Improving Categorisation in Social Media Using Hyperlinks to Structured Data Sources*

Sheila Kinsella[1], Mengjiao Wang[1], John G. Breslin[1,2], and Conor Hayes[1]

[1] Digital Enterprise Research Institute, National University of Ireland, Galway
firstname.lastname@deri.org
[2] School of Engineering and Informatics, National University of Ireland, Galway
john.breslin@nuigalway.ie

Abstract. Social media presents unique challenges for topic classification, including the brevity of posts, the informal nature of conversations, and the frequent reliance on external hyperlinks to give context to a conversation. In this paper we investigate the usefulness of these external hyperlinks for categorising the topic of individual posts. We focus our analysis on objects that have related metadata available on the Web, either via APIs or as Linked Data. Our experiments show that the inclusion of metadata from hyperlinked objects in addition to the original post content significantly improved classifier performance on two disparate datasets. We found that including selected metadata from APIs and Linked Data gave better results than including text from HTML pages. We investigate how this improvement varies across different topics. We also make use of the structure of the data to compare the usefulness of different types of external metadata for topic classification in a social media dataset.

Keywords: social media, hyperlinks, text classification, Linked Data, metadata.

1 Introduction

Social media such as blogs, discussion forums, micro-blogging services and social-networking sites have grown significantly in popularity in recent years. By lowering the barriers to online communication, social media enables users to easily access and share content, news, opinions and information in general. Recent research has investigated how microblogging services such as Twitter enable real-time, first-hand reporting of news events [15] and how question-answering sites such as Yahoo! Answers allow users to ask questions on any topic and receive community-evaluated answers [1]. Social media sites like these are generating huge amounts of user-generated content and are becoming a valuable source of information for the average Web user.

* The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

G. Antoniou et al. (Eds.): ESWC 2011, Part II, LNCS 6644, pp. 390–404, 2011.

However, navigating this wealth of information can be challenging. Posts are unstructured with little or no metadata and are usually much shorter than a typical Web document. Due to the casual environment and minimal curation in social media sites, the quality is highly variable [1]. Much of the information contained in social media is generated during conversations between people who have a shared context that allows them to communicate without explicitly stating all relevant information. In many cases, vital pieces of information are provided not within the text of the post, but behind hyperlinks that users post to refer to a relevant resource. For example, a poster may recommend a book by posting a hyperlink to a webpage where you can buy it, rather than using the traditional method of providing the book title and name of the author. In the message board dataset described in Section 4, we found that 65% of posts that linked to books mentioned neither the complete title nor the complete author name, and at least 11% did not contain even a partial title or author name.

Such hyperlinks to external resources are a potential source of additional information for information retrieval in online conversations. Hyperlinks to sources of structured data are particularly promising because the most relevant metadata can be extracted and associated with the original post content. The resulting rich representations of posts can be used for enhanced search and classification. Recently, there has been a growing amount of structured information available on the Web. Many of the most popular websites such as Amazon and YouTube provide developer APIs that can be used to programmatically access metadata about resources, and there is also a growing amount of RDFa and Linked Data being published. The Linking Open Data [5] project in particular has resulted in many structured datasets from diverse domains becoming available on the Web, some of which we use as data sources in our experiments.

In this paper, we focus on the task of improving categorisation in social media using metadata from external hyperlinks, building on initial experiments reported in [14]. To ensure that our conclusions are valid across different websites, we investigate datasets from two disparate types of social media, a discussion forum and a micro-blogging website. We compare the results of topic classification based on post content, on text from hyperlinked HTML documents, on metadata from external hyperlinks, and on combinations of these. Our experiments show that incorporating metadata from hyperlinks can significantly improve the accuracy of categorisation of social media items. We make use of the structure of the data to empirically evaluate which metadata types are most useful for categorisation. We also investigate how the results vary by topic, in order to determine the circumstances where this approach would add the most benefit. Our results demonstrate that thanks to the linked nature of the Web, structured data can be useful to improve classification even in non-structured data.

