STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVIII**, Number 2, 2013

ON THE STUDY OF REDUCING THE LEXICAL DIFFERENCES BETWEEN SOCIAL KNOWLEDGE SOURCES AND TWITTER FOR TOPIC CLASSIFICATION

ANDREA VARGA[(1)], AMPARO CANO[(2)], FABIO CIRAVEGNA[(1)], AND YULAN HE[(2)]

Abstract. State-of-the-art approaches on *cross-source topic classification* (TC) of Tweets rely on building a supervised machine learning classifier on *Social Knowledge Sources* (KSs) (such as DBpedia and Freebase) for detecting topics of Tweets. These approaches typically employ various lexical, syntactical or semantic features derived from the content of these documents or Tweets, often ignoring other indicators to external data sources (e.g. URL), which can provide additional background information for cross-source TC. In order to address these limitations, in this paper we analyse various such indicators, and evaluate their impact on cross-source TC. Our experiments, evaluating the proposed TC in the context of Violence Detection (VD) and Emergency Response (ER) tasks, indicate that the *Twitter specific information (indicators)* contain valuable information; and thus incorporating them into a TC can improve the performance over previous approaches not considering them.

1. Introduction

Topic classification (TC) of Tweets has only started to gain attention very recently. It provides an efficient and effective way of organising and searching Tweets, which can then be useful for various tasks e.g. *relating topics to events* (such as an Airplane crash, Egypt revolution, Mexican drug war, etc.) ([11]), *summarisation* ([12]), *question answering* ([6]), *content filtering* ([16]) etc. State-of-the-art approaches on *cross-source topic classification* (TC) of Tweets rely on building a supervised machine learning classifier on *Social Knowledge*

Received by the editors: April 15, 2013.

2010 *Mathematics Subject Classification.* 68T50, 03H65.

1998 *CR Categories and Descriptors.* I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text Analysis.*

Key words and phrases. cross-source topic classification, linked knowledge sources, violence detection, emergency response.

This paper has been presented at the International Conference KEPT2013: Knowledge Engineering Principles and Techniques, organized by Babeş-Bolyai University, Cluj-Napoca, July 5-7 2013.

Sources (KSs) (such as *DBpedia* and *Freebase*) for detecting topics of Tweets [1, 7, 17]. These approaches typically employ various lexical (BoW) and entity-based features (BoE) derived from the sole content of these documents or Tweets, often ignoring other features which can act as indicators of specific external data sources (e.g. URL, hashtags), providing additional background information for cross-source TC. In order to address these limitations, in this paper we analyse various such *external data sources' indicators*, and evaluate their impact on cross-source TC.

To better understand how the Twitter specific information impacts a cross-source TC, we conducted an in-depth analysis of Tweets collected over a period of three months belonging to Violence Detection (VD) and Emergency Response (ER) situations, and answered the following research question: *Do information derived from external data sources' indicators play an important role in TC of Tweets?*

The main contribution of our work are thus as follows: i) *we investigate the impact of enhancing the content of a Tweet by leveraging external data-sources obtained from Twitter content indicators, namely URLs and hashtags* ii) *we provide a detailed analysis and comparison on three topics (i.e. War, Disaster and Accident, and Law and Crime) related to the VD & ER scenarios.*

Before studying the above research questions, in Section 2 we review related work on TC; in Section 3 we introduce and describe the main methodology we followed to enrich KSs and Tweets with additional background information. The experimental results are described in Section 4, and the main challenges that we faced are presented in Section 5. Conclusions are then drawn in Section 6.

2. Related Work

Existing approaches to topic classification of Tweets can be divided into two main strands: approaches utilising a single data source (single source TC) (e.g. data from Twitter or blogs) for TC and approaches utilising *social knowledge sources* (multi-source or cross-source TC) (such as DBpedia or Freebase) for TC.

