Linked knowledge sources for topic classification of microposts: a semantic graph-based approach

Journal Item

Linked Knowledge Sources for Topic Classification of Microposts: A Semantic graph-based approach

Andrea Varga[a], Amparo E. Cano[b], Matthew Rowe[c], Fabio Ciravegna[a], Yulan He[b]

[a] *Organisations, Information and Knowledge Group, The University of Sheffield, UK*
[b] *School of Engineering and Applied Science, Aston University, UK*
[c] *School of Computing and Communications, Lancaster University, UK*

Abstract

Short text messages a.k.a Microposts (e.g. Tweets) have proven to be an effective channel for revealing information about trends and events, ranging from those related to Disaster (e.g. hurricane Sandy) to those related to Violence (e.g. Egyptian revolution). Being informed about such events as they occur could be extremely important to authorities and emergency professionals by allowing such parties to immediately respond.

In this work we study the problem of topic classification (TC) of Microposts, which aims to automatically classify short messages based on the subject(s) discussed in them. The accurate TC of Microposts however is a challenging task since the limited number of tokens in a post often implies a lack of sufficient contextual information.

In order to provide contextual information to Microposts, we present and evaluate several graph structures surrounding concepts present in linked knowledge sources (KSs). Traditional TC techniques enrich the content of Microposts with features extracted only from the Microposts content. In contrast our approach relies on the generation of different weighted semantic meta-graphs extracted from linked KSs. We introduce a new semantic graph, called category meta-graph. This novel meta-graph provides a more fine grained categorisation of concepts providing a set of novel semantic features. Our findings show that such category meta-graph features effectively improve the performance of a topic classifier of Microposts.

Furthermore our goal is also to understand which semantic feature contributes to the performance of a topic classifier. For this reason we propose an approach for automatic estimation of accuracy loss of a topic classifier on new, unseen Microposts. We introduce and evaluate novel topic similarity measures, which capture the similarity between the KS documents and Microposts at a conceptual level, considering the enriched representation of these documents.

Extensive evaluation in the context of Emergency Response (ER) and Violence Detection (VD) revealed that our approach outperforms previous approaches using single KS without linked data and Twitter data only up to 31.4% in terms of F1 measure. Our main findings indicate that the new category graph contains useful information for TC and achieves comparable results to previously used semantic graphs. Furthermore our results also indicate that the accuracy of a topic classifier can be accurately predicted using the enhanced text representation, outperforming previous approaches considering content-based similarity measures.

Keywords: linked knowledge sources, semantic concept graphs, topic classification

1. Introduction

Social media posts, and in particular Microposts collected from Twitter, have been found to contain useful information for many applications including disaster detection [1], seasonal mood level changes [2], tracking influenza rates [3], box-office revenue forecast [4], political elections [5], stock market prediction [6], etc. For instance, during the widespread protest in Egypt in 2013, Microposts were found to provide early warning signals of violent events; such events were reported much faster than traditional media sources[1]. The real-time identification of such events could be extremely important to authorities and emergency professionals by allowing such parties to immediately respond.

However, the classification of such messages poses unique challenges, due to the special characteristics of the messages i) the limited length of Microposts (up to 140 characters), restricting the contextual information necessary to effectively understand and classify them; ii) the noisy lexical nature of Microposts, where new terminology and jargon emerges as different events are discussed; iii) the large topical coverage of Micropost.

Existing approaches have addressed these challenges by proposing the use of social knowledge sources (KSs). These sources provide additional textual data on a growing number of topics, which can alleviate the sparsity of Microposts's content ([7, 8, 9, 10, 11, 12]). Furthermore these topic classifiers typically enhance the *lexical* (e.g. *bag-of-words* (*BoW*)) representation of text by incorporating additional contextual information about Microposts in the form of *semantic* (*bag-of-*

Email address: a.varga@dcs.shef.ac.uk (Andrea Varga)

[1] http://irevolution.net/2013/07/07/twitter-political-polarization-egypt/

entities (*BoE*)) features extracted from the content of Microposts only. Unlike these approaches, recently in [13] we proposed a TC framework which generates contextual information from graph structures surrounding concepts in multiple complementary linked KSs. Among the several useful graph structures defined in KSs ([8]), such as the *resource meta graph* providing course grained classification of concepts by their type, or the *category meta graph* which groups similar concepts together by their topic, our original framework exploited the *resource meta-graph* for context generation. Moreover, in [14] we also studied different content-based topic similarity (also called domain similarity or dataset similarity [15]) measures, which quantify the similarity between the KS data and Twitter data, serving as a proxy for the performance of a topic classifier on Twitter data. These content based features correspond to simple *BoW* and *BoE* features derived from the Micropost content only.

Unfortunately, current approaches still present some limitations. The majority of the approaches model the entities using very generic concept types. For example, in the case of the entity *Obama*, the generic class *Person* is considered. When detecting Microposts related to the *war* topic, however, a more fine grained categorisation of this entity, such as *President of United States* (*Presidents_of_the_United_States*), could be more useful.

Further, the use of fine grained information in KSs provided by the *category meta-graph* has been exploited for many other problems, such as document classification [8], entity disambiguation [16], and semantic relatedness [17], and shown that it carries rich semantic information. However, to date no study has been conducted to investigate the usefulness of this *category meta-graph* graph structure for TC.

In this paper we thus present an extension of our TC framework [13], which exploits this new semantic graph, called *category meta-graph*, providing a more fine grained classification of concepts based on their topics. We introduce a set of novel semantic features derived from this graph, and present a comparative evaluation against those obtained from the *resource meta-graph*.

Furthermore our goal is also to understand which semantic feature contributes to the performance of a topic classifier. For this reason we propose an algorithm for automatic estimation of accuracy loss of a topic classifier. We introduce novel topic similarity measures, which in contrast to our previous content-based similarity measures ([14]), aim to measure the similarity between the KS documents and Microposts at a conceptual level, considering the enriched representation of these documents.

To evaluate the usefulness of exploiting this new *category meta-graph* for both TC and topic similarity, we present an extensive analyses of our extended framework using a ground truth data in the Emergency Response (ER) and Violence Detection (VD) domains.

The main research questions we investigate are the following:

- *How does the performance of a topic classifier vary using different concept graphs? Which concept graph provides the most useful semantic features for TC of Microposts?*
- *Are there differences in the roles (generalisation patterns) of the concept graphs in the different TC scenarios?*
- *Can we predict the performance of a topic classifier? Which topic similarity measure provides best estimate on the performance of a topic classifier?*

1.1. Contributions

To address the above research questions, we present an approach which facilitates the exploitation of multiple semantic meta-graphs from linked KSs for TC of Microposts. In particular, in contrast to our previous work ([13]), in this paper our main focus is to understand the differences between the different semantic concept graphs, and present a comparative evaluation of these graphs at different stages of our three-stage approach. The main stages of our approach can be summarised as follows: i) *context modelling*; ii) *topic classification* and iii) *topic similarity analysis*.

The *context modelling* stage enriches the text using different concept abstraction techniques. For this reason we extract various semantic features about entities appearing in the text from two distinct concept graphs built from linked KSs.

The second stage *topic classification* involves the creation of statistical TC models, which incorporate the various semantic features obtained in the context modelling step. In this stage we investigate two different scenarios: the Twitter only scenario in which we build a topic classifier on Twitter data only, and the cross-source TC case where we make use of the information from multiple linked KSs. This allows us to analyse which concept graph provides better semantic features for TC, and also whether the role of the semantic features differ according to the TC scenarios. In particular, we investigate whether the same semantic features which account for modelling the specificity of the topic in the Twitter only scenario, serve the same role in the cross-source scenarios.

The final stage *topic similarity analysis* uses the enhanced representation of the documents (in both the KSs and Twitter) following context modelling to provide an estimate on the performance of the topic classifier on new, unseen Micropost data. This allows us to analyse which semantic concept graph is better suited to measure the topic similarity between KS documents and Microposts for TC. In this stage, we also investigate whether this novel representation of the documents provides a better measure for topic similarity than our previous content based statistical measures ([14]).

The main contributions of this paper are four fold:

- We introduce a novel set of semantic features derived from the *category meta-graph* of KSs;
- We present a systematic comparison of different semantic concept graphs for TC of Microposts;
- We present an analysis of the different roles of semantic concept graphs on ground truth data in the VD and ER domains;

- We propose a novel set of topic similarity measures for estimating the performance of a topic classifier.

The remainder of this paper is structured as follows: Section 2 provides an overview of the employed linked KSs, and explains their relevance for TC of Microposts; Section 3 presents the related work on TC. Section 4 then provides an overview of the original TC framework employed in this work, and describes its extension for the newly introduced *category meta-graph*. Section 5 continues by introducing a novel set of topic similarity measures. Next, Section 6, presents the gold standard datasets used in our experiments, and Section 7 describes the baseline models employed. After that, in Section 8 a comparative evaluation of the extended TC framework and new topic similarity measures are discussed. This is followed by a discussion on the shortcomings of our approach and possible future extensions in Section 9. Finally, conclusions are drawn in Section 10.

2. An overview of DBpedia and Freebase Linked Knowledge Sources

The Linked Open Data (LOD) cloud[2] consists of a large number of interlinked KSs, covering a range of different topics. Among these KSs, DBpedia[3] ([18]) and Freebase[4] ([19]) constitute some of the largest datasets built in a collaborative manner. The main advantages of these KSs are: i) they provide plentiful amount of data on a growing number of topics, ii) they contain factual information about a large number of entities, covering these topics. This semantic information is also structured according to their own KS ontology.

Exploiting these KSs can thus be useful to support topic classification of Microposts, as a KS's data can be used to provide additional labelled data to train supervised topic classifiers of Microposts. For example, for the topic *violence* we can consider DBpedia's violence category (i.e `http://dbpedia.org/page/Category:Violence`). This category provides a large number of resources associated with it (e.g. `<http://dbpedia.org/page/Counter-terrorism>`). Each of such resources include a short description of its content (the abstract). This data can then be used for enhancing the lexical feature representation for the *violence* topic. Further, the rich semantic information within a KS ontology can be used to generate violence-related features at the graph level. The same approach can be described for the representation of an entity (e.g. `http://dbpedia.org/page/Barack_Obama`). An entity's representation extracted from these KSs can therefore provide additional contextual information. This information can be used to enhance a Micropost representation mentioning the entity.

To provide context for our motivating example, we first present statistics of the KSs used in this paper (i.e. DBpedia and Freebase), summarised in Table 1.

[2]http://lod-cloud.net

[3]DBpedia, http://dbpedia.org

[4]Freebase, http://frebase.org

The first KS, DBpedia (*dbKS*), is derived from Wikipedia[5]. In DBpedia each resource is harvested from a Wikipedia article which is semantically structured into a set of DBpedia (*dbOwl*)[6] and YAGO2 (*yago*)[7] ontologies, with the provision of links to external knowledge sources such as Freebase, OpenCyc[8], and UMBEL[9]. The Wikipedia articles are furthermore grouped into categories, which are represented using the SKOS vocabulary[10]. The DBpedia dump version 3.8 classifies 2.35 million resources into DBpedia's ontology classes (*dbOwl*). These classes comprises 359 distinct classes, and 740,000 SKOS categories (*dbCat*), which form a subsumption hierarchy and are described by 1,820 different properties. Conversely, the *yago* ontology ([20]) is a much bigger and fine grained ontology. It contains over 447 million facts about 9.8 million entities which are classified into 365,372 classes, and 104 manually defined properties.

Semantic Features	DBpedia (*dbKS*)			Freebase (*fbKS*)
	dbOwl	*dbCat*	*yago*	*fbOnt*
Resource	2.35×10^6		447×10^6	3.6×10^6
Property *(P)*	1,820		104	7,000
Class *(Cls)*	359	NA	365,372	1,450
Category *(Cat)*	NA	740,000	NA	NA

Table 1: Statistics about *dbOwl*, *dbCat*, *yago*, *fbOnt* KS ontologies.

In contrast to DBpedia, Freebase (*fbKS*) is a large online knowledge base which users can edit in a similar manner to Wikipedia. In Freebase, resources are also harvested from multiple sources such as Wikipedia, ChefMoz, NNDB and MusicBrainz[11] along with data individually contributed by users. These resources are semantically structured into Freebase's own ontologies (*fbOnt*), which consist of 1,450 classes and more than 7,000 unique properties.

In summary, these different KS ontologies (i.e. *dbOwl*, *yago*, *fbOnt*) provide a rich source of semantic information about entities in many domains and across topics. Further, in these KSs each entity resource is related to different ontological classes or concepts which can provide additional contextual information for that resource. This contextual information allows us to exploit various semantic structures of these resources.

Consider the two Micropost examples displayed in Figure 1. In these examples, the entity *Obama* is mentioned in two different contexts, each of them corresponding to different roles this entity plays (for e.g. *president* in Microposts (1); and *husband* in Micropost (2)). In such cases, the role of this entity is defined by the contextual information provided on the content of each Microposts.

[5]Wikipedia, http://wikipedia.org

[6]http://wiki.dbpedia.org/Ontology

[7]http://www.mpi-inf.mpg.de/yago-naga/yago/

[8]OpenCyc, http://sw.opencyc.org/

[9]UMBEL, http://www.umbel.org/

[10]http://www.w3.org/2004/02/skos/

[11]Freebase Datasources, http://wiki.freebase.com/wiki/Data_sources

Figure 1: Tweets exposing different contexts involving the same entitiy.

In this paper, our main goal is to exploit this semantic contextual information about entities. In particular, we study different semantic graph structures defined in linked KSs, and provide a comparative evaluation of their usefulness for TC of Microposts. Section 4 introduces our approach and the semantic metagraphs exploited in this work.

In addition, we explore the role of the semantic features derived from these different semantic graphs in the representation of a topic. For example, looking at the Micropost in Figure 1, and the semantic representation of the resource *Barack Obama* in Figure 2, we can observe that the semantic features derived about *Obama*'s resource (*Barack Obama*) can be indicative of the topic *war*. Our aim is thus to investigate whether the different semantic structures of a KS can aid in identifying which semantic features are more representative of this topic. For this reason, Section 5 introduces different metrics for analysing different semantic feature-based topic similarity measures.

3. Related Work

We classify the related approaches into two main strands: research pertaining to generic *topic detection approaches* and the *use of KSs for linking topics* to Microposts.

3.1. Topic Detection in Twitter

The closest task to our multi-label topic classification task is *topic detection* which aims to assign a set of topics or labels to a given Micropost. Recent approaches for topic detection on Twitter stream can be classified into: *descriptive characterisation of Microposts*, *topic models*, and *classification models*.

The first approach, *descriptive characterisation of topics*, employs various *lexical (bag-of-word)*, *syntactic* (hashtags) and *semantic features* (named entities) extracted from the content of Microposts.

Kwak et al. [21] utilised words, phrases and hashtags as indicators for trending topics in Twitter. Laniado and Mika. [22] proposed different hashtag-based metrics for identifying community interests and trending topics in Twitter. Their study revealed that a great proportion of hastags (more than 50%) can be associated to Freebase concepts. Out of these, 40% were found to correspond to named entities. Other work has shown, however, that hashtags can be ambiguous and their meaning can differ geographically [23, 24].

The second approach is based on *topic models*, which rely on the popular probabilistic Latent Dirichlet Allocation (LDA) model introduced in [25]. Zhao et al. [26] proposed an extended version of the LDA model, called TwitterLDA, which aims to detect the topics of short messages using only unlabelled data. Their approach relies on distinguishing between background word (words which occur in every topic), and content words (words specific to a topic). Experiments comparing TwitterLDA with traditional news media (e.g. New York Times) showed promising results outperforming various other topic models. Mehrotra et al. [27] proposed various pooling schemas for improving the performance of the original LDA model for topic classification. These pooling strategies aim to aggregate Microposts into longer documents (called "macro-documents"), which are more suitable for training LDA based models. The evaluated pooling strategies were: author-wise pooling (pooling Microposts according to an author), burst-score wise pooling (pooling Microposts according to a burst-score), temporal pooling (pooling Microposts which are posted during major events by a large number of users), hashtag-based pooling (pooling Microposts according to a hashtag). Experimental results on three different datasets suggest that hashtag-based pooling leads to drastically improved topic modelling over unpooled schemes.