2 Related Work

The enhancement of Web documents with external information for information retrieval is a long-established technique, an early example being Google's use of

anchor text for Web search [17]. Previous studies in the field of Web document categorisation have proven that the classification of webpages can be boosted by taking into account the text of neighbouring webpages ([2], [16]). Our work differs in that we focus on social media and rather than incorporating entire webpages or parts of webpages, we include specific metadata items that have a semantic relation to the objects discussed in a post. This approach enables us to compare the effectiveness of different metadata types for improving classification.

Other work has looked at classifying particular types of Web objects using metadata. Our work is related to that of Figueiredo et al. [8], who assess the quality of various textual features in Web 2.0 sites such as YouTube for classifying objects within that site. They do not use any external data. Yin et al. [21] propose improving object classification within a website by bridging heterogeneous objects so that category information can be propagated from one domain to another. They improve the classification of Amazon products by learning from the tags of HTML documents contained within ODP[1] categories.

There is previous work on classifying social media using the metadata of the post itself. Berendt and Hanser [4] investigated automatic domain classification of blog posts with different combinations of body, tags and title. Sun et al. [19] showed that blog topic classification can be improved by including tags and descriptions. Our work differs from these because we use metadata from objects on the Web to describe a social media post that links to those objects.

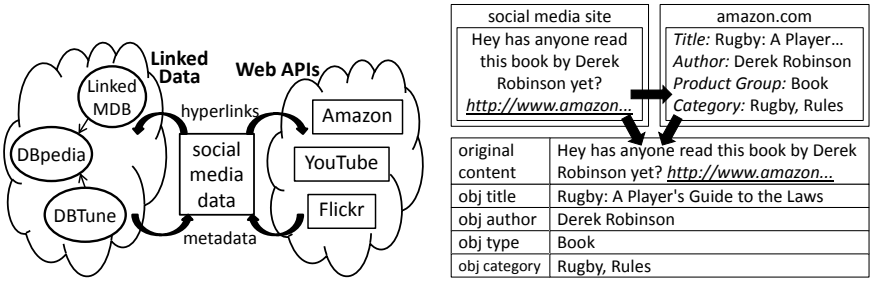
There has also been related work in the area of classifying Twitter messages. Garcia Esparza et al. [9] investigated tweet categorisation based on content. Jansen et al. [12] classified posts relating to consumer brands according to their sentiment. Irani et al. [11] studied the problem of identifying posters who aim to dishonestly gain visibility by misleadingly tagging posts with popular topics. They build models that correspond to topics in order to identify messages that are tagged with a topic but are in fact spam. They take advantage of hyperlinks by augmenting their models with text from webpages linked to within posts. Our work differs in that we focus on the potential offered by structured Web data and show that extracting relevant metadata gives superior results to using entire webpages. In the Semantic Web domain, Stankovic et al. [18] proposed a method for mapping conference-related posts to their corresponding talks and then to relevant DBpedia topics, enabling the posts to be more easily searched.

A relevant study that used structured data from hyperlinks in social media was performed by Cha et al. [7] who used hyperlinks in a blog dataset to study information propagation in the blogosphere. They downloaded the metadata of YouTube videos and analysed the popularity of categories, the age distribution of videos, and the diffusion patterns of different categories of videos.

3 Enhanced Post Representations

Hyperlinks are often an integral part of online conversations. Users share videos or photos they have seen, point to products or movies they are interested in,

[1] Open Directory Project, `http://www.dmoz.org/`, accessed March 2011.



(a) (b)

Fig. 1. (a) Web sources that were used to enrich social media data; **(b)** example of a how a social media post can be enhanced with external structured data relating to it

and use external articles as references in discussions. These external resources can provide useful new data such as author information in the case of books, or genre information in the case of movies. Many of these hyperlinks are to websites that publish metadata about objects, such as videos (YouTube) or products (Amazon), and make this metadata available via an API or as Linked Data. These websites are particularly useful since they allow particular pieces of relevant data to be identified and extracted, along with their relationship to the hyperlinked resource. In some cases the metadata is published by external sources, *e.g.*, DBpedia [3] provides a structured representation of Wikipedia. Figure 1(a) gives a graphical representation of how a social media dataset can be enhanced with information from various Web sources. The sources shown are those that will be used in experiments later in this paper.