In the former case, Genc et al. [3] proposed a latent semantic topic modelling approach, which mapped each Tweet to the most similar Wikipedia articles based on lexical features extracted from Tweets' content only. Song et al. [13] mapped a Tweet's terms to the most likely resources in the Probbase KS. These resources were used as additional features in a clustering algorithm which outperformed the simple BoW approach. Munoz et al. [10] proposed an unsupervised vector space model for detecting topics in Tweets in Spanish.

They used syntactical features derived from PoS (part-of-speech) tagging, extracted entities using the Sem4Tags tagger ([2]) and assigned a DBpedia URI for those entities by considering the words appearing in the context of the entity inside the Tweets. Vitale et al. [18] proposed a clustering based approach which augmented the BoW features with BoE features extracted using the Tagme system, which enriches a short text with Wikipedia links by pruning n-grams unrelated to the input text, showing significant improvement over the BoW features. Tao et al. [14] studied various Twitter dataset specific features (including whether a tweet contains a hashtag or a URL) for identifying whether a tweet is relevant to a topic, and showed that incorporating these features can help TC.

Considering the approaches exploiting data from KSs ; Michelson et al. [8] proposed an approach for discovering Twitter user' topics of interest by first extracting and disambiguating the entities mentioned in a Tweet. Then a subtree of Wikipedia category containing the disambiguation entity is retrieved and the most likely topic is assigned. Milne et al. [9] also assigned resources to Tweets. In their approach they make use of Wikipedia as a knowledge source, and consider a Wikipedia article as a concept, their task then is to assign relevant Wikipedia article links to a Tweet. They proposed a machine learning approach, which makes use of Wikipedia n-gram and Wikipedia link-based features. Xu et al. [19] proposed a clustering based approach which linked terms inside Tweets to Wikipedia articles, by leveraging Wikipedia's linking history and the terms' textual context information to disambiguate the terms meaning. In Varga et al. [17], we studied the similarity between KSs and Twitter using both BoW and BoE features, showing that DBpedia and Freebase KSs contain complementary information for TC of Tweets, with the lexical features achieving the best performance. More recently, in Cano et al. ([1]) we demonstrated that exploiting the semantic information about entities from DBpedia and Freebase is beneficial, and incorporating additional semantic information about entities in terms of properties and concepts can furthermore improve the performance of TC against the sole Twitter data approach. Consequent work, classifying blog posts into topics ([5]) has also demonstrated that selecting data from Freebase using distant supervision in addition to incorporating features about named entities is beneficial for TC.

Whilst previous work already focused on incorporating lexical and semantic features into TCs, these features were extracted from the sole content of Tweets. However, due to the length constraints of Twitter messages, these short messages often contain various other information (e.g. URLs or hashtags), which can further help the understanding of the content of the messages. The usefulness of Twitter specific features (such as "has_URL", "has_ hashtag") has already been shown to be beneficial for single source TC case ([14]).

However, to date no study has been conducted to validate whether these data-source specific indicators are also useful in cross-source TC scenarios as well.

3. Enriching KSs and Twitter with additional background information

In previous work [1, 17] we have investigated the use of *Social Knowledge Sources* (e.g. DBpedia, Freebase) for building cross-source topic classifiers, which can aid in the topic classification of Tweets. In such approaches we leverage the entities appearing in both KS and Twitter content for deriving semantic features, enabling to reduce the distributional differences across datasets. One of the main reasons for the distributional differences lies in the variation in vocabulary, writing style and format of the documents across data sources.

In this paper, however, we introduce a novel approach which leverages two main type of external source indicators for reducing the differences for both lexical and entity features across datasources. Figure 1, presents a tweet highlighting entities, links and hashtags.

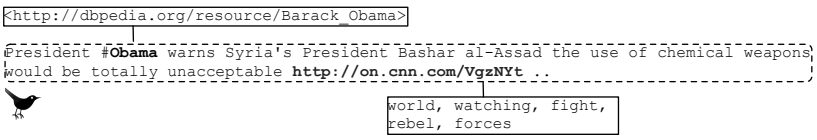

Figure 1. Enriching tweet content by using hashtags and links as indicators of external sources.