Ramage et al. [28, 29], on the contrary, utilised annotated data for topic modelling. Ramage et al. [28] introduced the LabelledLDA model, which extends the original LDA model by defining a one-to-one correspondence between LDA's latent topics and social media tags. Experimental results on a credit attribution problem, extracting tag-specific snippets from del.icio.us, showed promising results, outperforming supervised classifiers such as Support Vector Machines.

Ramage et al. [29] further performed an extrinsic evaluation of the LabelledLDA model on a user recommendation task. In this case, the Microposts were classified according

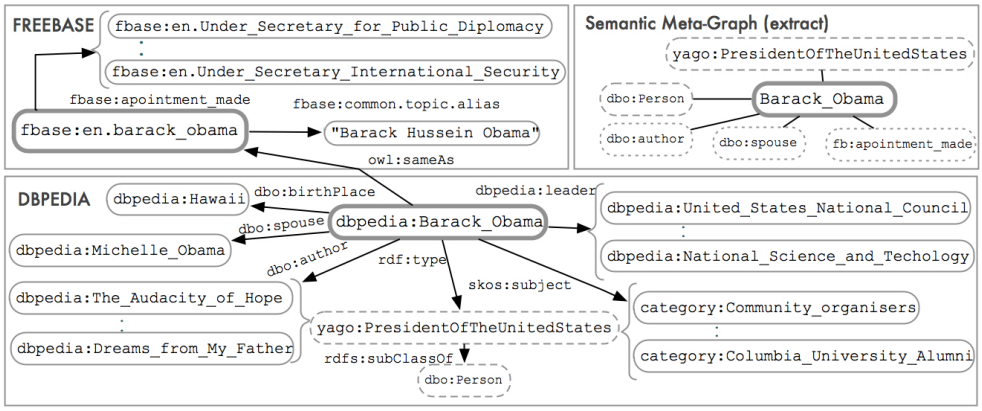

Figure 2: Deriving a semantic meta-graph from multiple linked LOD KSs.

to several dimensions including, e.g., style, substance, status, and other social characteristics of posts. Experimental results showed promising results, achieving a performance comparable to those obtained using term frequency-inverse document frequency (TF-IDF) feature vectors built on tokenised Microposts.

The third approach, *classification models*, is based on discriminative machine learning algorithms. Lin et al. [30] proposed to combine a language model with a supervised classifier for predicting the hashtags characterising a Twitter post. The features used for classification consisted of the perplexity of the unseen Microposts. Tao et al. [31] studied different topic-dependent and topic-independent features for topic classification. The topic-dependent features aimed to capture the relevance of the features to a topic (using keyword-based (lexical) and semantic-based relevance features). While the topic-independent features exploited various syntactic (e.g. has hashtag) and semantic (number of entities, number of distinct entity types) Microposts characteristics. Experimental results in the context of microblog search revealed that the topic-dependent features (the semantic relevance features) play an important role in this task, outperforming approaches which do not consider them.

3.2. Semantic Linking using Social Knowledge Sources

In this section we review the approaches which rely on the use of Knowledge Sources (DBpedia or Freebase) for semantic linking.

In light of the classification we used in Cano et al. [13], we distinguish between two main groups: approaches exploiting *local features*, and approaches exploiting the *link structure of KSs*.

In the first class of approaches, Genc et al. [9] proposed a model for mapping Microposts to the most similar Wikipedia articles, employing a simple *BoW* representation for the text content. Their approach comprises two steps: mapping Microposts to Wikipedia pages; and computing the semantic distance between Microposts. For the computation of semantic distance, a new measure is proposed, which approximates the distance between Microposts by the link distance measure computed between the corresponding Wikipedia pages. Experimental results showed that this new distance measure outperforms the String Edit Distance ([32]) and Latent Semantic Analysis ([33]).

Song et al. [34] proposed a probabilistic approach for mapping the terms within Microposts to the most likely resources in Probase KS ([35]). These resources were furthermore used as additional features in a clustering algorithm, achieving superior results to the simple *BoW* approach.

Shin et al. [36] proposed a graph-based approach for detecting persistent topics (PT) from Microposts, which correspond to topics of long-term, steady interest to a user. For their graph based approach they introduced two novel scoring functions that measure the properties inherent to PT terms: regularity and topicality. They allow to distinguish between terms that represent persistent topics, and terms which appear in static documents. Experimental results showed that this approach outperformed other existing alternatives (including LDA and keyword extraction models).

Muñoz García et al. [10] proposed an unsupervised approach for assigning topics to entities within Microposts written in Spanish. Their approach first employs the Sem4Tags *POS* tagger ([37]) for assigning POS tags to a Micropost. Following this process, a list of key phrases are identified, and the corresponding topics (DBpedia resource URIs) assigned to them. This topic recognition phase further exploits only local metadata, such as *BoW* features extracted from the keywords and contextual information in the form of neighbouring words to the keyword.

Vitale et al. [38] proposed a clustering-based approach which enriches the *BoW* representation of the Micropost using named entities extracted by the proposed Tagme system. The main idea behind Tagme is to assign the most likely topic to an entity, by taking into account the similarity between the topics returned by Tagme and Wikipedia categories for top-few categories. Experimental results showed that incorporating these new *BoE* features into topic classification significantly outperformed approaches using *BoW* features only.

P. N. Mendes and Sheth. [39] proposed the Topical Social Sensor system, which allows users to subscribe to hashtags and DBpedia concepts to receive updates regarding these topics. Their approach relies on linking a Micropost to DBpedia concepts derived from the entities contained in it. One of the main applications of the system is to detect the peak of a topic defined a priori.

Recently, in Varga et al. [14], we studied the similarity between linked KSs and Twitter using different content-based similarity measures. Their approach employs *BoW* and *BoE* features extracted from multiple linked KSs (such as DBpedia and Freebase). Experimental results demonstrated that these KSs contain complementary information for TC of Microposts, with the lexical features achieving the best performance.

For the second class of approaches, exploiting the link structure of KSs, Michelson and Macskassy [40] proposed a model that discovers topics of interest of Twitter users based on their Microposts. Their approach relies on first extracting and disambiguating the entities mentioned within a Micropost. Following this process, a sub-tree of Wikipedia categories is retrieved for each entity and the most likely topic assigned.

Milne and Witten. [7] proposed an approach for assigning Wikipedia resources to key concepts within Microposts. In their approach a Wikipedia article is considered as a concept. Following this representation, a machine learning approach is presented, which employs different Wikipedia n-gram and Wikipedia link-based features.

Xu and Oard [41] proposed a clustering-based approach which maps terms in Microposts to Wikipedia articles. To achieve this, their approach leverages the linking history of Wikipedia and the terms' textual context information to disambiguate the terms' meaning.

Recently, in Cano et al. [13], we demonstrated that exploiting the semantic information about entities from DBpedia and Freebase is beneficial. In particular, incorporating additional semantic information about entities in terms of properties and concepts can further improve the performance of a topic classifier against the approach using Twitter data only.

There is little work on classifying blogposts into topics ([42]). Husby and Barbosa [42] demonstrated that selecting data from Freebase using distant supervision, in addition to incorporating features about named entities is beneficial for TC.

Although previous work have focused on exploiting the semantic information from linked KSs for TC, the majority of these approaches still exploit a single KS. There is only little work which exploits multiple, linked KSs ([13, 14]). In Varga et al. [14], we presented an approach which makes use only of the data within KSs, ignoring the semantic information about entities present in KSs. In Cano et al. [13], we exploited a specific semantic graph defined in KSs (called *resource meta-graph*) which groups entities together by their types. In contrast to our previous work, in this paper we focus on understanding the usefulness of different semantic graphs defined in KSs. For this reason we extended our previous framework ([13]) by exploiting semantic information from these different semantic meta-graphs, and examined their usefulness for TC and measuring topic similarity.

In particular, our framework proposes novel weighting strategies for the explored semantic graphs. These weights can further be used to filter on KS semantic features relevant to a Micropost. This feature selection strategy also largely differs from state-of-the-art feature selection techniques (Forman et al. [43]) used in text classification, as they typically make use of the scores obtained for the features based on the text content only (e.g. occurrences of a feature in training positive- and negative-class training examples separately).

4. Framework for Topic Classification of Microposts

We now describe an extension of our TC framework proposed in [13]. This extension exploits a new type of semantic graph structure defined in KSs, named *category meta-graph*, and employs a novel set of semantic features derived from this graph for TC.

As depicted in Figure 3, our framework makes use of multiple linked KSs (from LOD) for TC of Microposts. The main stages of this framework can be summarised as follows:

1) *dataset collection and content modelling*

2) *context modelling*

   2.1) *dataset enrichment*

   2.2) *semantic feature derivation from different semantic meta-graphs*

3) *construction of a topic classifier based on the semantic features obtained.*

4.1. Dataset Collection and Content Modelling

In the first stage of our framework, *data collection*, data from both Twitter and KSs is retrieved. The Twitter dataset comprises a set of topically annotated Microposts. Conversely, the KSs dataset is build from a set of articles relevant to a given topic extracted from multiple, linked LOD KSs. This study considers two linked KSs (from LOD), namely DBpedia (DB) and Freebase (FB), which are applied both independently and as a merged KS. Therefore we consider three cross-source scenarios for the use of these KS articles: i) *DB* - from DBpedia only; ii) *FB* - from Freebase only; and iii) *DB-FB* - from both DBpedia and Freebase.

Having the documents selected from both KS and Twitter, a simple *BoW* representation is employed for modelling the content of these documents. This allows these datasets to be represented based on what is discussed in these documents. In order to capture the importance of each word mentioned in these documents, the TF-IDF weighting schema is applied.

4.2. Context Modelling

The second step of our framework aims to enrich the representation of both KS and Twitter documents using information about the entities and concepts mentioned in these documents.

In order to achieve this, two main steps are first performed: (i) *entity extraction* - employing the OpenCalais[12] and Zemanta[13] services for extracting the named entities in the documents; and (ii) *semantic mapping* - where the obtained named entities are mapped to their KS resource counterpart if it exists[14].

Following this process, different *semantic meta-graphs* are exploited from the different KSs, and a set of semantic features derived, leveraging the rich semantic information about entities described in these semantic meta-graphs. This stage comprises two steps: (i) *semantic meta-graph construction* and (ii) *semantic feature creation*. In the following subsections we discuss each of these steps.

4.2.1. Semantic Meta-Graphs Construction

The mapping of entities to DBpedia and Freebase URIs allows the incorporation of rich semantic representations into a topic classifier. In particular, the presented DB and FB KSs provide a rich source of structured information about concepts.

Figure 2 presents an overview of the semantic features extracted for the entity *Obama*. Similar to our previous work ([13]), rather than focusing on the <*subject, predicate, object*> instances associated with a resource, we focus on each triple's semantic structure at a meta-level, and for that we introduce two semantic meta-graphs: the *resource meta-graph* and the *category meta-graph*.

The first proposed in our original framework ([13]), *resource meta-graph*, exploits semantic information about an entity's KS resource. The second is the *category meta-graph* which exploits the semantic information extracted from the Wikipedia categories to which an entity belongs. This second graph can be effectively considered as a subset of the first one, as it groups similar entities belonging to the same topic under the same label. The *category meta-graph* thus categorises entities into more granular taxonomies.

We define these semantic meta graphs as follows:

Definition (Resource Meta Graph)**.** *is a sequence of tuples* $\mathtt{G} := (R, P, Cls, Y)$ *where*

- $\mathtt{R}$, $\mathtt{P}$, $\mathtt{Cls}$ *are finite sets whose elements are* `resources`, `properties` *and* `classes` *;*
- $\mathtt{Y}$ *is the ternary relation* $\mathtt{Y} \subseteq \mathtt{R} \times \mathtt{P} \times \mathtt{Cls}$ *representing a hypergraph with ternary edges. The hypergraph of a Resource Meta Graph* $\mathtt{Y}$ *is defined as a tripartite graph* $\mathtt{H(Y)} = \langle \mathtt{V}, \mathtt{D} \rangle$ *where the vertices are* $\mathtt{V} = \mathtt{R} \cup \mathtt{P} \cup \mathtt{Cls}$, *and the edges are:* $\mathtt{D} = \{\{r, p, cls\} \mid (r, p, cls) \in Y\}$.

A resource meta-graph provides information regarding the set of ontologies and properties used in the semantic definition of a given resource. The meta-graph of a given entity e can be represented as the sequence of tuples $G(e) = (R, P, Cls, Y')$, which is the aggregation of all resources, properties and classes related to this entity. In addition, we introduce two further notations: $R(cls) = \{e_1, \ldots, e_n\}$ for referring to the set of all entity resources whose *rdf:type* is class *cls*; and $R'(cls) = \{e_1, \ldots, e_m\}$ for denoting the set of entity resources whose types are specialisations of *cls*'s parent type (i.e. resources whose *rdf:type*s are siblings of *cls*).

Definition (Category Meta Graph)**.** *represents a qualified subset of the resource meta graph G in which all classes are of type dbCat:concept (skos:Concept). We define it as follows:* $G_{cat} := (R, P, Cat', Y)$ *where Cat' is a finite set whose elements are classes of type dbCat:concept.*

Class	Category
dbOwl:Person	*dbCat*:Presidents_of_the_United_States
dbOwl:Author	*dbCat*:Obama_family
dbOwl:OfficeHolder	*dbCat*:Harvard_Law_School_alumni
yago:LivingPeople	*dbCat*:Democratic_Party_Presidents_of_the_United_States
yago:President	*dbCat*:United_States_presidential_candidates,_2012

Table 2: Top 5 features extracted from the DBpedia KS for the entity *Obama* of type Person.

For the sake of comparison, Table 2 and Table 3 present the top few class and category features derived from these graphs for two different entity types (*Obama* of type Person, and *Syria* of type Country). As we can observe, the *dbCat* features group entities by topic, while the *dbOwl* features group entities by type[15].

Class	Category
dbOwl:Place	*dbCat*:Countries_of_the_Mediterranean_Sea
dbOwl:PopulatedPlace	*dbCat*:Arabic-speaking_countries_and_territories
yago:Country	*dbCat*:Eastern_Mediterranean_countries
yago:YagoGeoEntity	*dbCat*:Member_states_of_the_United_Nations
yago:MiddleEasternCountries	*dbCat*:Western_Asian_countries

Table 3: Top 5 semantic features extracted from the DBpedia KS for the entity *Syria* of type Country.

In light of the proposed three KS cross-source scenarios we construct three different *resource meta-graphs*: (i) one from *DB* using the *dbOwl* and *yago* ontologies; (ii) one from *FB* using the *fbOnt* ontology; and (iii) another one from *DB-FB* using the joint ontologies. For the joint scenario we use the concepts from *dbOwl* ontology together with the the classes obtained after mapping the *yago* and *fbOnt* ontologies[16]. For the *category meta-graphs* we derived a concept graph only from DBpedia, given that there is no category structure defined in Freebase. The three category meta graphs in this case correspond to (i) one from *DB* using the *dbCat* categories; (ii) one from *FB* using the *dbCat* categories obtained after mapping the FB URIs to DB URIs (iii) another one from *DB-FB* using *dbCat* categories.

[12] OpenCalais, http://www.opencalais.com

[13] Zemanta, http://zemanta.com

[14] Following this process, the percentage of entities without a deferenced URI is 35% in DBpedia, 40% in Freebase, and 36% in Twitter.

[15] Further statistics about these semantic features are provided in Table 5.

[16] The mapping of Freebase entity classes to the most likely Yago classes was done by a combined element and instance based technique (`www.l3s.de/~demidova/students/master_oelze.pdf` ([44]) and is available at http://iqp.l3s.uni-hannover.de/yagof.html).

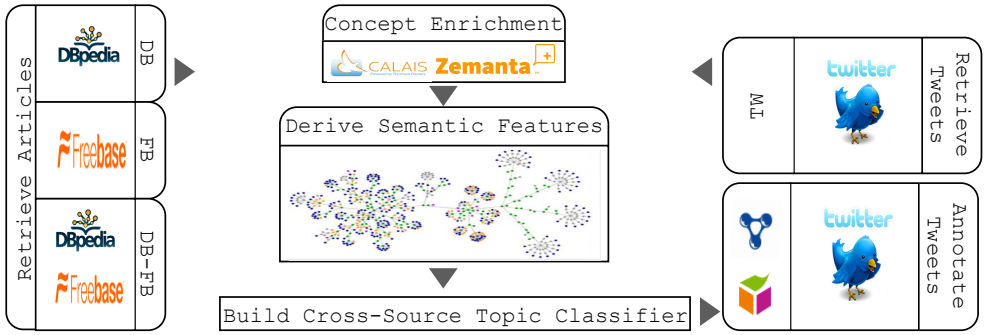

Figure 3: Architecture of cross-source TC framework using semantic features derived from semantic meta-graphs.