Figure 1(b) gives an example of a post where additional useful information can be gained by considering the metadata of a hyperlinked object. Some of the information such as the author is redundant, but the title of the book and the categories are new. Both the title and the categories can be of use for classifying this post, which was posted in a Rugby forum. The name of the book could be useful for information retrieval, for example in a search scenario where a user queries for a book title. It is likely that certain metadata types will generally be more useful than others - for example, while the name of a book's publisher may sometimes indicate the topic of a book, in many cases it will not be helpful.

In order to integrate information from structured data sources in our post representations we make use of the vector space model so that we can re-use established methods that operate on vector models. A document d_i is represented as a vector of terms $d_i = \{w_{i,1}, w_{i,2}...w_{i,t}\}$ where $w_{i,j}$ denotes the weight of term j in document i. Functions such as tf-idf and document length normalisation can be applied to reduce the impact of variations in term frequency and document length. For each post, we create one feature vector based on the text from the original post, another based on hyperlinked HTML documents, and another based on hyperlinked object metadata. We experiment with different ways of combining the text sources into unified feature vectors.

4 Data Corpus

In our experiments, we use datasets originating from two different types of social media: *Forum*, from an online discussion forum, and *Twitter*, from the microblogging site[2]. We examined the domains linked to in each dataset and identified the most common sources of structured data. We extracted the posts that contained hyperlinks to these sources, and for each hyperlink we retrieved the related metadata as well as the corresponding HTML page. An identical pre-processing step was applied to each text source - post content, metadata and HTML documents. All text was lower-cased, non-alphabetic characters were omitted and stopwords were removed. Table 1 shows the average number of unique tokens remaining per post for each text source after preprocessing. The discrepancy in lengths of metadata and HTML between datasets is due to differences in the distribution of domains linked to in each dataset. Wikipedia articles, for example, tend to have particularly long HTML documents.

We now describe each stage of the data collection in more detail. The process of metadata collection for the forum dataset is described further in [13].

***Forum* dataset.** We use the corpus from the 2008 `boards.ie` SIOC Data Competition[3], which covers ten years of discussion forum posts represented in the SIOC (Semantically-Interlinked Online Communities) format [6]. Each post belongs to a thread, or conversation, and each thread belongs to a forum, which typically covers one particular area of interest. Our analysis considers only the posts contained in the final year of the dataset, since the more recent posts contain more links to structured data sources. From the most common domains in the dataset we identified MySpace, IMDB and Wikipedia as sources of Linked Data, via third-party data publishers detailed later in this section. We identified Amazon, YouTube and Flickr as sources of metadata via APIs. We use forum titles as categories for the classification experiments, since authors generally choose a forum to post in according to the topic of the post. We selected ten forums for these experiments based on the criteria that they were among the most popular forums in the dataset and they each have a clear topic (as opposed to general "chat" forums). The percentage of posts that have hyperlinks varies between forums, from 4% in Poker to 14% in Musicians, with an average of 8% across forums. These are a minority of posts; however, we believe they are worth focusing on because the presence of a hyperlink often indicates that the post is a

Table 1. Average unique tokens from each text source (± standard deviation)

Dataset	Content	Metadata	HTML
Forum	37.8 ± 42.6	19.6 ± 26.0	597.0 ± 659.1
Twitter	5.9 ± 2.6	26.9 ± 28.1	399.7 ± 302.9

[2] `http://twitter.com/`, accessed March 2011.
[3] `http://data.sioc-project.org/`, accessed March 2011.

useful source of information rather than just chat. Of the posts with hyperlinks, we focus on the 23% that link to one or more of the structured data sources listed previously. For the 23% of posts that have a title, this is included as part of the post content. Since discussion forums are typically already categorised, performing topic classification is not usually necessary. However, this data is representative of the short, informal discussion systems that are increasingly found on Web 2.0 sites, so the results obtained from utilising the class labels in this dataset should be applicable to similar uncategorised social media sites.

