, and proposed a method that filters the concepts which are less likely to be discriminative.

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.

[17] `https://wordnet.princeton.edu/`

[18] `http://wiki.dbpedia.org/`

4. Cristianini, N. and Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.
5. Gao, H., Barbier, G., & Goolsby, R. Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intelligent Systems, 26(3), 10-14, 2011.
6. 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.
7. 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.
8. Karimi, S., Yin, J. and Paris, C. December. Classifying microblogs for disasters. Australasian Document Computing Symposium, Brisbane, QLD, Australia, 2013.
9. 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.
10. 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.
11. 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.
12. Olteanu, A., Vieweg, S. and Castillo, C. February. What to expect when the unexpected happens: Social media communications across crises. Conf. on Computer Supported Cooperative Work & Social Computing (CSCW '15), Vancouver, Canada, 2015.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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), Atlanta, GA, USA, 2010.
18. 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/
19. Yin, J., Lampert, A., Cameron, M., Robinson, B. and Power, R. Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 27(6), pp.52-59, 2012.
20. Zhang, S. and Vucetic, S.. Semi-supervised Discovery of Informative Tweets During the Emerging Disasters. arXiv preprint arXiv:1610.03750, 2016.