4.2.2. Semantic Feature Creation

Once a semantic meta graph has been constructed for a given entity, three main features can be derived from it: *class*, *category* and *property* features. Among these features the *class* and *category* features are particular to a semantic meta-graph: *class* being extracted from the *resource meta-graph*, and *category* being derived from the *category meta-graph*; while the *property* features are common to both meta-graphs.

We now describe each semantic feature as follows:

- Semantic class features (*Cls*): Extracted from the *resource meta-graph*, this feature set consists of all the classes an entity refers to. This set captures fine-grained information about this entity. For example, for Barack Obama these features would be *yago*:PresidentsOfTheUnitedStates, *fbOnt*:/book/author, *yago*:LivingPeople, and *dbOwl*:Person. Our main intuition is that the relevance of an entity to a given topic could be inferred from an entity's class type. For example, the class *yago*:PresidentsOfTheUnitedStates could be considered more relevant to the topic *violence*, than the class *yago*:Singer.

- Semantic category features (*Cat*): Extracted from the *category meta-graph*, this feature set captures the Wikipedia categories an entity is related to. Similar to the semantic classes, these categories provide additional fine-grained information about entities, as entities about similar topics are grouped together in categories. For example, for Barack Obama these category features would be *dbCat*:American_political_writers, *dbCat*:People_from_Honolulu,_Hawaii.

- Semantic property features (*P*): Common to both semantic meta-graphs, this feature set captures all the properties an entity is associated with. Our intuition is that given a context, certain properties of an entity may be more indicative of this entity's relevance to a topic than others. For example, given the role of Tahrir Square in the Egyptian revolution, properties such as *dcterms:subject* could be more topically informative than *geo:geometry*. The relevance of a property to a given topic can be derived from the semantic structure of a KS graph by considering the approach proposed in Subsection 4.3.1.

4.3. Supervised Topic Classifier Creation

The final stage of the framework aims to build supervised topic classifiers corresponding to the different cross-source scenarios, which make use of the generated KS semantic features. The Support Vector Machine (SVM) ([45]) with polynomial kernel was selected as a base classifier, which we detail in Section 8.

For incorporating the presented KS semantic features into a topic classifier, this framework employs different *weighting strategies* for the semantic features and feature combinations, as well as different *semantic augmentation strategies* for extending the initial feature spaces of both KS and Twitter documents. In this section we review these two strategies -originally proposed in ([13]) - and present their adoption for the *category meta-graph*.

4.3.1. Semantic Feature Weighting Strategies

The goal of the feature weighting strategies is to capture the importance of the semantic features for a given topic, based on the structure of the KS ontologies.

In the following we present an overview of the weighting strategies applied for the different semantic features derived from the *resource meta-graph* and *category meta-graph*:

- Semantic Feature Frequency (*W-Freq*): This weighting strategy provides a light-weight approach for weighting the different semantic features f (ontological *class*, *category* and *property* features) derived for entities. This weighting function aims to enrich the feature space of a document (i.e KSs' article, or Tweet) x by considering all the semantic meta-graphs extracted from the entity resources appearing in this document.

  Formally, the frequency of a semantic feature f in a given document x with Laplace smoothing can be defined as follows:

$$W - Freq_x(f) = \frac{N_x(f) + 1}{\sum_{f' \in F} N_x(f') + |F|} \quad (1)$$

  where $N_x(f)$ is the number of times feature f appears in all the semantic meta-graphs associated to document x; and F

is the semantic feature's vocabulary. This weighting function captures the relative importance of a document's semantic features against the rest of the corpus; while the normalisation prevents bias toward longer documents.

While the *W-Freq* (semantic feature frequency) weighting function depends on the occurrences of features in a particular document, other generalised weighting information can be derived from a KS's semantic structure to characterise a semantic meta-graph. To derive a weighted semantic meta-graph the following W-SG weighting strategy is proposed.

- Class/Category-Property Co-Occurrence Frequency (*W-SG*): The rationale behind this weighting strategy is to model the relative importance of a property p (e.g. *dbOwl*:leader) to a given class cls (*yago*:President) or category cat (*dbCat*:United_States_presidential_candidates,_2012), together with the generality of the property in a KS's graph.

  This weighting function computes how specific and how general a *property* is to a given *class* or *category* based on a set of semantically related resources derived from a KS's graph.

  In particular, given the semantic meta-graph of an entity e (i.e. $G(e)$), the relative importance of a property $p \in G(e)$ to a given class $cls \in G(e)$ in a KS graph $\mathcal{G}_{KS}$ can be computed by first defining the specificity of p to cls as follows:

$$specificity_{KS}(p, cls) = \frac{N_p(R(cls))}{N(R(cls))} \tag{2}$$

  where $N_p(R(cls))$ is the number of times property p appears in all resources of type cls in the KS graph $\mathcal{G}_{KS}$, and $N(R(cls)$ is the number of resources of type c in $\mathcal{G}_{KS}$. This measure captures the probability of the property p being assigned to an entity resource of type cls.

  For example for the *Obama* entity, considering the *dbOwl*:leader property and *yago*:President class, the specificity value of *dbOwl*:leader in the DBpedia graph $\mathcal{G}_{DB}$ is computed as follows:

$$\begin{aligned} &\text{specificity\_DB(dbOwl:leader, yago:President)} \\ &= \{| <?headofstate, dbOwl:leader, ?leader>, \\ &<?headofstate, rdf:type, yago:President> \in \mathcal{G}\_DB|\} / \\ &\{| <?headofstate, rdf:type, yago:President> \in \mathcal{G}\_DB|\} \end{aligned} \tag{3}$$

  As indicated in Equation 2, the computation of the specificity value is independent of the entity e and differs according to the KS graph from which it is derived[17]. Higher specificity values indicate that the property p occurs frequently in resources of the given class cls.

[17] It might be worth mentioning that for each entity resource the specificity values for the properties are the same, capturing in this way the generalisation of the property for the same concept type.

  Conversely, the generality measure captures the specialisation of a property p to a given class cls, by computing the property's frequency within other semantically related classes $R'(cls)$. The generality measure of a property p to a class cls in a KS graph $\mathcal{G}_{KS}$ is defined, as follows:

$$generality_{KS}(p, cls) = \frac{N(R'(cls))}{N_p(R'(cls))} \tag{4}$$

  where $N(R'(cls)$ is the number of resources whose type is either cls or a specialisation of cls's parent classes. This measure captures the relative generalisation of a property p to a broader set of specialised sibling classes derived from cls, and its computation is independent of the entity e. In this case the generality of property *dbOwl*:leader given the class *yago*:President for the *DB* graph is computed as:

$$\begin{aligned} &\text{generality\_DB(dbOwl:leader, yago:President) =} \\ &\{| <yago:President, rdf:subClassOf, ?parent>, \\ &<?group, rdf:subClassOf, ?parent> \\ &<?agroup, rdf:type, ?group> \\ &\in \mathcal{G}\_DB|\} / \\ &\{| <yago:President, rdf:subClassOf, ?parent>, \\ &<?group, rdf:subClassOf, ?parent> \\ &<?agroup, rdf:type, ?group> \\ &<?agroup, dbOwl:leader, ?leader> \in \mathcal{G}\_DB|\}| \} \end{aligned} \tag{5}$$

  Higher generality values indicate that a property spans over multiple classes, and is less specific to a given class cls. These two measures (generality and specificity) of a property p to a given class cls are combined as follows:

$$W\text{-}SG(p, cls) = specificity(p, cls) \times generality(p, cls) \tag{6}$$

4.4. Incorporating Semantic Features into TC's Feature Space

This section provides an overview of the semantic augmentation strategies supported by the TC framework proposed in [13], and presents its extension to the *category meta-graph*. Examples for the various semantic features, feature combinations and semantic augmentation strategies employed for the entity *Obama* are provided in Table 4.

4.4.1. Semantic augmentation

This strategy (F'_{A1}) augments the initial lexical features (e.g *BoW* and *BoE* features) of the datasets with additional semantic information extracted for the entities appearing in them.

In the case of the *resource meta-graph*, for both *Cls* and *P* features, the original lexical feature set F has been extended with a set of unique *Cls* (including for e.g. *dbOwl*:Author) and *P* (including for e.g. *dbOwl*:writer) features derived from this graph. In this case, the expanded feature space vocabulary size becomes $|F'_{A1_{Cls}}| = |F| + |F_{cls}|$ for the *Cls* features and $|F'_{A1_P}| = |F| + |F_p|$ for the *P* features, where $|F_{cls}|$ denotes the

	augmentation strategy	feature name	feature value
F	Baseline	*BoW*	*Obama*
F'_{A1}	P(W-Freq)	P_1	f_{W-Freq}(*dbOwl*:leader)
	P(W-Freq)	P_2	f_{W-Freq}(*dbOwl*:writer)
	P(W-SG)	P_1	f_{W-SG}(*dbOwl*:leader, *yago*:President)
	P(W-SG)	P_2	f_{W-SG}(*dbOwl*:writer, *dbOwl*:Author)
	Cls(W-Freq)	Cls_1	f_{W-Freq}(*yago*:President)
	Cls(W-Freq)	Cls_2	f_{W-Freq}(*dbOwl*:Author)
	Cls+P(W-SG)	Cls_1+P_2	f_{W-Freq}(*yago*:President), f_{W-SG}(*dbOwl*:writer, *yago*:President)
	Cls+P(W-SG)	Cls_2+P_1	f_{W-Freq}(*dbOwl*:Author), f_{W-SG}(*dbOwl*:leader, *dbOwl*:Author)
F'_{A2}	parent(Cls)(W-Freq)	parent(Cls_1)	f_{W-Freq}(*yago*:HeadOfState)
	parent(Cls)(W-Freq)	parent(Cls_2)	f_{W-Freq}(*dbOwl*:Thing)
	parent(Cls)(W-Freq)+P(W-SG)	parent(Cls_1)+P_2	f_{W-Freq}(*yago*:HeadOfState), f_{W-SG}(*dbOwl*:writer, *yago*:President)
	parent(Cls)(W-Freq)+P(W-SG)	parent(Cls_2)+P_1	f_{W-Freq}(*dbOwl*:Thing), f_{W-SG}(*dbOwl*:leader, *dbOwl*:Author)

Table 4: Example semantic augmentation strategies for the entity *Obama* using semantic features derived from *resource meta-graph*. The first column stands for the augmentation strategies used to incorporate semantic features into a TC classifier, the second column provides example features to which the augmentation strategies are applied, while the third column gives examples of possible values for each such feature.
As possible semantic features two different features are considered: P_1, P_2 corresponding to top semantic property features, and Cls_1, Cls_2 referring to top semantic class features for *Obama*. These features are considered alone as well as in combination (for e.g. $Cls_1 + P_2$). For the sake of completeness, in the first row, the original feature space denoted by F, consisting of *BoW* features, is also presented. For this feature representation no augmentation strategy is applied.
For the semantic features further two different augmentation strategies are presented: F'_{A1} extending the F features with semantic features, and F'_{A2} augmenting the F features with semantic features derived from the class hierarchies of KSs (e.g. considering the parent classes of a class (parent(Cls))). For both augmentation strategies two different weighting strategies are presented: W-Freq corresponding to the semantic feature frequency weighting, and W-SG corresponding to the class-property co-occurrence weighting. When these strategies are applied for the feature combinations (e.g. $Cls_1 + P_2$), two additional features are added to the TC classifier (e.g. f_{W-Freq}(*yago*:President), f_{W-SG}(*dbOwl*:writer, *yago*:President)).

total number of unique class features added and $|F_p|$ denotes the total number of unique property features added. Furthermore, for the combined *Cls*+ *P* feature set this augmentation strategy creates the novel feature set $F'_{A1_{Cls+P}}$, in which the feature set F is expanded with the properties' $< p, cls >$ tuple features derived from the semantic meta-graphs. In this case, the size of the expanded feature set is: $|F'_{A1_{Cls+P}}| = |F| + |F_p| \times |F_{cls}|$ (see the Cls_1+P_2 and Cls_2+P_1 examples in Table 4).

Similarly, for the *category meta-graph*, the expanded feature set becomes $|F'_{A1_{Cat}}| = |F| + |F_{cat}|$ for the *Cat* features, and $|F'_{A1_P}| = |F| + |F_p|$ for the *P* features. In this case, $|F_{cat}|$ refers to the total number of unique category features and $|F_p|$ denotes the total number of unique property features derived from this graph. Furthermore, for the combine *Cat*+ *P* feature set this augmentation strategy creates the novel feature set $F'_{A1_{Cat+P}}$, in which the feature set F is expanded with the properties' $< p, cat >$ tuple features derived from this semantic meta-graph. In this case, the size of the expanded feature set is: $|F'_{A1_{Cat+P}}| = |F| + |F_p| \times |F_{cat}|$.

4.4.2. Semantic augmentation with generalisation

This augmentation strategy (F'_{A2}) aims to further improve the generalization of a TC by exploiting the subsumption relation among classes within the DBpedia or Freebase ontologies.

In the case of the *resource meta-graph*, the feature set F is enhanced with the set of parent classes of *cls* where $cls \in Cls$. Therefore the size of the enhanced feature set $F'_{A2_{Cls}}$ is computed as $|F'_{A2_{Cls}}| = |F| + |F_{parent(cls)}|$, where $|F_{parent(cls)}|$ denotes the total number of unique parent classes of *cls*. Similarly, the enhanced feature set $F'_{A2_{Cls+P}}$ which uses the *Cls*+*P* features is built by adding the $< p, parent(cls) >$ tuple features. The size of the $F'_{A2_{Cls+P}}$ is therefore: $|F'_{A2_{Cls+P}}| = |F| + |F_p| \times |F_{parent(cls)}|$, where $|F_{parent(cls)}|$ denotes the total number of unique $parent(cls)$ classes derived from this graph.

When applying this strategy over the *category meta-graph*, however, the subsumption relations among the SKOS categories are considered. In this case, the expanded feature set size for the *Cat* features is $|F'_{A2_{Cat}}| = |F| + |F_{parent(cat)}|$, and for the combined *Cat*+*P* features is $|F'_{A2_{Cat+P}}| = |F| + |F_p| \times |F_{parent(cat)}|$ features. In this case $|F_{parent(cat)}|$ stands for the number of unique parent SKOS classes of *cat*, and $|F_p|$ denotes the number of unique properties extracted from this *category meta-graph*.

The following section introduces a set of metrics we proposed for analysing the relevance of these semantic features to the performance of a topic classifier.

5. Measuring Topic Similarity

In the previous section we presented different semantic features extracted from two semantic meta-graphs, which can be used to enhance the representation of documents with additional contextual information. While feature expansion can be beneficial in some cases, in others it rather undermines the performance of a topic classifier. In order to understand the relevance of the proposed semantic features to the performance of a topic classifier, in this section we study different topic similarity (also called domain similarity [15]) measures which can provide an estimate of the usefulness of these structures for TC.

Designing such topic similarity measures can be extremely important for a cross-source topic classifier, as they could help in providing an estimation of usefulness of a KS graph to previously unseen lexical data. One such example could be, the application of our model to a different genre, longer posts e.g. blogposts or Facebook comments. Another situation could be,

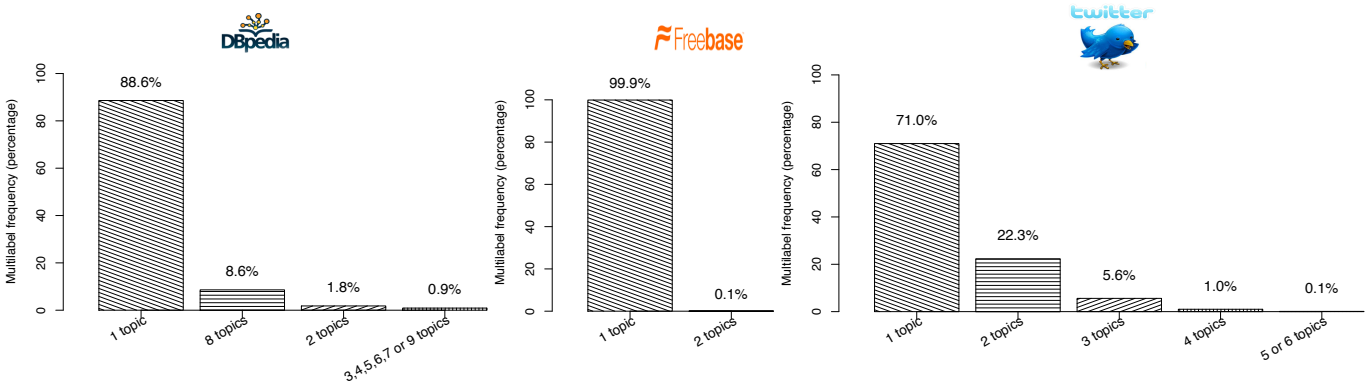

Figure 4: The multi-label distribution of the three gold standard datasets: DBpedia, Freebase and Twitter datasets. The numbers on the x axis represent the number of topics assigned to a document, raging from 1 topic to 9 topics. The numbers on the y axis correspond to the percentage of documents labelled with different topics ([14]).

in building a topic classifier for a new topic (e.g. Politics), in which case we want to have an a priori estimate of the similarity between KS data and Twitter data.