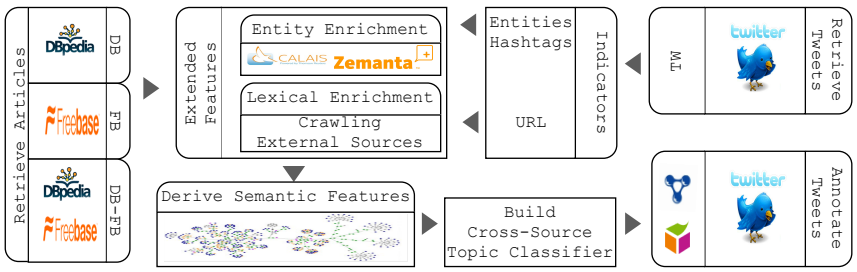

Figure 2. Architecture of cross-source TC based on lexical and entity-based feature enrichment derived from specific indicators (e.g. hashtags, links and entities).

In our approach we propose to reduce the lexical and entity differences between KSs and Twitter by: i) incorporating lexical features derived from external sources pointed out by *links indicators*; and by ii) incorporating KS entity-based features derived from resources embedded in *hashtags indicators* (e.g. #Egypt, #Obama).

We propose the architecture summarised in Figure 2 for building a cross-source TC. The proposed architecture involves the use of three datasets, two of them consisting of articles derived from DBpedia (DB) and Freebase (FB) KSs, and a third one consisting of a collection of tweets (TW). This architecture comprises the following stages:

1) *KS Data collection* - Given a topic c, KS-derived datasets (DB and FB) are generated by SPARQL querying for those articles whose categories and subcategories are c.
2) *Feature Extraction* - For both KS-datasets and the TW dataset, lexical features represented by a bag of words (BoW) are extracted and weighted using a TF-IDF weighting function, keeping only the top 1000 words. For the KS-datasets, a bag of entities (BoE) features is also generated, using the OpenCalais[1] and Zemanta[2] name entity recognition services.
3) *Indicator Extraction* - For the TW dataset, URL and hashtag (HSH) indicators are extracted. These indicators will be referred to as bag of links (BoL) and bag of hashtags (BoH) respectively.
4) *Incorporating background information from indicator features* - In order to overcome the number of character limitation posed by Tweets, the feature space representing a tweet is extended via the BoL and BoH indicators as follows:
   BoL based features. - Each URI from the BoL is resolved and the content of the referenced website is parsed. For each link the following lexical features are kept: i) the title of the page (BoL (T)); ii) the first paragraph of the page (BoL (1)); and iii) the last paragraph of the page (BoL (L)).
   BoH based features. - In order to assign a semantic meaning to a hashtag we implemented a series of regular expressions, which relate a term inside a hashtags to a DBpedia or Freebase resource URI (e.g. #egypt will be associated with `dbpedia.org/resource/Egypt` and `freebase:Egypt`). These resources, which we refer to as bag of resources are later on used as pointers to enable semantic enrichment.
5) *Semantic Feature Enrichment* - The semantic enrichment consists on extending a feature space with ontological classes and properties which characterise a KS resource URI. For example, the resource `dbpedia.org/resource/`

[1] http://opecalais.com
[2] http://zemanta.com

`Barack_Obama` is related to rdf types such as *yago:PresidentsOfTheUnitedStates* and *yago:NobelPeacePrizeLaureates*, and is characterised by properties such as *dbpedia-owl:commander* and *dbpedia-owl:knownFor*. The described semantic enrichment was applied to both KS BoE features and to the TW bag of resources derived from the bag of hashtags. We weighted the class features by frequency, while the property features were weighted following the specificity-generality weighting function introduced in [1].