***Twitter* dataset.** The Twitter dataset[4] comes from Yang and Leskovec [20], and covers 476 million posts from June 2009 to December 2009. Twitter is a microblogging site that allows users to post 140 character status messages (tweets) to other users who subscribe to their updates. Due to the post length restriction, Twitter users make frequent use of URL shortening services such as bit.ly[5] to substantially shorten URLs in order to save space. Therefore for this dataset it was necessary to first decode short URLs via cURL[6]. From the most common domains we identified Amazon, YouTube and Flickr as sources of metadata via APIs. Like many social media websites, but in contrast to the previous dataset, Twitter does not provide a formal method for categorising tweets. However, a convention has evolved among users to tag updates with topics using words or phrases prefixed by a hash symbol (#). We make use of these hashtags to create six categories for classification experiments. Our approach borrows the hashtag-to-category mappings method from Esparza et al. [9] to identify tweets that relate to selected categories. We reuse and extend the hashtag categories of [9]; Table 2 shows the mappings between hashtags and categories. These categories were chosen because they occur with a high frequency in the dataset and they have a concrete topic. Tweets belonging to more than one category were omitted, since our goal is to assign items to a single category. All hashtags were removed from tweets, including those that do not feature in Table 2, since they may also contain category information. Any URLs to websites other than the selected metadata sources were eliminated from tweets. Finally, to avoid repeated posts caused by users retweeting (resending another post), all retweets were omitted.

External metadata. Amazon product, Flickr photo and YouTube video metadata was retrieved from the respective APIs. MySpace music artist information was obtained from DBTune[7] (an RDF wrapper of various musical sources including MySpace), IMDB movie information from LinkedMDB[8] (a movie dataset with links to IMDB) and Wikipedia article information from DBpedia[9]. The latter three services are part of the Linking Open Data project [5]. The number

[4] `http://snap.stanford.edu/data/twitter7.html`, accessed March 2011.
[5] `http://bit.ly`, accessed March 2011.
[6] `http://curl.haxx.se/`, accessed March 2011.
[7] `http://dbtune.org/`, accessed March 2011.
[8] `http://linkedmdb.org/`, accessed March 2011.
[9] `http://dbpedia.org/`, accessed March 2011.

Table 2. Categories and corresponding hashtags in the *Twitter* dataset

Category	#hashtags
Books	book, books, comic, comics, bookreview, reading, readingnow, literature
Games	game, pcgames, videogames, gaming, gamer, xbox, psp, wii
Movies	movie, movies, film, films, cinema
Photography	photography, photo
Politics	politics
Sports	nfl, sports, sport, football, f1, fitness, nba, golf

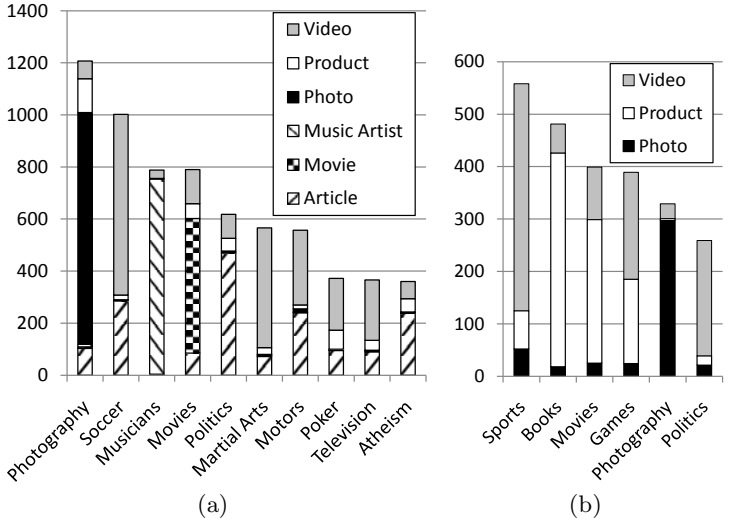


Fig. 2. No. of posts containing links to each type of object for (a) *Forum*, (b) *Twitter*

of posts containing links to each type of object in the *Forum* dataset is shown in Figure 2(a), and the number of posts containing links to each type of object for *Twitter* is shown in Figure 2(b). For the *Forum* dataset, hyperlinks to music artists occur mainly in the Musicians forum, movies in the Films forum, and photos in the Photography forum. The other object types are spread more evenly between the remaining seven forums. In total, *Forum* contains 6,626 posts and *Twitter* contains 2,415 posts. Note that in rare cases in *Forum*, a post contains links to multiple object types, in which case that post is included twice in a column. Therefore the total counts in Figure 2(a) are inflated by approximately 1%. For our analysis, we select only the most commonly available metadata types in order to make comparisons between them, but our method could be applied using arbitrary metadata. The metadata types that we chose were Title, Category (includes music/movie genre), Description (includes Wikipedia abstract), Tags and Author/Director (for Amazon books and IMDB movies only).