In light of the semantic features ($f = \{Cls, P, Cat\}$) and feature combinations ($f = \{Cls{+}P, Cat{+}P\}$) introduced in Section 4.2.1, we thus propose a set of entropy-based measures for topic similarity.

Entropy is an information theoretic measure which defines a probability distribution p[18] over a random variable X, capturing the dispersion of the variable f among the different classes in a given dataset T: $H_T(f) = -\sum_{f \in X} p(f) \log p(f)$. In our context, we introduce this measure, as it allows to capture the semantic ambiguity and uninformativeness of a topic based on the entities mentioned in the documents and the KS structure[19]. That is, entities that are evenly distributed over multiple KS concepts/categories will have high entropy and thus topics mentioning these entities are less focused (more ambiguous) in the subject(s) they discuss.

A summary of the proposed measures can be given as follows:

1. Topic-Class bag entropy (*Class Entropy*) : We took the class bag for each topic derived from the *resource meta-graphs* and measured the entropy of that class bag, capturing the dispersion of classes used for a particular topic. In this context, *low entropy* indicates a focused topic, while *high entropy* indicates an unfocused topic, which is more random in the subjects that this topic discusses. We define this measure as follows:
   $H_T(Cls) = -\sum_{j=1}^{|Cls_T|} p(cls_j) \log p(cls_j)$, where $p(cls_j)$ denotes the conditional probability of a concept cls_j, within the topic's concept bag Cls_T.

2. Topic-Category bag entropy *(Category Entropy)*: We constructed the category bag for each topic derived from the *category meta-graphs* and measured the entropy of that category bag, capturing the dispersion of categories used for a particular topic. In this context, *low entropy* indicates a focused topic, while *high entropy* indicates an unfocused topic, which is more random in the subjects that this topic discusses. We define this measure as follows:
   $H_T(Cat) = -\sum_{j=1}^{|Cat_T|} p(cat_j) \log p(cat_j)$, where $p(cat_j)$ denotes the conditional probability of a category cat_j, within the topic's category bag Cat_T.

3. Topic-Property bag entropy *(Property Entropy)*: We considered the property bag for each topic derived from the KS graphs and measured the entropy of that property bag, capturing the dispersion of properties used for a particular topic. In this context, *low entropy* indicates a focused topic, while *high entropy* indicates an unfocused topic which is more random in the subjects that this topic discusses. We define this measure as follows:
   $H_T(P) = -\sum_{j=1}^{|P_T|} p(p_j) \log p(p_j)$, where $p(p_j)$ denotes the conditional probability of a property p_j, within the topic's property bag P_T.

4. Topic-Entity bag entropy *(Entity Entropy)*: We took the entity bag for each topic extracted by the named entity recogniser and measured the entropy of that entity bag, capturing the dispersion of entities used for a particular topic. In this context, *low entropy* indicates a focused topic, while *high entropy* indicates an unfocused topic which is more random in the subjects that this topic discusses. We define this measure as follows:
   $H_T(Ent) = -\sum_{j=1}^{|Ent_T|} p(e_j) \log p(e_j)$, where $p(e_j)$ denotes the conditional probability of an entity e_j, within the topic's entity bag Ent_T.

5. Entity-Class entropy *(Entity-Class Entropy)*: We computed this measure for each topic, by considering the class bags for each entity mentioned in a topic, based on the extracted *resource meta-graphs*. This measure captures the

[18] In this paper we used the shorthand notation p for $Pr_p(X = f)$. We reserve the capital P for the property features.

[19] Compared to previous content-based similarity measures (e.g. cosine), these measures can explicitly measure the informativeness of a topic by capturing the dispersion of the entities among different KS classes/categories according to the various semantic meta-graphs presented.

dispersion of the entities in each class. That is, *low entropy* indicates that the topic is less ambiguous, consisting of entities belonging to few classes, while *high entropy* refers to higher ambiguity at the level of entities.
$H_T(Cls|E) = -\sum_{j=1}^{|E_T|} p(e_j)H_T(Cls|E = e_j)$, where $p(e_j)$ denotes the conditional probability of an entity e_j within the topic's entity bag E_T, and $H_T(Cls|E = e_j)$ refers to topic class entropy given the entity e_j.

6. Entity-Category entropy *(Entity-Category Entropy)*: In an analogy with the Entity-Class entropy, we computed this measure for each topic, by considering the category bags for each entity mentioned in a topic based on the extracted *category meta-graph*s. In this case, *low entropy* indicates that the topic is less ambiguous, consisting of entities belonging to few categories, while *high entropy* refers to higher ambiguity at the level of entities.
$H_T(Cat|E) = -\sum_{j=1}^{|E_T|} p(e_j)H_T(Cat|E = e_j)$, where $p(e_j)$ denotes the conditional probability of an entity e_j within the topic's entity bag E_T, and $H_T(Cat|E = e_j)$ refers to topic category entropy given the entity e_j.

7. Entity-Property entropy *(Entity-Property Entropy)*: Similarly, we took the property bag for each entity mentioned in a topic based on the extracted KS graphs. In this context, *low entropy* indicates that the topic is less ambiguous, consisting of entities being associated to few properties, while *high entropy* refers to higher ambiguity at the level of entities.
$H_T(P|E) = -\sum_{j=1}^{|E_T|} p(e_j)H_T(P|E = e_j)$, where $p(e_j)$ denotes the conditional probability of an entity e_j within the topic's entity bag E_T, and $H_T(P|E = e_j)$ refers to topic property entropy given the entity e_j.

8. Class-Property entropy *(Class-Property Entropy)*: We measured this by taking the property bag for each class appearing in each topic derived from the *resource meta-graphs*. In this context, *low entropy* indicates that a topic is less ambiguous, few properties spanning over multiple classes, while *high entropy* reveals high property diversity. The corresponding measure is defined as follows:
$H_T(P|Cls) = -\sum_{j=1}^{|Cls_T|} p(cls_j)H_T(P|Cls = cls_j)$, where $p(cls_j)$ denotes the conditional probability of a class cls_j within the topic's class bag Cls_T, and $H_T(P|Cls = cls_j)$ refers to topic property entropy for the class cls_j.

9. Category-Property entropy *(Category-Property Entropy)*: Similarly, we computed the category property entropy for each topic. In this context, *low entropy* indicates that a topic is less ambiguous, few properties spanning over multiple categories, while *high entropy* reveals high property diversity. The corresponding measure is defined as followed:
$H_T(P|Cat) = -\sum_{j=1}^{|Cat_T|} p(cat_j)H_T(P|Cat = cat_j)$, where $p(cat_j)$ denotes the conditional probability of a category cat_j within the topic's class bag Cat_T, and $H_T(P|Cat = cat_j)$ refers to topic property entropy for the category cat_j.

Considering that our aim is to estimate the performance of a topic classifier on a new unseen test dataset, we furthermore define the *entropy difference* (DE) measure for capturing the differences between a training dataset (T_{train}) - used to train a topic classifier-, and a test dataset (T_{test}) - used to test a topic classifier. Let T_{train} and T_{test} be the probability distributions estimated from the training and test datasets. For instance, given the *Cri* topic, and the cross-source topic classifier built on DBpedia KS data, the T_{train} training dataset corresponds to a dataset collected for the *Cri* topic from DBpedia, while the T_{test} dataset corresponds to the dataset collected from Twitter. According to the above entropy measures, for each semantic feature (e.g. $f = P$) and feature combination (e.g. $f = Cat + P$)[20], we define the entropy difference measure as follows:

$$DE(f, T_{train}, T_{test}) = |H_{T_{train}}(f) - H_{T_{test}(f)}|. \quad (7)$$

Intuitively, having features (e.g. *Cls* or *Cat*) with low DE values means that the features have similar values with respect to the training and test datasets. It is also expected that the lower the DE values are, the better the performance of a topic classifier.

These measures will be examined in Section 8 by correlating them with the performance of different topic classifiers. Our approach was evaluated in the Emergency Response and Violence Detection domains. The following section introduces the datasets in which the proposed framework and topic similarity metrics were tested.

6. Dataset

Our experiments make use of a Twitter dataset and two KSs datasets previously introduced in ([13, 14]). These datasets belong to the Emergency Response (ER) and Violence Detection (VD) domains. This section provides a brief description of these datasets.

The Twitter dataset was derived from Abel et al. [46]. It comprises Microposts collected over a period of two months starting in November 2010. Some of the notable Emergency events discussed in these messages are "Mexican drug war", "Egyptian revolution" and "Indonesia Volcano Eruption". This dataset was annotated with 17 OpenCalais topics[21]. Each of these topics consists of 1,000 randomly selected Tweets excluding re-tweets.

This resulted in a collection of 10,189 Tweets[22], which were manually re-annotated by two annotators, achieving an inter-annotator Kappa score of 71.25%. The final dataset is a multi-labelled dataset, consisting of Tweets annotated with up to 6

[20]For clarity we mention here that for the feature combinations (e.g. $f = Cat + P$) we employ the conditional entropy measure (e.g. $H_{T_{train}}(f = P|Cat)$), as this provides a natural way for capturing the relationships among multiple semantic features.

[21]The full list of topics include: Business & Finance, Disaster & Accident, Education, Entertainment & Culture, Environment, Health & Medical & Pharma, Hospitality & Recreation, Human Interest, Labor, Law & Crime, Politics, Religion & Belief, Social Issues, Sports, Technology & Internet, Weather and War & Conflict.

[22]For each given topic (e.g. *Cri*) then the number of positive instances is 1,000 and the number of negative instances is 9,189.

labels (as shown in Figure 4). For the purpose of these experiments we considered the following three topics related to ER and VD domains: "War & Conflict" (*War*), "Law & Crime" (*Cri*) and "Disaster & Accident" (*DisAcc*) topics.

The two KS datasets were compiled by querying each KS for 1,000 randomly selected resources for each of these three specific topics.

In the case of DBpedia, we SPARQL[23] queried for all resources whose categories (*dcterms:subject*) and sub-categories (*skos:narrower*) are similar to the topic of interest. The final DBpedia dataset comprised 9,465 articles[24]. While the majority of these articles belong to a single topic, less than 1.% of them are annotated with 3,4,5,6,7 or 9 topics (as shown in Figure 4). For querying the Freebase KS we used the Freebase Text Service API[25]. The final Freebase dataset comprises 16,915 articles[26], where the majority belong to a single topic (as shown in Figure 4). From the returned resources, we kept each resource's abstract or title to build the annotated dataset for the given topic.

Looking at the overall distribution of the entities in the three datasets, we observe that the KS datasets contain more entities than the Twitter dataset; the DBpedia dataset contains on average 22.24 entities per article, the Freebase dataset contains 8.14 entities per article, and the Twitter dataset contains on average 1.73 entities. The distribution of the top 15 entity types is presented in Figure 5, indicating that the most frequent entity types are Country, Person, Organization, Natural Feature, Position and City (as reported in [14]).

6.1. Dataset Pre-processing

The pre-processing steps for generating lexical features (i.e. *BoW*) included: i) removal of stopwords; ii) transformation of each word to lowercase iii) stemming each word using the Lovins stemmer ([47]).

For generating the semantic features we used the *BoE* features derived from a document. For each entity feature we looked for its resource representation in both KSs (DBpedia and Freebase). Using these KS resources we then generated different semantic meta-graphs (i.e. $\mathcal{G}_{DB}$ and $\mathcal{G}_{FB}$) as indicated in Subsection 4.2.1. In addition, for generating the semantic meta-graphs we disregarded properties which contained general information about an entity. Examples of such properties from DBpedia include: *rdfs:comment*, *abstract*, *wikiPageExternalLink*, and from Freebase include: *type/object*.

These feature spaces were also reduced by considering only the top 5 entity classes and top 5 properties derived from the different KS graphs for each OpenCalais' entity type (e.g. Person). The same strategy was used for reducing the number of category features. The statistics of the lexical and semantic features derived for these datasets is summarised in Table 5.

[23] http://www.w3.org/TR/rdf-sparql-query/

[24] For each given topic (e.g. *Cri*) then the number of positive DBpedia instances is 1,000, and the number of negative DBpedia documents equals to 8,465.

[25] Freebase, http://frebase.org

[26] For each given topic (e.g. *Cri*) then the number of positive Freebase instances is 1,000 and the number of negative Freebase instances is 15,915.

Comparing the statistics obtained for the *resource meta-graph* and *category meta-graph*, we observe that the frequency of *dbCat* categories are generally higher than those of *dbOwl* and *yago* classes. In addition, the average number of distinct categories for an entity (*cat*/*ent*) double the number of distinct classes per entities (*cls*/*ent*). This indicates that the categories form much larger clusters than the classes.

In addition, we observe that in all the three datasets the number of unique categories is higher than the number of unique classes. This indicates that the datasets are more diverse in terms of categories than in terms of classes.

Following the concept generalisation process, in the *resource meta-graph* the number of unique *dbClass* classes reduced by 76%, the number of unique *yagoClass* classes reduced by 92%, and the number of unique $fbClass$ classes by 88%. While in the *category meta-graph* the number of unique *dbCat* classes reduced by 42%.

7. Baseline Feature Sets and Models

We compared the performance of the proposed framework against several baseline models corresponding to state-of-the-art approaches for TC. These baseline models employ three baseline features namely: *BoW*, *BoE*, *part-of-speech* (*POS*). These features are typical baseline features for TC and have been evaluated in previous work ([13, 14]). In addition, a new baseline feature set (*bag-of-concepts* (*BoC*)) is also introduced. The *BoC* feature set consists of a collection of OpenCalais-derived semantic classes which are assigned to entities. As opposed to the semantic classes from KS semantic graphs, classes derived with the OpenCalais service represent more generic concepts.

A summary of these baseline features is given as follows:

- Bag-Of-Words Features (*BoW*): The first baseline feature set consists of simple unigram features, which captures our natural intuition to utilise what we know about a particular topic. The *BoW* features consists of a collection of words weighted by term frequency-inverse document frequency (TF-IDF). This weighting metric captures the relative importance of a word in a document to its use in the whole corpus. This feature set is very competitive, previous work on cross-source TC has shown that this features set outperforms on average the *BoE* features presented below [14].

- Bag-Of-Entities Features (*BoE*): This feature set extends the lexical *BoW* features with entities and concepts extracted using available annotation services, e.g. OpenCalais API, weighted by TF-IDF. These web services annotate each entity with generic types. For example in the case of *Obama*, rather than recognise it as being of type *dbOwl*:President the majority of these services will annotate this entity with the label *Person* ([48]). In this case the value of the *BoE* feature thus captures the co-occurrence of the entity and concept pairs $f_{BoE}(BarackObama \wedge Person)$.