6) *Building Cross-source topic classifier* - For our experiments, we considered as the base cross-source classifier a supervised TC classifier (SVM DB-FB) trained on the joint DBpedia (DB) and Freebase (FB) KSs, which was found to perform best for the topic classification task ([1]). This classifier takes into account both lexical, entity and semantic features introduced in the above stages of this architecture.
7) *Annotating Tweets* Finally tweets are annotated as belonging or not to the given topic c.

4. Experiments

To understand how the different information provided by external sources influence the performance of a TC, we evaluated our framework on a series of experiments, and conducted an analysis on a corpus of Tweets compiled over three months.

In our analysis we investigated the research questions of *Do information derived from external data sources indicators play an important role in TC of Tweets?*

4.1. Dataset characteristics.

For building our single source and cross-source TCs, we used the same dataset collected in our previous work ([1]), consisting of 9,465 articles from DB, 16,915 articles from FB and 10,189 tweets from Twitter (TW), covering multiple topics including three specific to ER &VD tasks (Disaster (*DisAcc*), Crime (*Cri*) and War (*War*)).

The general statistics about the TW dataset are summarised in Table 1. As we can observe, in the TW dataset the frequency of hashtags (HSH) and URLs is relatively low, indicating that only a small number of Tweets contain external data source specific information. In total 2,386 (23,41%) tweets contain at least one hashtag; and 3,348 (32.85%) Tweets contain at least one URLs. The number of unique hashtags is 1,784; while the number of unique URLs is 1,902.

The concept statistics derived for each entity in the KSs dataset are summarised in Table 2.

Topic	%Hsh	#Hsh	%URL	#URL	#dbCls (HSH)	#yagoCls (HSH)	#fbClass (HSH)
DisAcc	1.85%	233	2.59%	154	29	150	316
Cri	2.78%	220	6.41%	411	23	169	312
War	2.10%	198	2.65%	139	20	171	215

TABLE 1. Twitter dataset statistics about external data-source indicators (HSH, URL). #dbCls(HSH) refers to the number of unique KSs concepts derived for HSH from DB ontology; #yagoCls (HSH) refers to the number of unique KSs concepts derived for HSH from Yago ontology; and #fbClass (HSH) stands for the number of unique FB concepts derived for HSH from Freebase ontology.

Topic	#dbCls (BoE)	#yagoCls (BoE)	fbClass (BoE)
DisAcc	119	3,865	1,289
Cri	119	3,865	1,289
War	124	3,864	1,215

TABLE 2. Concept statistics in the multi-source DB-FB KS dataset. #dbCls (BoE) refers to the number of unique DB concepts derived for the named entities from DB ontology; #yagoCls (BoE) refers to the number of unique KSs concepts derived for entitites from Yago ontology; and #fbClass (BoE) stands for the number of unique FB concepts derived for entities from Freebase ontology.

When augmenting the single source TC classifier with concept and property features, we used a reduced vocabulary consisting of 180 unique concepts and properties from KSs, which we empirically set.

4.2. **Results.** We employed SVM classifiers for both single source (SVM TW), and cross-source (SVM DB-FB) classifiers to classify tweets into relevant topics. When training the classifiers, we split the TW dataset up into a training/testing set using an 80:20 split. This resulted in that the SVM TW classifier was trained on 80% of the TW dataset, while the SVM DB-FB classifier was trained on the full KSs data together with 80% of TW data. The test set

in both cases consists of 20% of TW data, and the results were averaged over five independent runs.

Given the sparse distribution of HSHs and URLs in Tweets, we performed two series of experiments. In the first set of experiments we utilised the full set of TW data (10,189 Tweets), which we denote as *Full*. In our second set of experiments, we only considered Tweets having at least one HSH or URLs (resulting in 4,778 Tweets), which we refer to *Filt*.