HTML documents. We crawled the corresponding HTML document for each hyperlink remaining in the datasets. For any cases where a HTML document was not retrievable, this object was removed from the dataset. We stripped out HTML tags and retained only the visible text of the webpage.

5 Analysis of the External Metadata

We now investigate some features of the metadata that was collected for the *Forum* dataset. Statistics are not reported for *Twitter* due to space constraints. Note that this analysis was performed after pre-processing the metadata text.

The first section of Table 3 shows the percentage of non-empty metadata for each type of object. This is of interest since a metadata type that occurs rarely will have limited usefulness. Due to the unique features of each website, not every object type can have every metadata type. There are large variations in the percentage of non-empty features for different metadata types. Titles are typically essential to identify an object and categories are typically required by a website's browsing interface, so these features are almost always present. For user-generated content, the frequency of non-empty fields is depends on whether the field is mandatory. For example, tags are often absent in Flickr because they are optional, while for videos they are almost always present because in the absence of user-provided tags, YouTube automatically assigns tags. For products, the author feature is often empty since this field is only available for books. For movies, the director feature is sometimes empty, presumably due to some inconsistencies in the various sources from which LinkedMDB integrates data.

The second section of Table 3 shows the average number of unique tokens found in non-empty metadata fields. These figures are an indicator of how much information each feature provides. In general, titles and authors/directors provide few tokens since they are quite short. For categories, the number of tokens depends on whether the website allows multiple categories (*e.g.*, Wikipedia) or single categories (*e.g.*, YouTube). The number of unique tokens obtained from descriptions and tags are quite similar across all object types studied.

The third section of Table 3 gives the average percentage of unique tokens from metadata that do not occur in post content. This section is important since it shows which features tend to provide novel information. Note that for article titles, the percentage is zero since all titles are contained within the article's URL. For music artist titles, the figure is low since bands often use their title as their username, which is contained within the artist's URL. All other object types have URLs that are independent of the object properties. This section also allows us to see how users typically describe an object. For example, 40% of the tokens from product titles are novel, indicating that posters often do not precisely name the products that they link to. For the subset of products that are books, 23% of tokens from titles were novel. Approximately 32% of the tokens from book authors and 43% of the tokens from movie directors are novel, showing that posters often mention these names in their posts, but that in many other cases this is new information which can aid retrieval.

Table 3. Properties of external metadata content for *Forum*

	Title	Category	Description	Tags	Author/ Director
Average % of text features that are non-empty after pre-processing					
Article	100.0	100.0	99.7	-	-
Movie	100.0	100.0	-	-	39.9
Music Artist	99.7	100.0	-	-	-
Photo	100.0	-	58.8	84.9	-
Product	100.0	100.0	-	75.2	65.4
Video	100.0	100.0	99.5	99.5	-
Average unique metadata tokens for non-empty fields (± standard deviation)					
Article	2.1 ± 0.9	13.6 ± 12.1	15.8 ± 8.3	-	-
Movie	1.7 ± 0.7	4.1 ± 1.8	-	-	2.2 ± 0.6
Music Artist	1.8 ± 0.9	2.7 ± 0.9	-	-	-
Photo	2.0 ± 1.1	-	10.9 ± 17.2	6.5 ± 4.9	-
Product	5.2 ± 3.0	11.5 ± 7.8	-	5.7 ± 2.1	2.0 ± 0.4
Video	3.7 ± 1.6	1.0 ± 0.0	13.1 ± 26.3	7.2 ± 5.0	-
Average % of unique metadata tokens that are novel (do not occur in post content)					
Article	0.0	78.5	68.4	-	-
Movie	17.4	76.2	-	-	43.3
Music Artist	10.1	85.4	-	-	-
Photo	72.5	-	50.3	74.6	-
Product	39.5	81.0	-	51.1	32.2
Video	62.0	95.7	78.5	74.4	-

6 Classification Experiments

In this section, we evaluate the classification of posts in the *Forum* and *Twitter* datasets, based on different post representations including the original text augmented with external metadata.