- Part-of-Speech Features (*POS*): Similar to the *BoE* feature set, this feature set aims to capture some generalisation

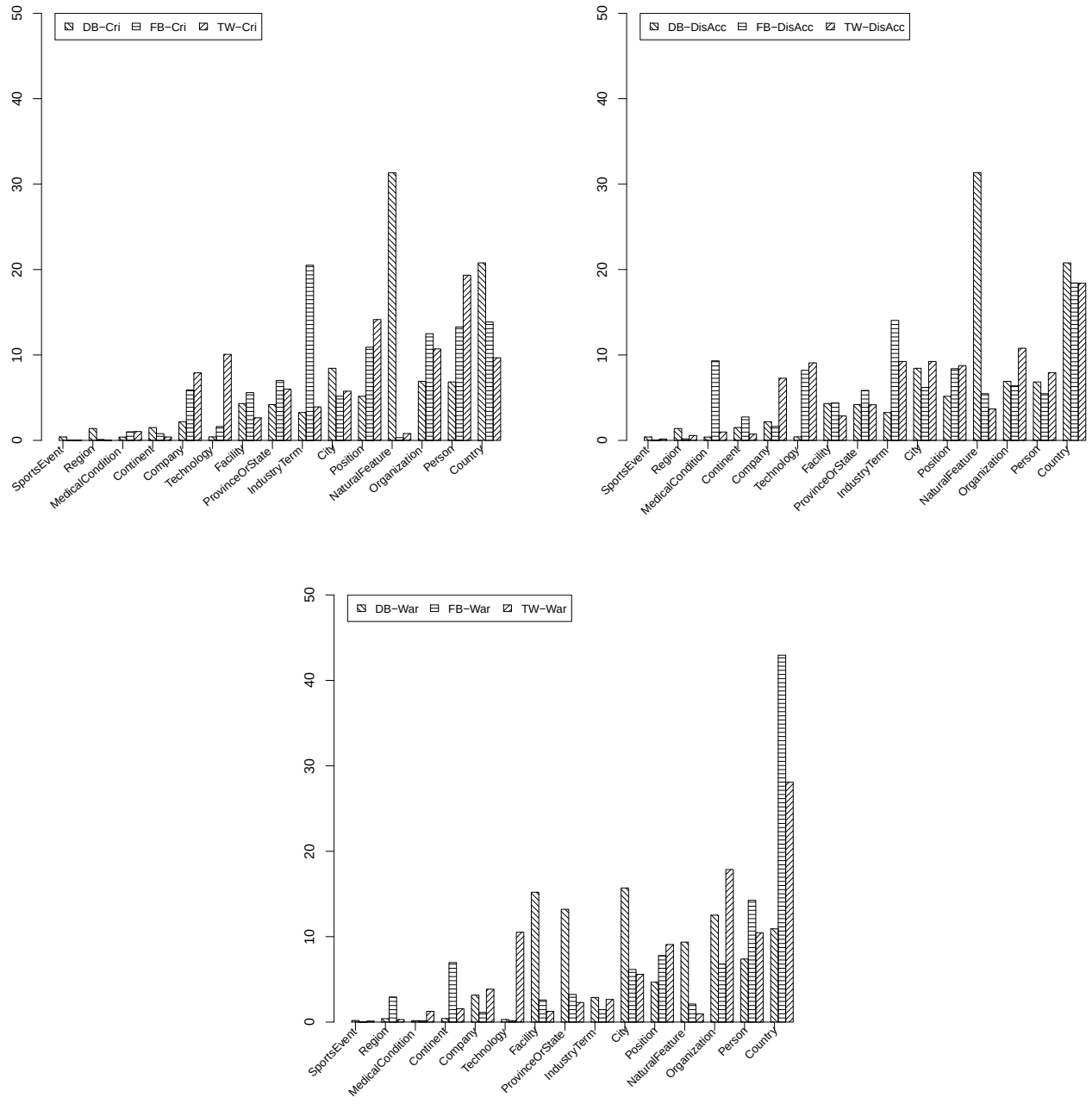

Figure 5: The distribution of top 15 concept types in the three gold standard datasets: DBpedia (DB), Freebase (FB) and Twitter (TW) datasets for the Crime (*Cri*), Disaster (*DisAcc*), and War (*War*) topics.

patterns for the words. For this reason, the syntactical patterns within the documents are considered and used to extend the lexical *BoW* features. In light of our previous work [13], we used Ritter et al.'s Twitter NLP Tool; this *POS* tagger has been trained on short text messages. For each POS tag again the TF-IDF weighting was assigned.

- Bag-Of-Concepts Features (*BoC*): This feature set extends the lexical *BoW* features with concepts extracted with the OpenCalais API. The API provides one single (often generic) concept type for each entity. For example assuming that Barack Obama is annotated as *Person* by OpenCalais, this feature set captures the presence of the Person class type $f_{BoC}(Person)$[27]. This new baseline feature set provides an alternative comparison between the newly proposed semantic meta-graph derived features (*Cls*) and those obtained from the OpenCalais service.

[27]This comparison also allows us to investigate whether modelling each entity with more than one KS concept (in our case 5) is more suitable for TC than with a single one.

	Statistics	*DB*			*FB*			*TW*		
		DisAcc	*Cri*	*War*	*DisAcc*	*Cri*	*War*	*DisAcc*	*Cri*	*War*
Lexical	BoW	8,837	8,837	8,504	2,078	4,596	2,574	3,218	3,197	2,781
	BoE	18,247	18,247	18,167	1,172	2,715	1,822	1,818	1,816	2,146
Semantic features	dbCls	119	119	124	39	47	48	80	85	68
	yagoCls	3,865	3,865	3,864	351	834	922	1,480	1,795	1,275
	fbCls	1,289	1,289	1,215	394	713	641	881	915	772
	dbCat†	9,275	9,275	8,796	783	1,844	1,807	3,252	3,878	3,087
	dbprop	4,105	4,105	4,215	1,229	1,849	1,871	2,544	2,457	2,422
	fbprop	1,090	1,090	1,065	420	586	554	834	869	696
	cls/ent	4.56	4.56	4.48	5.55	4.21	6.33	5.73	6.02	5.80
	cat/ent†	5.45	5.49	5.34	7.76	5.80	8.89	7.49	8.20	8.72
	prop/ent	26.56	26.56	26.29	39.65	33.97	41.78	36.99	32.62	36.17
	fbcls/ent	7.30	7.30	7.12	15.89	12.68	15.57	11.98	11.66	12.49
	fbprop/ent	10.08	10.08	9.76	23.44	17.06	23.05	16.93	16.65	17.97

Table 5: General statistics for the DBpedia (DB), Freebase (FB) and Twitter (TW) datasets used in the context of ER and VD for the two semantic meta-graphs analysed (*resource meta-graph* and *category meta-graph*). The rows labelled as *BoW* and *BoE* represent the size of the vocabulary of the *BoW* and *BoE* (without *BoW*) features. Statistics about the *resource meta-graph* (as reported in [13]): dbCls, yagoCls and fbCls stand for the unique number of classes extracted from the DBpedia and Freebase knowledge graphs. dbprop counts the number of unique DBpedia properties, and correspondingly fbprop counts the number of unique Freebase properties. Considering the *category meta-graph*: dbCat refers to the unique number of categories extracted from DBpedia knowledge graph.
cls/ent refers to the average number of *dbOwl* and *yago* classes per entity; cat/ent quantifies the average number of *dbCat* categories per entity, while fbcls/ent denotes the average number of *fbOnt* classes per entity. Similarly prop/ent denotes the average number of *dbOwl* and *yago* properties per entity, and fbprop/ent refers to the average number of *fbOnt* properties per entity.
The statistics highlighted with † correspond to the new *category meta-graph* analysed in this paper.

Considering the above baseline features, two strong baseline supervised machine learning models are employed[28]:

- *TW single source* topic classifier, in which an SVM topic classifier is built on Microposts only (TW)
- *KS cross source* topic classifier, in which an SVM topic classifier is built on KS (DBpedia and/or Freebase) data only.

8. Experimental Setup

In this section we present a series of experiments to evaluate the TC framework and topic similarity measures using the two semantic meta-graphs introduced in Subsection 4.2.1. In particular, these experiments aim to compare and contrast the results obtained for the *resource meta-graph* (used in our previous experiments [13]) with the results obtained for the newly introduced *category meta-graph*.

For evaluating the TC framework for the different single-source and cross-source scenarios, we took the commonly used one-vs-all approach ([50]). In this approach we decompose the multi-label problem into multiple independent binary classification problems.

Following this approach, each TC system was evaluated using 5-fold cross-validation. The training dataset for the *TW* TC system consisted of 80% of the original Twitter data. For the *KS* classifier the training set consisted of the full KS data. For the *KS + TW* classifier the full KS data was combined with 80% of Twitter data. Using Weka Software[29] we applied different discriminative classifiers including the Maximum Entropy, Perceptron and Support Vector Machine. After comparing the results of these classifiers we found the SVM with polynomial kernel to be that which achieved the best results. Therefore, in this paper we only report results for the SVM classifier.

In order to asses the usefulness of the different semantic meta-graph structures for TC we conducted a series of experiments. In the first set of experiments, we compared the performance of the topic classifiers using the *resource meta-graph* and *category meta-graph*. First, the results obtained for the single-source TC case are discussed in Subsection 8.1.1. Then we discuss the results obtained for the cross-source TC case in Subsection 8.1.2. The main research questions that we aim to address are *How does the performance of a topic classifier vary*

[28]The motivation behind the selection of these discriminative models, is that they correspond to typical baseline methods used in cross-source (multi-source) learning ([15]). Previous approaches on single-source TC are not directly comparable with our current setting and results. For instance, the majority of the generative LDA based approaches (e.g. TwitterLDA) were trained on unlabelled data only using simple *BoW* features. However our approach exploits labelled data from KSs, and further focuses on the evaluation of the usefulness of different semantic features for TC.

[29]Weka Software,http://www.cs.waikato.ac.nz/ml/weka/

Dataset	Semantic graph	Features	TW(*dbKS+fbKS*)			TW(*dbKS*)			TW(*fbKS*)		
			P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
War	Baseline	BOW	0.867	0.743	0.800	0.867	0.743	0.800	0.867	0.743	0.800
		POS	0.844	0.757	0.798	0.844	0.757	0.798	0.844	0.757	0.798
		BOE	0.857	0.761	**0.806**	0.857	0.761	**0.806**	0.857	0.761	**0.806**
		BOC†	0.839	0.735	0.784	0.839	0.735	0.784	0.839	0.735	0.784
	Resource	Cls(W-Freq)	0.864	0.727	0.790	0.867	0.736	0.796	0.873	0.744	0.803
		parent(Cls)(W-Freq)	0.859	0.734	0.792	0.862	0.730	0.791	0.874	0.743	0.803
		P(W-Freq/Cls)	0.874	0.743	0.803	0.872	0.739	0.800	0.869	0.742	0.800
		Cls+P(W-SG)	0.869	0.746	0.803	0.880	0.748	0.808	0.868	0.749	0.804
		parent(Cls)+P(W-SG)	0.871	0.745	0.803	0.868	0.745	0.802	0.873	0.754	0.809
		P(W-SG/Cls)	0.885	0.777	**0.828**	0.885	0.759	**0.817**	0.881	0.759	**0.816**
	Category†	Cat(W-Freq)†	0.882	0.763	0.818	0.882	0.767	0.820	0.879	0.763	0.817
		parent(Cat)(W-Freq)†	0.887	0.770	**0.824**	0.876	0.775	**0.822**	0.885	0.764	**0.820**
		P(W-Freq/Cat)†	0.871	0.759s	0.811	0.871	0.759	0.811	0.871	0.756	0.809
		Cat+P(W-SG)†	0.871	0.759	0.811	0.871	0.759	0.811	NA	NA	NA
		parent(Cat)+P(W-SG)†	0.879	0.762	0.816	0.879	0.762	0.816	NA	NA	NA
		P(W-SG/Cat)†	0.880	0.773	0.823	0.877	0.767	0.818	0.878	0.771	0.821
Cri	Baseline	BOW	0.715	0.521	0.602	0.715	0.521	0.602	0.715	0.521	0.602
		POS	0.667	0.541	0.597	0.667	0.541	0.597	0.667	0.541	0.597
		BOE	0.736	0.534	**0.619**	0.736	0.534	**0.619**	0.736	0.534	**0.619**
		BOC†	0.677	0.523	0.590	0.677	0.523	0.590	0.677	0.523	0.590
	Resource	Cls(W-Freq)	0.705	0.518	0.597	0.714	0.516	0.599	0.715	0.525	0.605
		parent(Cls)(W-Freq)	0.716	0.523	0.604	0.723	0.518	0.603	0.724	0.523	0.607
		P(W-Freq/Cls)	0.711	0.525	0.604	0.712	0.524	0.604	0.718	0.524	0.606
		Cls+P(W-SG)	0.709	0.521	0.601	0.712	0.517	0.599	0.717	0.522	0.604
		parent(Cls)+P(W-SG)	0.716	0.522	0.604	0.709	0.521	0.601	0.716	0.526	0.607
		P(W-SG/Cls)	0.729	0.547	**0.625**	0.716	0.534	**0.612**	0.731	0.532	**0.616**
	Category†	Cat(W-Freq)†	0.694	0.549	0.613	0.700	0.545	0.613	0.702	0.538	0.609
		parent(Cat)(W-Freq)†	0.698	0.547	0.613	0.698	0.547	0.613	0.693	0.536	0.605
		P(W-Freq/Cat)†	0.701	0.541	0.610	0.701	0.541	0.610	0.704	0.535	0.608
		Cat+P(W-SG)†	0.701	0.541	0.610	0.701	0.541	0.610	NA	NA	NA
		parent(Cat)+P(W-SG)†	0.710	0.543	**0.616**	0.710	0.543	**0.616**	NA	NA	NA
		P(W-SG/Cat)†	0.690	0.551	0.613	0.686	0.542	0.606	0.691	0.553	**0.614**
DisAcc	Baseline	BOW	0.800	0.637	0.709	0.800	0.637	0.709	0.800	0.637	0.709
		POS	0.746	0.652	0.696	0.746	0.652	0.696	0.746	0.652	0.696
		BOE	0.798	0.670	**0.728**	0.798	0.670	**0.728**	0.798	0.670	**0.728**
		BOC†	0.772	0.608	0.680	0.772	0.608	0.680	0.798	0.644	0.713
	Resource	Cls(W-Freq)	0.790	0.636	0.705	0.800	0.632	0.707	0.792	0.631	0.703
		parent(Cls)(W-Freq)	0.793	0.634	0.705	0.799	0.632	0.706	0.795	0.635	0.706
		P(W-Freq/Cls)	0.779	0.620	0.690	0.793	0.636	0.706	0.797	0.628	0.703
		Cls+P(W-SG)	0.799	0.629	0.704	0.804	0.636	0.710	0.797	0.637	0.708
		parent(Cls)+P(W-SG)	0.810	0.629	0.708	0.804	0.635	0.709	0.797	0.630	0.704
		P(W-SG/Cls)	0.808	0.656	**0.724**	0.811	0.644	**0.718**	0.800	0.646	**0.715**
	Category†	Cat(W-Freq)†	0.786	0.651	0.712	0.798	0.646	0.714	0.800	0.639	0.710
		parent(Cat)(W-Freq)†	0.788	0.655	0.716	0.788	0.655	0.716	0.788	0.655	**0.716**
		P(W-Freq/Cat)†	0.796	0.649	0.715	0.796	0.649	0.715	0.796	0.642	0.711
		Cat+P(W-SG)†	0.796	0.649	0.715	0.796	0.649	0.715	NA	NA	NA
		parent(Cat)+P(W-SG)†	0.805	0.650	**0.719**	0.805	0.650	**0.719**	NA	NA	NA
		P(W-SG/Cat)†	0.777	0.662	0.715	0.795	0.655	0.718	0.786	0.647	0.709

Table 6: The performance of the single-source TW SVM topic classifiers using different KSs ontologies (DBpedia *dbKS*'s ontologies, and Freebase *fbKS*'s ontology) and two semantic meta-graphs derived from these KSs (*resource meta-graph* (Resource) and *category meta-graph* (Category)). The results obtained for the semantic features derived for the *resource meta-graph* (reported in [13]) using the W-Freq weighting schema correspond to: class (Cls(W-Freq)), upper-class (parent(Cls)(W-Freq)) and property (P(W-Freq/Cls)); while using the W-SG weighting schema are: class-property co-occurrence (Cls+P(W-SG)), upper-class-property co-occurrence (parent(Cls)+P(W-SG)) and property (P(W-SG/Cls)). The results obtained for the semantic features derived for the *category meta-graph* using the W-Freq weighting schema are: category (Cat(W-Freq)), upper-category (parent(Cat)(W-Freq)) and property (P(W-Freq/Cat)); while using the W-SG weighting schema are: category-property co-occurrence (Cat+P(W-SG)), upper-category-property co-occurrence (parent(Cat)+P(W-SG)) and property (P(W-SG/Cat)).
The baseline models (Baseline) employed are bag-of-words (BOW), bag-of-entities (BOE), part-of-speech (POS) and bag-of-concepts (BOC).
The results marked with † correspond to the new results obtained for the newly introduced *category meta-graph*.

using different concept graphs? Which concept graph provides the most useful semantic features for TC of Microposts?