Case	Features	*DisAcc*			*Cri*			*War*		
		P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Full	BOW	0.776	0.635	0.699	0.724	0.489	0.584	0.842	0.745	0.791
	BoL(1)	0.791	0.644	***0.710***	0.720	0.522	0.605	0.880	0.737	0.802
	BoL(L)	0.796	0.640	***0.710***	0.716	0.525	0.605	0.885	0.739	0.806
	BoL(T)	0.778	0.626	0.694	0.723	0.507	0.596	0.872	0.732	0.796
	BoH(Cls)	0.773	0.644	0.703	0.687	0.540	0.605	0.861	0.745	0.799
	BoH(P)	0.783	0.649	0.709	0.695	0.544	***0.610***	0.882	0.752	***0.812***
Filt	BOW-Filt	0.877	0.498	0.635	0.749	0.400	0.522	0.955	0.624	0.755
	BoL(1-Filt)	0.801	0.509	0.623	0.725	0.474	0.574	0.839	0.698	0.762
	BoL(L-Filt)	0.801	0.509	0.623	0.727	0.474	0.574	0.839	0.698	0.762
	BoL(T-Filt)	0.813	0.497	0.617	0.766	0.488	0.596	0.874	0.714	0.786
	BoH(Cls-Filt)	0.810	0.523	***0.636***	0.733	0.488	0.586	0.868	0.724	0.790
	BoH(P-Filt)	0.796	0.515	0.625	0.747	0.526	***0.617***	0.892	0.746	***0.813***

TABLE 3. The performance of the SVM TC using extrenal *data source indicators*.

Table 3 summarises the results obtained for the single-source TC case. When considering the full TW corpus, we can observe that the classifier built using BoL and BoH features improve upon the baseline classifier considering words only (BoW), except for the *DisAcc* topic using *BoL (T)* features. The best overall results were obtained by the *BoH (Prop)*, achieving an improvement of 2.6% over the baseline for the *Cri* topic, and an improvement of 1.5% for the *War* topic. These results are also in agreement with our previous findings ([1]), showing that the property features provide useful information for TC, and also incorporating them into TC is more beneficial than utilising concept features.

Considering the results on the filtered TW corpus, we again found the *BoH (Prop)* features to perform the best, except for the *DisAcc* topic. The improvement over the baseline classifier, however, was much bigger in this case: 4.3% for the *Cri* topic, and 5.8% for the *War* topic. An explanation for the small improvement for the *DisAcc* topic can be understood by the fact that the Tweets belonging to the *DisAcc* topic contain the less number of HSHs and URLs, and therefore less number of Tweets are semantically enriched.

Looking at the individual features derived from the URLs, in the Filt case, when most of the tweets have a URL inside them, the Title of the articles was

Case	Features	*DisAcc*			*Cri*			*War*		
		P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
Full	BOW	0.955	0.869	0.910	0.944	0.857	0.898	0.955	0.861	0.905
	BoL(1)	0.955	0.867	0.908	0.943	0.857	0.898	0.958	0.871	0.913
	BoL(L)	0.955	0.866	0.908	0.945	0.859	0.900	0.958	0.868	0.911
	BoL(T)	0.953	0.862	0.905	0.944	0.854	0.897	0.959	0.868	0.911
	BoH(Cls)	0.959	0.979	***0.969***	0.946	0.974	***0.960***	0.964	0.984	***0.974***
	BoH(P)	0.955	0.900	0.927	0.947	0.895	0.920	0.958	0.902	0.929
Filt	BOW-Filt	0.842	0.409	0.550	0.711	0.386	0.500	0.956	0.823	0.885
	BoL(1-Filt)	0.766	0.673	0.716	0.917	0.813	0.862	0.956	0.827	0.887
	BoL(L-Filt)	0.957	0.841	0.895	0.914	0.805	0.856	0.958	0.824	0.886
	BoL(T-Filt)	0.958	0.834	0.892	0.917	0.814	0.863	0.961	0.822	0.886
	BoH(Cls-Filt)	0.953	0.986	***0.969***	0.918	0.966	***0.941***	0.964	0.981	***0.973***
	BoH(P-Filt)	0.956	0.876	0.914	0.919	0.842	0.879	0.960	0.868	0.912

TABLE 4. The performance of the DB-FB cross-source SVM TC using various external *datasource indicators* .

found to be more informative of a topic. However, in the Full case, the first and the last paragraphs of the webpages were found more beneficial than the title of the webpages.