Next, our second set of analyses aim to investigate whether there are any differences in the roles (generalisation patterns) of semantic features derived from the two semantic concept graphs in TC. In this case, we address the research question *Are there differences in the roles of the concept graphs in the different TC scenarios?*

Finally, in the third set of experiments, we look at the roles of the semantic features in predicting the performance of a topic classifier. For this reason we proposed and compared various entropy-based measures using the semantic features which characterise a topic. We then correlated these entropy-based measures with the performance of SVM topic classifiers. In this case we investigate the questions of *Can we predict the performance of a topic classifier? Which topic similarity measure provides a better estimate on the performance of a topic classifier?*

8.1. Comparison of Multiple Semantic Structures for Topic Classification

We start our analysis by assessing the usefulness of the different semantic meta-graphs in both single-source TC (Section 8.1.1) and cross-source TC (Section 8.1.2) scenarios.

8.1.1. Evaluation of Semantic Concept Graphs in Single-Source Topic Classification

This section details the results obtained for the single-source TC case. In particular, it compares and contrasts the results reported in our previous work for the *resource meta-graph* ([13]) with the results obtained for the *category meta-graph* introduced in this paper.

In our experiments we employed three different single-source TW classifiers. These classifiers make use of a single KS ontol-

ogy: TW(*dbKS*) and TW(*fbKS*); and the combined KS ontologies: TW(*dbKS+fbKS*). In particular, in the case of the *resource meta-graph*, *dbKS* denotes the *dbOwl +yago* ontologies, while in the case of the *category meta-graph*, *dbKS* stands for the *dbCat* ontology. These classifiers are evaluated against several baseline models, as presented in Table 6.

Looking at the performance of the baseline models, we observe that the best performance was achieved by the *BoE* features, which performed better than the *BoC* and *BoW* features. Further, the *POS* features did not improve on the baseline model using only *BoW* features. An explanation for this could be that the language in Tweets is quite complex, and exhibits less regularity than longer texts used from KSs (KS abstracts).

Comparing the results obtained for the best baseline feature -*BoE* feature- with those for the semantic features derived from the two semantic meta-graphs, we observe that the best results were obtained for the *resource meta-graph* for the combined *TW*(*dbOwl+yago+fbOnt*) scenario using the *P* features with the *W-SG* weighting strategy, which significantly outperforms the baseline lexical features (t-test with $\alpha < 0.05$). As reported in our previous work [13], in the case of the *War* category, the F1 measure increases with 2.8% with respect to the *BoW* features and 2.2% with respect to the *BoE* features; in the case of the *Cri* category the F1 measure increases with 2.3% with respect to the *BoW* feature and 0.6% with respect to the *BoE* features, while in the case of *DisAcc* an improvement of 1.5% over the *BoW* features can be observed. Further, for both semantic meta-graphs, our novel class-property co-occurrence weighting schema (*W-SG*) for the properties (*P(W-SG)*) shows a significant improvement over the feature frequency strategy (*P(W-Freq)*) (t-test with $\alpha < 0.01$). These results demonstrate that capturing the importance of the property within a given semantic meta-graph (with respect to concepts in the *resource meta-graph* or to categories in the *category meta-graph*), improves the generality of the properties and the performance of the TC classifier for each topic.

While employing the *P* features have been shown to provide a positive gain over the baseline features for most of the topics, the usefulness of the semantic features and augmentation strategies merely depend on a number of factors. For instance, one of the factors which influences the performance of a TC classifier is the number of entities identified in a Micropost. For instance, in the case of the *War* topic, a higher number of entities have been extracted than for the other two topics. This can explain the higher gain achieved for this topic, resulted from a larger number of Microposts being enriched. Further, the lower performance achieved by the *Cls* features, could be due to the level of ambiguity (measured as cls/ent) of the *Cls* features and their discriminative power for a given topic. Looking at the Table 5, it can be observed that there are a larger number of property features defined in KSs for an entity ($prop/ent$) than for a class (cls/ent, $fbcls/ent$). This allows the incorporation of very fine grained information into TC, which indeed seems to improve the performance of the classifier upon the baseline features. In order to capture these factors and provide an insight into the usefulness of these features for topic classification, the reminder of the reader, we employed a set of topic similarity measures which we will evaluate in Subsection 8.3.

Inspecting the results obtained for the different taxonomies, we observe similar trends for the *resource meta-graph* and *category meta-graph*. That is, for both semantic graphs the *dbKS* ontologies (*dbOwl+yago* for *resource meta-graph*; and *dbCat* for *category meta-graph*) provide a significant improvement over the semantic features derived from *fbKS* ontology for the *War* and *DisAcc* topics, except for *Cri* (t-test with $\alpha < 0.05$). This could be explained by the fact that in the *Cri* topic the entities extracted from the *dbKS* graph are more ambiguous than those found within the *War* and *DisAcc* topics (see cls/ent values in Table 5). Similarly, the entities extracted from the *fbKS* are less ambiguous in the *Cri* topic than in the other two topics (see $fbcls/ent$ values in Table 5). The best overall results were obtained by the combined *dbOwl+yago+fbOnt* and *dbCat+fbOnt* ontologies using the property features, indicating that the three ontologies contain complementary information (properties) about the entities.

Further, we found that the augmentation strategies are beneficial for both semantic graphs. In the case of the *resource meta-graph*, we found different trends for the *fbOnt* and *dbOwl+yago* ontologies. When using *fbOnt* ontology, both (*parent(Cls)(W-Freq)* and *parent(Cls) + P(W-SG)*) showed a consistent improvement over the initial non-generalisation case (*Cls(W-Freq)* and *Cls+ P(W-SG)*) for each topic. However, when using the *dbOwl + yago* ontology encoding the very specific classes of the entities were found to be more beneficial for some topics (e.g. *War*). These results are understandable because after generalisation, the entities which have the same parent class in the KS graphs will be unified to the same semantic concept type, losing as a result the very specific meaning of the entity. In the case of *yago* ontology, the number of unique classes reduces with 92% after generalisation, while in *fbOnt*, the number of unique classes becomes 88% less. In the case of the *category meta-graph*, further, we found that the *parent(Cat)(W-Freq)* and *parent(Cat) + P(W-SG)* features significantly improved over the *Cat(W-Freq)* and *Cat+ P(W-SG)* features for each topic (t-test with $\alpha < 0.05$).

8.1.2. Evaluation of Semantic Concept Graphs in Multi-Source Topic Classification

This section continues with the description of the results obtained for the cross-source TC case. In particular, it compares and contrasts the results reported in our previous work for the *resource meta-graph* ([13]) with the results obtained for the *category meta-graph* introduced in this paper.

Based on the three scenarios analysed, in our experiments we employed six different cross-source TW classifiers. Among these cross-source classifiers, four make use of individual KS ontologies: DB making use of *dbKS*'s ontologies, FB making use of *fbKS*'s ontology, DB+TW exploiting *dbKS*'s ontologies, FB+TW employing *fbKS*'s ontology. The remaining two cross-source TC classifiers make use of the combined KS ontologies: DB+FB and DB+FB+TW. In particular, in the case of the *resource meta-graph*, *dbKS* denotes the *dbOwl +yago* ontologies, while in the case of the *category meta-graph*, *dbKS* stands for

Dataset	Semantic graph	Features	DB+FB			DB+FB+TW			DB			DB+TW			FB			FB+TW		
			P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
War	Baseline	BOW	0.420	0.011	**0.022**	0.955	0.861	0.905	0.208	0.049	**0.080**	0.877	0.723	0.793	0.678	0.136	0.226	0.851	0.722	**0.781**
		POS	0.217	0.006	0.013	0.952	0.880	**0.914**	0.258	0.034	0.061	0.859	0.744	0.797	0.597	0.148	**0.237**	0.809	0.746	0.776
		BOE	0.903	0.007	0.014	0.842	0.761	0.799	0.490	0.040	0.080	0.856	0.754	**0.802**	0.767	0.12	0.207	0.753	0.801	0.776
		BOC †	0.038	0.013	0.020	0.748	0.658	0.700	0.176	0.061	0.091	0.837	0.698	0.761	0.486	0.125	0.199	0.807	0.704	0.752
	Resource	Cls(W-Freq)	0.370	0.009	0.017	0.957	0.878	0.916	0.221	0.045	0.075	0.881	0.720	0.792	0.678	0.136	0.226	0.844	0.699	0.765
		parent(Cls)(W-Freq)	0.426	0.011	**0.022**	0.957	0.880	0.917	0.206	0.047	0.077	0.877	0.727	0.795	0.678	0.136	0.226	0.846	0.718	0.776
		P(W-Freq/Cls)	0.364	0.009	0.017	0.956	0.871	0.911	0.222	0.054	0.086	0.876	0.717	0.789	0.683	0.136	**0.227**	0.845	0.712	0.773
		Cls+P(W-SG)	0.422	0.011	0.021	0.956	0.864	0.908	0.195	0.043	0.071	0.878	0.726	0.795	0.673	0.136	0.226	0.844	0.723	0.779
		parent(Cls)+P(W-SG)	0.406	0.011	0.021	0.955	0.863	0.907	0.244	0.040	0.069	0.874	0.716	0.787	0.683	0.136	**0.227**	0.844	0.715	0.774
		P(W-SG)	0.902	0.006	0.013	0.967	0.879	**0.921**	0.303	0.062	**0.103**	0.874	0.731	**0.796**	0.670	0.136	0.226	0.850	0.732	**0.787**
	Category †	Cat(W-Freq)†	0.395	0.011	0.021	0.959	0.872	0.914	0.302	0.046	0.080	0.878	0.745	**0.806**	0.693	0.147	**0.243**	0.854	0.741	0.793
		parent(Cat)(W-Freq)†	0.418	0.013	0.024	0.960	0.874	0.915	0.326	0.064	**0.107**	0.875	0.734	0.799	0.679	0.147	0.242	0.852	0.736	0.789
		P(W-Freq/Cat)†	0.401	0.012	0.023	0.959	0.865	0.910	0.225	0.054	0.088	0.878	0.723	0.793	0.679	0.147	0.242	0.850	0.723	0.781
		Cat+P(W-SG)†	0.431	0.013	**0.026**	0.960	0.867	0.911	0.221	0.054	0.087	0.880	0.724	0.794	NA	NA	NA	NA	NA	NA
		parent(Cat)+P(W-SG)†	0.387	0.011	0.021	0.961	0.869	0.913	0.230	0.056	0.090	0.878	0.730	0.797	NA	NA	NA	NA	NA	NA
		P(W-SG/Cat)†	0.441	0.013	0.026	0.960	0.878	**0.917**	0.341	0.055	0.094	0.877	0.742	0.804	0.679	0.147	0.242	0.852	0.744	**0.794**
Cri	Baseline	BOW	0.489	0.013	0.025	0.944	0.857	0.898	0.071	0.006	**0.011**	0.718	0.477	0.573	0.747	0.143	0.240	0.723	0.489	0.583
		POS	0.448	0.013	0.025	0.950	0.860	**0.902**	0.069	0.005	0.009	0.676	0.527	0.592	0.695	0.150	**0.247**	0.667	0.517	0.582
		BOE	0.353	0.028	**0.052**	0.814	0.626	0.708	0.049	0.004	0.008	0.744	0.502	**0.600**	0.656	0.108	0.186	0.733	0.498	**0.593**
		BOC †	0.037	0.014	0.021	0.655	0.446	0.531	0.051	0.014	0.022	0.700	0.457	0.553	0.638	0.155	0.250	0.692	0.469	0.559
	Resource	Cls(W-Freq)	0.616	0.011	0.021	0.944	0.873	**0.907**	0.083	0.009	0.016	0.702	0.471	0.564	0.691	0.140	0.233	0.722	0.486	0.581
		parent(Cls)(W-Freq)	0.586	0.010	0.019	0.944	0.873	**0.907**	0.082	0.008	0.014	0.705	0.477	**0.569**	0.740	0.141	0.237	0.728	0.489	0.585
		P(W-Freq/Cls)	0.628	0.011	0.021	0.944	0.866	0.904	0.096	0.011	**0.019**	0.705	0.473	0.566	0.710	0.145	**0.241**	0.724	0.484	0.580
		Cls+P(W-SG)	0.663	0.013	0.026	0.945	0.858	0.899	0.062	0.006	0.011	0.703	0.464	0.559	0.726	0.140	0.235	0.728	0.490	0.586
		parent(Cls)+P(W-SG)	0.617	0.012	0.024	0.945	0.858	0.899	0.067	0.006	0.011	0.706	0.469	0.563	0.738	0.143	0.240	0.716	0.490	0.582
		P(W-SG/Cls)	0.666	0.014	**0.028**	0.944	0.864	0.903	0.126	0.009	0.016	0.699	0.468	0.560	0.713	0.138	0.232	0.739	0.496	**0.593**
	Category †	Cat(W-Freq)†	0.606	0.011	0.021	0.948	0.860	**0.902**	0.061	0.007	0.013	0.674	0.502	0.575	0.758	0.156	0.258	0.689	0.509	0.586
		parent(Cat)(W-Freq)†	0.472	0.011	0.021	0.947	0.858	0.901	0.088	0.010	**0.017**	0.690	0.512	0.588	0.763	0.145	0.244	0.690	0.512	0.588
		P(W-Freq/Cat)†	0.458	0.011	0.022	0.947	0.855	0.899	0.069	0.007	0.013	0.690	0.499	0.579	0.785	0.157	**0.261**	0.688	0.505	0.582
		Cat+P(W-SG)†	0.461	0.011	0.022	0.948	0.856	0.900	0.057	0.005	0.009	0.684	0.486	0.568	NA	NA	NA	NA	NA	NA
		parent(Cat)+P(W-SG)†	0.457	0.012	0.022	0.946	0.856	0.899	0.086	0.009	0.016	0.702	0.507	0.589	NA	NA	NA	NA	NA	NA
		P(W-SG/Cat)†	0.606	0.012	**0.024**	0.947	0.859	0.901	0.090	0.007	0.014	0.695	0.515	**0.592**	0.740	0.150	0.250	0.699	0.517	**0.594**
DisAcc	Baseline	BOW	0.216	0.002	0.004	0.955	0.869	**0.910**	0.584	0.059	**0.107**	0.782	0.608	0.684	0.835	0.090	**0.162**	0.819	0.605	**0.696**
		POS	0.322	0.009	0.017	0.951	0.860	0.903	0.273	0.029	0.052	0.746	0.630	0.688	0.719	0.090	0.159	0.744	0.625	0.679
		BOE	0.875	0.04	**0.076**	0.810	0.629	0.708	0.494	0.043	0.079	0.806	0.653	**0.722**	0.909	0.048	0.092	0.744	0.648	0.692
		BOC †	0.059	0.014	0.023	0.743	0.506	0.602	0.366	0.098	0.155	0.785	0.564	0.656	0.626	0.103	0.176	0.740	0.544	0.627
	Resource	Cls(W-Freq/Cls)	0.293	0.002	0.004	0.951	0.881	0.915	0.553	0.070	**0.125**	0.783	0.599	0.679	0.835	0.090	0.162	0.805	0.605	0.691
		parent(Cls)(W-Freq)	0.267	0.002	0.004	0.953	0.883	**0.917**	0.568	0.060	0.109	0.789	0.611	**0.689**	0.835	0.090	**0.162**	0.814	0.601	0.692
		P(W-Freq/Cls)	0.238	0.002	0.004	0.953	0.871	0.910	0.519	0.070	0.123	0.777	0.600	0.677	0.835	0.090	**0.162**	0.805	0.591	0.681
		Cls+P(W-SG)	0.237	0.002	0.004	0.953	0.866	0.907	0.570	0.067	0.120	0.786	0.606	0.684	0.835	0.090	**0.162**	0.812	0.598	0.689
		parent(Cls)+P(W-SG)	0.268	0.002	**0.005**	0.953	0.866	0.908	0.578	0.062	0.112	0.785	0.615	**0.689**	0.835	0.090	**0.162**	0.816	0.602	0.693
		P(W-SG/Cls)	0.248	0.002	0.004	0.954	0.873	0.912	0.643	0.059	0.109	0.800	0.603	0.688	0.835	0.090	**0.162**	0.815	0.607	**0.695**
	Category †	Cat(W-Freq)†	0.233	0.002	0.003	0.957	0.869	0.911	0.521	0.057	**0.102**	0.801	0.625	**0.702**	0.787	0.093	0.166	0.815	0.621	**0.705**
		parent(Cat)(W-Freq)†	0.271	0.002	0.004	0.957	0.869	0.911	0.546	0.048	0.088	0.792	0.617	0.694	0.775	0.091	0.162	0.802	0.602	0.688
		P(W-Freq/Cat)†	0.266	0.002	0.004	0.956	0.863	0.907	0.562	0.052	0.096	0.770	0.611	0.681	0.775	0.091	0.162	0.801	0.612	0.694
		Cat+P(W-SG)†	0.261	0.002	0.004	0.958	0.862	0.908	0.555	0.052	0.095	0.761	0.629	0.689	NA	NA	NA	NA	NA	NA
		parent(Cat)+P(W-SG)†	0.327	0.003	0.005	0.959	0.865	0.909	0.552	0.054	0.098	0.782	0.613	0.687	NA	NA	NA	NA	NA	NA
		P(W-SG/Cat)†	0.312	0.002	**0.005**	0.956	0.874	**0.913**	0.661	0.047	0.088	0.787	0.609	0.686	0.775	0.091	0.162	0.790	0.601	0.683