The results corresponding to the cross-source scenario are presented in Table 4. We can observe different trends compared to the single source scenario. For the case of the full TW dataset, the best cross-source feature for all the three scenarios was the *BoH (Cls)* feature. The highest improvement of 6.2% being achieved for the *Cri* topic. We can furthermore notice, that the results for the *BoH (Prop)* features are also outperforming the results obtained by the URL features. These results indicate, that incorporating semantic information derived from KSs are very important in reducing the lexical gap between KSs and TW. In particular, the addition of new words derived from the external URL websites were found worst or achieved little improvement over the baseline BoW case (for *DisAcc*, *Cri*). With respect to the URL features, we can notice that the performance of the classifier does not change drastically when utilising the first, last or the title of external URL websites. The difference in the performances is less than 1%.

Examining the results obtained for the Filtered case, we can observe similar trends, where again the *BoH (Cls)* feature exhibit the highest performance for each topic, which is then followed by the *BoH (Prop)* features. Considering the URL features, however, we can notice that the title of the websites seems to be more beneficial for TC, than the first or the last paragraphs. An explanation for this could be, that in this Filtered scenario more tweets are affected by feature augmentation than in the Full scenario. In light with the results for the single-source scenario, we can also observe a bigger improvement (up to 44.2% for *Cri*) in the Filt case than in the Full case.

5. Challenges and Limitations

In this work we enriched the representation of short text messages with information from external websites with the goal of reducing the lexical differences between KSs and Twitter.

One of such external link indicators were the hashtags from a Tweet, for which we assigned a DBpedia and Freebase URI using a simple word matching approach. We encountered various challenges, given that hashtags can often contain: (1) abbreviations (e.g. #nkorea `http://dbpedia.org/resource/North_Korea`); (2) contain compound words (e.g. #flightdelay `http://dbpedia.org/resource/Flight_delay`); and (3) some of the hastags may contain new abbreviations not present in KSs (e.g. #emfrmf) . For those cases no semantic meaning was assigned to them. In addition, one hashtag as any other word (#beirut) may have multiple meanings (e.g. the capital city of Lebanon; or a Lebanese governorate), and thus in order to assign the correct DB and FB URI one may apply a word sense disambiguation algorithm ([15]) first, which takes into account not only the lexical form of a hashtag but also the context of the hashtag.

The automatic extraction of sentences and paragraphs from external websites also poses challenges. One of the main problems considering this task is the accurate identification of boundaries on a page, since different websites employ different formats for describing the content of their pages. Particularly in pages where users can add comments (e.g. newswire articles and forum-like pages) the identification of the last paragraphs becomes challenging. In our work we parsed a full webpage as a whole, independently of whether the last part of the page referred to users' comments or not.

6. Conclusion and Future work

This study presented an approach for incorporating various background information into cross-source TCs built on multiple linked KSs. The goal of our study was to investigate whether incorporating such information can furthermore reduce the lexical differences between KSs and Twitter -imposed by the short length nature of Tweet messages-, thus allowing the creation of more accurate TC of Tweets.

We looked at two Twitter specific indicators including hashtags and URLs, for which we derived additional lexical and semantic features for training cross-source TCs. Our results on both sole Twitter and cross-source settings reveal that the indicator which provides a better feature enrichment, and therefore better classification performance was the hashtags.

Our future effort will consist on investigating alternative ways for bridging the gap between KSs and Twitter. One possible future direction could be to

investigate the impact of tweet normalisation approaches for cross-source TC, aiming at resolving the abbreviation, misspelling to standard English words ([4])[3]

7. References

[1] Amparo Elizabeth Cano, Andrea Varga, Matthew Rowe, Fabio Ciravegna, and Yulan He. Harnessing linked knowledge sources for topic classification in social media. In *24th ACM Conference on Hypertext and Social Media, HT '13*. ACM, 2013.