Table 7: The performance of the DB, FB and DB-FB cross-source SVM topic classifiers using different KSs ontologies (DB -using *dbKS*'s ontologies, FB - using *fbKS*'s ontology) and two semantic meta-graphs derived from these KSs (*resource meta-graph* (Resource) and *category meta-graph* (Category)). The results obtained for the semantic features derived for the *resource meta-graph* (reported in [13]) using the W-Freq weighting schema correspond to: class (Cls(W-Freq)), upper-class (parent(Cls)(W-Freq)) and property (P(W-Freq/Cls)); while using the W-SG weighting schema are: class-property co-occurrence (Cls+P(W-SG)), upper-class-property co-occurrence (parent(Cls)+P(W-SG)) and property (P(W-SG/Cls)). The results obtained for the semantic features derived for the *category meta-graph* using the W-Freq weighting schema are: category (Cat(W-Freq)), upper-category (parent(Cat)(W-Freq)) and property (P(W-Freq/Cat)); while using the W-SG weighting schema are: category-property co-occurrence (Cat+P(W-SG)), upper-category-property co-occurrence (parent(Cat)+P(W-SG)) and property (P(W-SG/Cat)).
The baseline models (Baseline) employed are bag-of-words (BOW), bag-of-entities (BOE), part-of-speech (POS) and bag-of-concepts (BOC).
The results marked with † correspond to the new results obtained for the newly introduced *category meta-graph*.

the *dbCat* ontology. These classifiers are evaluated against several baseline models, as presented in Table 7.

Looking at the performance of the baseline models, we observe a different trend compared to the TW only scenario. The syntactic classes provided by the *POS* taggers, in this cross-source scenario, were found to be more beneficial, compared to the *BoW* cases. While for the *BoE* and *BoC* features, we did not obtain an improvement on the baseline *BoW* features. An explanation for this could be that the entities which appear in the TW dataset could be quite different from the entities appearing in the KS data for each topic, in which case exploiting the semantic information from KSs seems to be more beneficial.

Inspecting the best overall performance for the various features, feature weighting strategies and augmentation strategies, we notice that the *resource meta-graph* achieved the best results using the $DB(dbOwl + yago) + FB(fbOnt) + TW$ topic classifier. As reported in our previous work ([13]), this classifier significantly outperformed the baseline single KS classifiers: by 11.9-30.7% (over $DB + TW$) and 13.4-31.4% (over $FB + TW$) (t-test with $\alpha < 0.05$). Considering the *category meta-graph*, the improvements were slightly smaller, a significant improvement of 11.5-30.2% was observed over $DB+TW$ and 13-30.9% over $FB + TW$ (t-test with $\alpha < 0.05$). Comparing the results against the TW baseline models, we observe a significant improvement of 9.3%-28.2% over the $TW(dbOwl+yago+fbOnt)$ when using the *resource meta-graph*, and 8.9%-27.7% over the $TW(dbCat + fbOnt)$ classifiers when using the *category meta-graph*.

Comparing the different enrichment strategies, we observed similar trends for both *resource meta-graph* and *category meta-graph*. The best enrichment that consistently improved over the baseline for both concept graphs was the *W-SG* for *P*, indicating that encoding the specificity of a property for each semantic concept graph is beneficial for TC. For the *W-Freq* features, however, we found that in the case of the *resource meta-graph*, the semantic augmentation by feature frequency (*Cls(W-Freq)*) and by generalisation (*parent(Cls)(W-Freq)*) (Table 7, column 8) worked consistently better than the baseline models. However, in the case of the *category meta-graph*, the performance of the *Cat(W-Freq)* and *parent(Cat)(W-Freq)* were only comparable to those of the baseline models.

Despite of the accuracy gain obtained with the *P* and *Cls* features for the $DB+FB+TW$ classifier, an interesting observation about these results is however, that the semantic features do not always improve upon the baseline models. For instance in the case of $DB + FB$ topic classifier, the results are comparable or slightly worst than those obtained by the *BoW* feature set ignoring semantic augmentation. An explanation for this could be that the distribution of entities in the DB and FB datasets may slightly be different to the one in Twitter. Further given that these classifiers do not make use of any Microposts data, this mismatch provides challenges for the topic classifier. A possible reason for this could be the level of ambiguity of the entities in the different datasets. In order to capture the differences between the datasets and provide an estimation on the usefulness of the different semantic features, the reminder of the reader, we employed a set of topic similarity measures which we will examine in Subsection 8.3.

Contrasting the results for all three topics, we observe, that the biggest overall improvement was achieved for the *Cri* topic using the *resource meta-graph*. In particular, the $DB+FB+TW$ achieved an improvement of 31.4% over $FB+TW$. For the case of the *category meta-graph*, the $DB + FB + TW$ achieved an improvement of 30.9% over $FB + TW$.

Also for the *Cri* topic, we observe, that the $FB + TW$ single KS classifier using *BoW* features performed better than the $DB + TW$ single KS classifier. However, when looking at the results obtained for the *BoE* features, we observe the opposite trend, the $DB+TW$ performed better than the $FB+TW$. An explanation for this could be that a relatively large number (3,377) of articles do not contain any entity, and thus are not semantically enriched.

Further, we noticed that the coverage of entities is lower in the Freebase than in DBpedia. For example from the total number of entities extracted by OpenCalais a large proportion (40%) of the entities were not found in the Freebase KS, while in the case of DBpedia 35% of the entities were not assigned any URI. Regardless of this, an improvement in F1 measure was obtained for both semantic graphs when combining the two linked KSs. This thus indicates that the two linked KSs complement each other well. In one hand, Freebase brings its strength in content coverage for the topics, while DBpedia brings useful semantic evidence about the entities which are covered ([13]).

In conclusion, considering the results obtained for both single-source and cross-source scenarios for the various semantic features derived from the three KS graphs, our findings are as follows:

1. Semantic meta-graphs (both *resource meta-graph* and *category meta-graph*) built from KSs contain useful semantic features about entities for TC. In particular, incorporating semantic features about properties (*P*) using our novel class-property co-occurrence weighting schema (*W-SG*) proved a significant improvement over previous state-of-the-art approaches.

2. Combining the evidence about the semantic features from multiple, linked KS taxonomies ($TW(dbKS + fbKS)$) is beneficial for TC, showing a significant improvement over approaches considering a single KS ($TW(dbKS)$, $TW(fbKS)$).

8.2. The Role of Semantic Concept Graphs in Single-Source and Cross-Source Topic Classification

In the previous section we compared the overall performance of a topic classifier using semantic features derived from two semantic meta-graphs (*resource meta-graph* and *category meta-graph*). In this section, we continue our discussion focusing on the differences in roles of these semantic features in different TC scenarios.

Looking at the results obtained for the individual semantic features (*Cat*, *Cls*, *P*) we observe different patterns for the single-source and cross-source TC scenarios.

Inspecting the results obtained for the single-source topic classifier, we notice that the performance of the SVM topic classifier was consistently higher using the *Cat* features than using the *Cls* features for both *W-Freq* and *W-SG* weightings (see Table 7) (t-test with $\alpha < 0.05$). These results indicate that the information about the category features seems to be more beneficial than the information about the classes in the single-source TC scenario. However, for the *P* features, we found that the weights obtained from the *resource meta-graph* are better than those obtained from the *category meta-graph*. This behaviour could be understood by the fact that the *category meta-graph* consists of a larger number of *Cat* than the number of *Cls* in the *resource meta-graph*, and in addition the *Cat* are more ambiguous (less focused) than the *Cls* in terms of the number of properties associated to them.

In contrast to these observations, in the cross-source TC scenarios we notice different trends for the *Cat* and *C* features (t-test with $\alpha < 0.05$). While for the TW only scenario, the *Cat* features worked better than the *Cls* features, in the cross-source scenario we observe the opposite trend, the *Cls* features are more useful than *Cat* features. An explanation for this could be that the different datasets contain a larger number of *Cat* features than the *Cls* features (compare *dbCat* with *dbClass*, *yago-Class* and *fbClass* in Table 5), making it harder for the cross-source classifiers to generalise over the *Cat* features than the *Cls* features.

In conclusion, considering the results obtained for both single-source and cross-source scenarios, our findings are as follows:

1. The semantic features derived from the *resource meta-graph* and *category meta-graph* exhibit different roles (generalisation patterns) in the different TC scenarios. The class features derived from the *resource meta-graph* exhibit better *generalisation patterns* in the cross-source setting, while the category features derived from the *category meta-graph* are better suited to encode the *specificity* of a topic in a single-source setting

2. Despite the differences in roles of the semantic features derived from the two semantic meta-graphs, incorporating semantic features from both semantic graphs is beneficial for TC, achieving performance superior to previous approaches utilising lexical features.

8.3. Evaluating Topic Similarity Measures

The previous sections analysed the benefit of using semantic features derived from KS graphs for the topic classification task in both single-source and cross-source scenarios. These results have also shown that there is variation in the performance levels between topics. This suggests that differences between the KS and Twitter datasets affects the performance levels. In order to understand these variations, we analysed the relevance of these semantic features for the representation of a given topic. For this reason, we computed the *entropy difference* values between the training and test datasets for each topic as introduced in Section 5.

In order to assess the relevance of a semantic feature type to the performance of a topic classifier, we analysed these metrics by considering the following cases:

1. Measuring entity dispersion (*Entity Entropy*) - Since this metric captures only the entity dispersion in topics, we correlated it against topic classifiers build on *BoE* features;

2. Measuring class dispersion (*Class Entropy*, *Entity-Class Entropy*) - In this case we took the topic classifiers trained using *Cls* features;

3. Measuring category dispersion (*Category Entropy*, *Entity-Category Entropy*) - In this case we considered the topic classifiers built using the *Cat* features; and

4. Measuring property dispersion (*Property Entropy*, *Entity-Property Entropy*, *Class-Property Entropy*, and *Category-Property Entropy*) - we considered the topic classifiers using *P* features.

Figure 6 presents the Pearson correlation values obtained for each topic. The correlation was calculated between the entity difference scores and the performance of the cross-source ($DB + FB + TW$) and single-source (TW(*dbKS+fbKS*)) classifiers in terms of F1 measure obtained using 80% of TW data for training (in addition to the KS data), and 20% TW data for test.

A positive correlation indicates that the performance increases as the entropy difference decreases (the distributions are more similar); while a negative correlation indicates that the performance increases as the divergence increases (the distributions are less similar).

These figures show that in the cross-source ($DB + FB + TW$) scenario, the *Entity-Property Entropy* yields the best correlation scores, over 70% in two out of three topics. When looking at the values obtained for *Class Entropy*, *Category Entropy*, *Property Entropy* and *Entity Entropy* measures, we observe, that the *Class Entropy* showed the highest correlation values with the performance of the cross-source topic classifiers. For the *DisAcc* and *War* these values were higher than 54%, however, for the *Cri* topic the correlation values were 11%. When examining the class dispersion measures, we see that the *Entity-Class Entropy* showed higher correlation than *Class Entropy*. In the case of the category dispersion values, for some topics (e.g. *DisAcc*) the *Entity-Category Entropy* was found to be better, while for others (e.g. *War*) the *Category Entropy* was more beneficial. Moreover, among the property dispersion values the *Entity-Property Entropy* values showed the highest correlation values.

Considering the results obtained for the single-source TC (TW(*dbKS+fbKS*)) case, the *Class-Property Entropy* yields the best correlation value, over 60% for all three topics. Among the *Class Entropy*, *Category Entropy*, *Property Entropy* and *Entity Entropy* measures, however, the *Property Entropy* values were found to be the best. As opposed to the cross-source case, among the class dispersion measures, the *Category Entropy* values were higher than the *Entity-Class Entropy* values. For the category dispersion measures, the *Category Entropy* values

were higher than the *Entity-Category Entropy* values in two out of three topics. These results indicate that in the single-source case analysing a single semantic feature (e.g. *P*, *Cls* or *Cat*) can provide a good estimate of the performance of the topic classifier. In the case of the cross-source scenario the representation of the topics seems to be more complex, requiring the modelling of the entropy of two semantic features (in our case in the form of conditional entropy values). Nonetheless, among the property dispersion values, the best results were obtained by the *Class-Property Entropy* values.

We also compared these results with the content-based similarity measures studied in our previous work ([14]). That is, we computed the $(\chi^2)^{-1}$ measure between the training and test datasets for the $DB+FB+TW$ and TW(*dbKS*+*fbKS*) classifiers using *BoW* and *BoE* features, and correlated these values with the performance of these classifiers. According to these results, in the single-source case the best correlation values obtained were: 21% (*BoE*) for *DisAcc*, 58% (*BoW*) for *Cri*, and 23% (*BoE*) for *War*; while in the cross-source case, these values were 14% (*BoW*) for *DisAcc*, 45% (*BoE*) for *Cri*, and 20% (*BoE*) for *War*. As we observe, our novel entropy based similarity measures (*Entity-Property Entropy* for cross-source TC and *Class-Property Entropy* for single-source TC) achieve better correlation with the performance of the topic classifier, showing the usefulness of incorporating semantic features from KSs for enhancing the representation of a topic.

Given the above observations, our general findings about the entropy-based measures are as follows:

1. The performance of a topic classifier can be accurately assessed following the proposed entropy-based measures. These measures when applied over a particular Topic's concept graphs generated from multiple linked KSs, outperform previous content based similarity measures derived from the sole text content.

2. The usefulness of these entropy based measures varies among different topics and TC scenarios. However, the property-based dispersion measures achieved best correlation values in both single-source and cross-source TC scenarios[30].

9. Discussion and Future Directions

Our three-stage approach for topic classification analysis of microposts functions by i) *context modelling*; ii) *topic classification* and iii) *topic similarity analysis*.

We now discuss the issues and findings from each stage.

9.1. Context modelling

The presented semantic meta-graphs (both *resource meta-graph* and *category meta-graph*) are capable of providing contextual information about concepts in short text. Our method for TC makes use of various semantic features that are constructed from these semantic meta-graphs. By extracting the named entities we were able to enhance the lexical feature space of a topic classifier with additional contextual information about these concepts. In addition, our approach takes into account the information about concepts (e.g. resource type-hierarchies, resource properties) present in multiple semantic concept graphs of multiple linked KSs.

The current framework employed two large coverage LOD KSs for demonstrating the usefulness of structured data in the TC task. However, LOD contains many other KSs interlinked with DBpedia, such as Geonames[31] or MusicBrainz[32]. A new LOD KS can easily be integrated into the current framework, by exploiting the data (if available) for training a topic classifier, and the semantic information present in the KS's ontology as additional semantic features.

For other KSs, which are not part of the LOD cloud (e.g. Wikidata[33]), the proposed framework could still be applied provided that a mapping between the DBpedia KS and the newly explored KS exists. A possible future direction could be to utilise the data from DBpedia, and derive contextual information about entities from the semantic meta-graph of the new KS.

One of the main factors which influence the performance of our approach, is the performance of the named entity recogniser (NER) used to extract the named entities from short text messages. In this paper we employed one of the most popular entity recognisers (OpenCalais and Zemanta) for this purpose. Although there have been several NER available ([48]) for extracting entities from textual data, these approaches were built on newswire corpora, and therefore to date it is not well understood which provides the best performance on Microposts. Our future work will thus concentrate in evaluating our framework using other NERs ([51]).

A second factor which has some drawback to the performance of our approach is the *incompleteness* and the *inconsistencies* within the KSs. For e.g. in Freebase the *fbOnt*:/crime/crime_accuser class is derived from a very generic *fbOnt*:/common/topic class, while another related class type *fbOnt*:/crime/convicted_criminal extends the *fbOnt*:/people/person class. In the case of the category structure of Wikipedia, we also note that the category tree is not a strict taxonomy and does not always contain an *is-a* relationship ([8]). Given that for both semantic concept graphs we applied concept generalisation strategies, this mismatch can affect the generalisation of the patterns learned by our topic classifier, in that entities which should be considered together might belong to different entity types. One possible solution to overcome this problem could be to perform a cross-consistency validation, by

[30] We also mention here that the inconsistencies for the entropy values (achieving both positive and negative correlations for a given entropy measure) may be the result of many different factors, such as the noisy lexical nature of Microposts or the distributional differences between the KS and TW datasets in term of entities. In order to understand these variations, a more in-depth analysis would need to be conducted, which we aim to investigate in the future.

[31] http://www.geonames.org

[32] http://musicbrainz.org

[33] http://www.wikidata.org

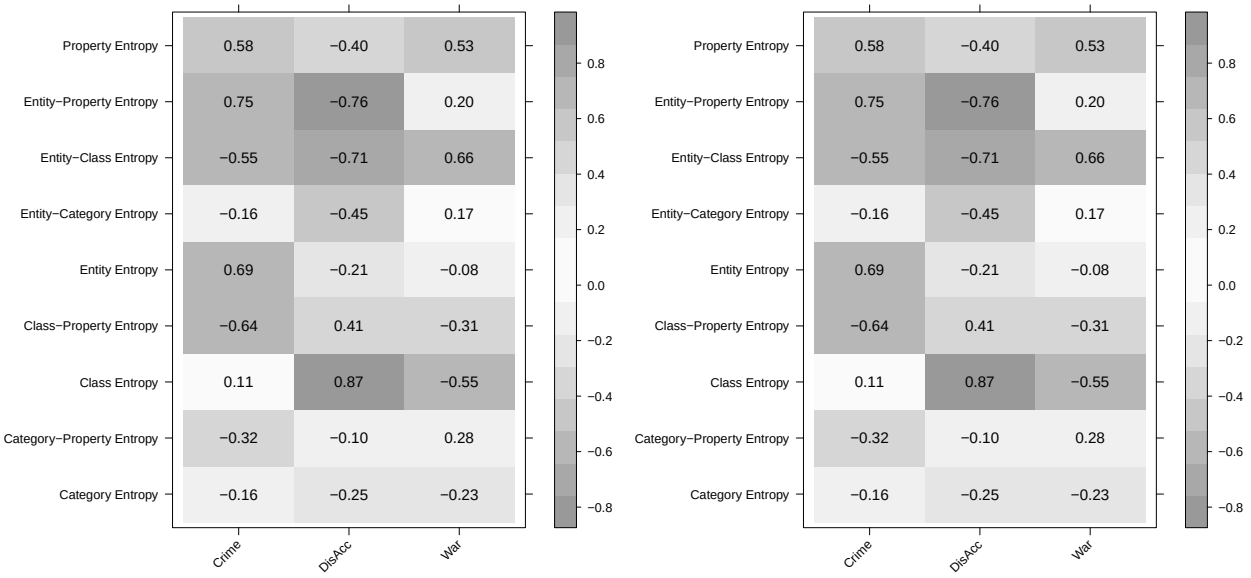

Figure 6: Pearson correlation values between the entropy difference measures and the performance of the DB+FB cross-source (left), and TW (*dbKS+fbKS*) single-source (right) topic classifiers.

investigating the overlapping properties between the entities assigned to the same entity classes, and consider the most likely entity classes ([52]).

9.2. Topic classification

The described method for topic classification uses supervised SVM machine learning models to detect the topic of Microposts. These models make use of the lexical (*BoW*) and semantic features extracted from different semantic meta-graphs.

Given the vocabulary differences between KSs and Tweets, one of the challenges faced by these models is the frequent usage of grammatically incorrect English in Microposts. Due to the restricted size of short messages, entities such as country names (e.g. *nkorea*) are often abbreviated, as in the following Tweet: *"nkorea prepared nuclear weapons holy war south official tells state media usa"*. These irregularities mean that current annotation services (including OpenCalais API) will ignore these entities, and therefore no semantic information will be exploited for these entities by the TC system. A possible solution to address these challenges is to apply lexical normalisers especially developed for Tweets ([53]) to normalise these words to standard English terms.

In addition, these methods model the content of text using simple 1-gram (unigram) features. A possible extension of our approach could be to incorporate other ngram features into these models also, for e.g. 2-grams or a combination of 1-grams with 2-grams ([54]).

Moreover, when building the cross-source topic classifiers, our models still require a large number of annotated Tweets to outperform the single-source Twitter models. Previous research on cross-domain (cross-source) learning has also shown that outperforming the target (Twitter) classifier is extremely difficult for many text classification tasks ([55, 15]). In order to further increase the effectiveness and robustness of the current model, our future work in this direction will thus focus on investigating unsupervised multi-source adaptation models which require less annotation from Twitter ([56]).

9.3. Topic Similarity Analysis

The examined entropy-based measures make use of the enhanced representation of topics exploiting contextual information from semantic concept graphs about concepts from linked KSs. This new representation led us to induce a new semantic feature space for a topic consisting of semantic features extracted from the semantic meta-graphs of multiple linked KSs. Our results on both single-source and cross-source scenarios show that this new semantic representation can be useful for providing a good estimate on the performance of a TC, achieving correlation values over 60% on the single-source scenario and over 70% on the cross-source scenario.

In contrast to our previous work using content based similarity measures ([14]) for topic similarity, we also showed an improvement in correlation values. These results provided further evidence of the benefit of exploiting the information from KSs for the representation of a topic. Considering, however, that our entropy-based measures also depend on the performance of a NER at hand, a promising future direction of this research could be to propose measures that combine the contribution of both content-based lexical similarity and entropy-based similarity measures (e.g. as in [57]).

10. Conclusion

The real-time classification of Tweets is important since they act as social sensors revealing emerging events occurring in the world. In this work we investigated the use of semantic concept graphs of linked knowledge sources for topic classification of social media posts. We demonstrated the feasibility of this approach by implementing classification models that make use of semantic graph structures in multiple knowledge sources (DBpedia and Freebase). In particular, we introduced and evaluated various semantic features derived from two distinct concept graphs (*resource-meta graph* and *category-meta graph*) of these KSs, and showed that they can help to build accurate topic classifiers of Tweets.

By exploring the research question *How does the performance of a topic classifier vary using different concept graphs?*, we found that although both semantic concept graphs contain useful information for TC of Tweets, the best overall performance was achieved by the features derived from the resource meta-graph. More importantly, for both concept graphs, we obtained a significant improvement over previous approaches using only lexical features derived from the single Twitter dataset content.

Through addressing the question *Are there differences in the roles (generalisation patterns) of the concept graphs in the different topics and TC scenarios?*, we compared the usefulness of the semantic features for two different scenarios: the *cross-source* scenario utilising KS data and the *single-source* scenario utilising only Twitter data. Our results in this respect revealed different roles (generalisation patterns) for the features derived from the two concept graphs. In particular, we found that some features from the category meta-graph were better used to encode the specificity of a topic, achieving the best performance in the single-source case, however, when considering the cross-source scenario other features derived from the resource meta-graph were found to be better. Nonetheless, despite the different roles of the features, our topic classifier exploiting multiple linked KSs achieved significant results over the baseline models.

These insights have provoked our final question *Can we predict the performance of a topic classifier?* To address this question, we introduced and evaluated various entropy-based measures defined over these semantic concept graphs and showed that the performance of a topic classifier can be predicted with reasonably high accuracy using the property dispersion entropy measures. Further, we showed a significant improvement over previous content-based lexical similarity measures proposed for TC.

Overall, our approach demonstrated that semantic meta-graphs derived from linked KSs: i) provide useful semantic features helpful in accurately detecting topics in Microposts ii) can be used as a measure for predicting the accuracy of a topic classifier.

11. Acknowledgement

The authors would like to express their gratitude to Aba-Sah Dadzie for proofreading this paper.

References

[1] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010.

[2] J. Bollen, H. Mao, A. Pepe, Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, in: Proceedings of the Fifth International Conference on Weblogs and Social Media, AAAI Press, 2011.

[3] V. Lampos, T. De Bie, N. Cristianini, Flu detector: tracking epidemics on twitter, in: Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III, Springer-Verlag, 2010.

[4] S. Asur, B. A. Huberman, Predicting the future with social media, in: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, IEEE Computer Society, 2010.

[5] B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith, From tweets to polls: Linking text sentiment to public opinion time series, in: Proceedings of the Fourth International Conference on Weblogs and Social Media, AAAI Press, 2010.

[6] X. Zhang, H. Fuehres, P. A. Gloor, Predicting stock market indicators through twitter 'i hope it is not as bad as i fear', Procedia - Social and Behavioral Sciences (2011).

[7] D. Milne, I. H. Witten. (Eds.), Learning to link with Wikipedia., 2008.

[8] E. Gabrilovich, S. Markovitch, Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge, in: Proceedings of Twenty-First AAAI Conference on Artificial Intelligence, AAAI Press, 2006.

[9] Y. Genc, Y. Sakamoto, J. 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, Springer-Verlag, 2011.

[10] O. Muñoz García, A. García-Silva, O. Corcho, M. de la Higuera Hernández, C. Navarro, Identifying Topics in Social Media Posts using DBpedia, in: Proceedings of the NEM Summit, Eurescom, the European Institute for Research and Strategic Studies in Telecommunications, 2011.

[11] S. P. Kasiviswanathan, P. Melville, A. Banerjee, V. Sindhwani, Emerging topic detection using dictionary learning, in: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011.

[12] E. Meij, W. Weerkamp, M. de Rijke, Adding semantics to microblog posts, in: Proceedings of the fifth ACM international conference on Web search and data mining, ACM, 2012.

[13] A. E. Cano, A. Varga, M. Rowe, F. Ciravegna, Y. He, Harnessing linked knowledge sources for topic classification in social media, in: Proceedings of the 24th ACM Conference on Hypertext and Social Media, ACM, 2013.

[14] A. Varga, A. E. Cano, F. 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, CEUR, 2012.

[15] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering (2010).

[16] R. C. Bunescu, M. Pasca, Using encyclopedic knowledge for named entity disambiguation, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, ACL, 2006.

[17] M. Strube, S. P. Ponzetto, Wikirelate! computing semantic relatedness using wikipedia, in: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2, AAAI Press, 2006.

[18] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, Dbpedia - a crystallization point for the web of data, Journal of Web Semantics (2009).

[19] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, 2008.
[20] J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum, Yago2: a spatially and temporally enhanced knowledge base from wikipedia, Artificial Intelligence Journal, Special Issue on Artificial Intelligence, Wikipedia and Semi-Structured Resources (2012).
[21] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media?, in: Proceedings of the 19th international conference on World wide web, ACM, 2010.
[22] D. Laniado, P. Mika., Making sense of twitter, in: Proceedings of the 9th international semantic web conference on The semantic web, Springer-Verlag,, 2010.
[23] S. C. W. Weerkamp, M. Tsagkias., How people usewitter in different languages., in: Proceedings of the Web Science 2011, ACM, 2011.
[24] K. M. T. J. Huang, E. N. Efthimiadis, Conversational tagging in twitter, in: Proceedings of the 21th ACM Conference on Hypertext and Social Media, ACM, 2010.
[25] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation. (2003).
[26] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: Proceedings of the 33rd European conference on Advances in information retrieval, Springer-Verlag, 2011.
[27] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, Improving lda topic models for microblogs via tweet pooling and automatic labeling, in: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM, 2013.
[28] D. Ramage, D. Hall, R. Nallapati, C. D. Manning, Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, ACL, 2009.
[29] D. Ramage, S. Dumais, D. Liebling, Characterizing Microblogs with Topic Models, in: Proceedings of the International AAAI Conference on Weblogs and Social Media, AAAI Press, 2010.
[30] J. Lin, R. Snow, W. Morgan, Smoothing techniques for adaptive online language models: topic tracking in tweet streams, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011.
[31] K. Tao, F. Abel, C. Hauff, G.-J. Houben, What makes a tweet relevant for a topic?, in: Making Sense of Microposts (#MSM2012), CEUR, 2012.
[32] V. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, Soviet Physics Doklady (1966).
[33] S. T. Dumais, Latent semantic analysis, Annual Review of Information Science and Technology (2004).
[34] Y. Song, H. Wang, Z. Wang, H. Li, W. Chen, Short text conceptualization using a probabilistic knowledgebase, in: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, AAAI Press, 2011.
[35] W. Wu, H. Li, H. Wang, K. Q. Zhu, Probase: A probabilistic taxonomy for text understanding, in: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, ACM, 2012.
[36] Y. Shin, C. Ryo, J. Park, Automatic extraction of persistent topics from social text streams, World Wide Web (2013).
[37] A. Garcia-Silva, O. Corcho, J. Gracia, Associating semantics to multilingual tags in folksonomies, in: Proceedings of the 17th International Conference on Knowledge Engineering and Knowledge Management, Springer, 2010.
[38] D. Vitale, P. Ferragina, U. Scaiella, Classification of short texts by deploying topical annotations, in: Proceedings of the 34th European Conference on IR Research, Springer, 2012.
[39] P. K. P. N. Mendes, A. Passant, A. P. Sheth., Linked open social signals., in: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society, 2010.
[40] M. Michelson, S. 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, CEUR, 2010.
[41] T. Xu, D. W. Oard, Wikipedia-based topic clustering for microblogs, Proceedings of the American Society for Information Science and Technology (2011).
[42] S. Husby, D. Barbosa, Topic classification of blog posts using distant supervision, in: Proceedings of the Workshop on Semantic Analysis in Social Media, ACL, 2012.
[43] G. Forman, I. Guyon, A. Elisseeff, An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research (2003).
[44] I. Oelze, Integration of Yago ontology in the IQP query construction system to support efficient query construction over a large-scale relational database, Technical Report, 2011.
[45] N. Cristianini, J. Shawe-Taylor, An introduction to support Vector Machines: and other kernel-based learning methods, Cambridge University Press, 2000.
[46] F. Abel, Q. Gao, G.-J. Houben, K. Tao, Analyzing user modeling on twitter for personalized news recommendations, in: Proceedings of the 19th international conference on User modeling, adaption, and personalization, Springer-Verlag, 2011.
[47] J. B. Lovins, Development of a stemming algorithm, Mechanical Translation and Computational Linguistics (1968).
[48] G. Rizzo, R. Troncy, Nerd : a framework for evaluating named entity recognition tools in the web of data, in: Proceedings of the 10th International Semantic Web Conference, Springer, 2011.
[49] A. Ritter, S. Clark, Mausam, O. Etzioni, Named entity recognition in tweets: An experimental study, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, ACL, 2011.
[50] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 2007.
[51] A. E. C. Basave, A. Varga, M. Rowe, M. Stankovic, A.-S. Dadzie, Making sense of microposts (#msm2013) concept extraction challenge, in: Making Sense of Microposts (#MSM2013) Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts', 22nd International World Wide Web Conference (WWW'13), CEUR, 2013.
[52] J. Dolby, A. Fokoue, A. Kalyanpur, E. Schonberg, K. Srinivas, Extracting Enterprise Vocabularies Using Linked Open Data, in: Proceedings of the 8th International Semantic Web Conference (ISWC2009), Springer, 2009.
[53] B. Han, T. 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, ACL, 2011.
[54] V. Lampos, Detecting events and patterns in large-scale user generated textual streams with statistical learning methods, CoRR, 2012.
[55] J. Blitzer, M. Dredze, F. Pereira, Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, in: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, ACL, 2007.
[56] W. Dai, G.-R. Xue, Q. Yang, Y. Yu, Transferring naive bayes classifiers for text classification, in: Proceedings of the 22nd National Conference on Artificial Intelligence, AAAI Press, 2007.
[57] N. Ponomareva, M. Thelwall, Biographies or blenders: Which resource is best for cross-domain sentiment analysis?, in: Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing, Springer-Verlag, 2012.