[2] A. Garcia-Silva, Oscar Corcho, and J. Gracia. Associating semantics to multilingual tags in folksonomies, 2010.

[3] Yegin Genc, Yasuaki Sakamoto, and Jeffrey V. Nickerson. Discovering context: classifying tweets through a semantic transform based on wikipedia. In *Proceedings of the 6th international conference on Foundations of augmented cognition: directing the future of adaptive systems*, FAC'11, pages 484–492, Berlin, Heidelberg, 2011. Springer-Verlag.

[4] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[5] Stephanie Husby and Denilson Barbosa. Topic classification of blog posts using distant supervision. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 28–36, Avignon, France, April 2012. Association for Computational Linguistics.

[6] Baichuan Li, Xiance Si, Michael R. Lyu, Irwin King, and Edward Y. Chang. Question identification on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2477–2480, New York, NY, USA, 2011. ACM.

[7] Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12*.

[8] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, New York, NY, USA, 2010.

[9] D. Milne and I. H. Witten., editors. *Learning to link with Wikipedia.* 2008.

[10] Óscar Muñoz García, Andrés García-Silva, Óscar Corcho, Manuel de la Higuera Hernández, and Carlos Navarro. Identifying Topics in Social Media Posts using DBpedia. In Meunier Jean-Dominique, Halid Hrasnica, and Florent Genoux, editors, *Proceedings of the NEM Summit*, pages 81–86. NEM Initiative, Eurescom ? the European Institute for Research and Strategic Studies in Telecommunications ? GmbH, September 2011.

[3] In this work, we performed some initial experiments applying a dictionary based approach for Tweet normalisation. We built a lexicon from *http://www.noslang.com/* website, consisting of 5,407 abbreviation word pairs, and replaced all abbreviations found in tweets with standard English terms. Our initial results, however, showed no improvement upon the baseline model without normalisation. Our future work will thus aim to investigate other tweet normalisation approaches too.

[11] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[12] Beaux Sharifi, Hutton, Mark-Anthony, and Jugal Kalita. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 685–688, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[13] Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. Short text conceptualization using a probabilistic knowledgebase. In *IJCAI*, pages 2330–2336, 2011.

[14] Ke Tao, Fabian Abel, Claudia Hauff, and Geert-Jan Houben. What makes a tweet relevant for a topic? In *Making Sense of Microposts (#MSM2012)*, pages 49–56, 2012.

[15] Doina Tatar. Word sense disambiguation by machine learning approach: A short survey. *Fundam. Inf.*, July 2004.

[16] Irina Temnikova, Dogan Biyikli, and Francis Boon. First steps towards implementing a sahana eden social media dashboard. In *Proceedings of the conference Social Media and Semantic Technologies in Emergency Response (SMERST 2013)*, Coventry, UK, 2013.

[17] Andrea Varga, Amparo Elizabeth Cano, and Fabio Ciravegna. Exploring the similarity between social knowledge sources and twitter for cross-domain topic classification. In *Proceedings of the Knowledge Extraction and Consolidation from Social Media, 11th International Semantic Web Conference (ISWC2012)*, 2012.

[18] Daniele Vitale, Paolo Ferragina, and Ugo Scaiella. Classification of short texts by deploying topical annotations. In *ECIR*, pages 376–387, 2012.

[19] Tan Xu and Douglas W. Oard. Wikipedia-based topic clustering for microblogs. *Proc. Am. Soc. Info. Sci. Tech.*, 48(1):1–10, 2011.

(1) The Organisations, Information and Knowledge (OAK) Group, Department of Computer Science, University of Sheffield, UK
E-mail address: a.varga@dcs.shef.ac.uk

(2) School of Engineering and Applied Science, Aston University, UK
E-mail address: ampaeli@gmail.com

E-mail address: f.ciravegna@dcs.shef.ac.uk

E-mail address: y.he9@aston.ac.uk