6.1 Experimental Setup

For each post, the following representations were derived, in order to compare their usefulness as sources of features for topic classification:

Content (without URLs): Original post content with hyperlinks removed.
Content: The full original content with hyperlinks intact.
HTML: The text parsed from the HTML document(s) to which a post links.
Metadata: The external metadata retrieved from the hyperlinks of the post.

Document length normalisation and tf-idf weighting were applied to each feature vector. We also generate aggregate feature vectors for the combinations of Content+HTML and Content+Metadata. An aggregate vector for two text sources is obtained by adding their individual feature vectors, after document length normalisation and tf-idf weighting. We tested two methods for combining different sources of textual information into a single vector:

Bag of words: The same term in different sources is represented by the same element in the document vector. For these experiments, we test different weightings of the two sources, specifically {0.1:0.9, 0.2:0.8, ... , 0.9:0.1}. Two vectors v_1 and v_2 are combined into a single vector v where a term i in v is given by, for example, $v[i] = (v_1[i] \times 0.1) + (v_2[i] \times 0.9)$.

Concatenate: The same term in different sources is represented by different elements in the feature vector - *i.e.*, "music" appearing in a post is distinct from "music" in a HTML page. Two vectors v_1 and v_2 are combined into a single vector v via concatenation, *i.e.*, $v = \langle v_1, v_2 \rangle$.

Classification of documents was performed with the Multinomial Naïve Bayes classifier implemented in Weka [10]. A ten-fold cross validation was used to assess the performance of the classifier on each type of document representation. K-fold cross validation involves randomly splitting a dataset into K folds, and using one fold as test data and the remaining $K-1$ folds as training data. The process is repeated so that each of the K folds is used as a test set exactly once. Duplication of hyperlinks across splits was disallowed, so the metadata of a particular object cannot occur in multiple folds. In order to avoid duplication of post content due to one post quoting another, *Forum* was split by thread so that multiple posts from one thread do not occur in separate folds. Duplication was not an issue in *Twitter* since retweets had been removed. These restrictions resulted in the omission of approximately 11% of the *Forum* posts from any fold.

6.2 Experimental Results

The accuracy of classification for each representation is measured using the F_1 measure, which takes into account both precision p and recall r and is defined as $F_1 = \frac{2.p.r}{p+r}$. Micro-averaged F_1 is calculated by averaging F_1 over each test instance and is therefore more heavily influenced by common categories, while macro-averaged F_1 is calculated by averaging F_1 over the result for each category and is therefore more heavily influenced by rare categories.

The results of the classification experiments for each post representation are shown in Table 4 with their 90% confidence intervals. For both datasets, classification results based on content improve when tokens from URLs within posts are included. Classification using only the HTML pages linked to by posts gives relatively poor results, while classification using only metadata from hyperlinked objects improves accuracy for *Forum*, but decreases accuracy for *Twitter*. Those differences are all statistically significant. For the combined representations, the bag-of-words representation gives slightly better results than concatenation. The results reported are for the best-performing weightings. For *Forum*, these were 0.9:0.1 for Content+HTML and 0.5:0.5 for Content+Metadata. For *Twitter*, these were 0.9:0.1 for Content+HTML and 0.8:0.2 for Content+Metadata. For both HTML and Metadata, a bag-of-words combination with Content outperforms results for Content alone. The Content+Metadata approach significantly outperforms the Content+HTML approach, for both datasets.

Table 4. Micro-averaged F_1 for (± 90% Confidence Interval)

Data Source	*Forum*		*Twitter*	
	Bag of Words	Concatenate	Bag of Words	Concatenate
Content (without URLs)	0.745 ± 0.009	-	0.722 ± 0.019	-
Content	0.811 ± 0.008	-	0.759 ± 0.015	-
HTML	0.730 ± 0.007	-	0.645 ± 0.020	-
Metadata	0.835 ± 0.009	-	0.683 ± 0.018	-
Content+HTML	0.832 ± 0.007	0.795 ± 0.004	0.784 ± 0.016	0.728 ± 0.016
Content+Metadata	0.899 ± 0.005	0.899 ± 0.005	0.820 ± 0.013	0.804 ± 0.018

Table 5. F_1 achieved by classifier for each category, ordered by performance

Forum				*Twitter*			
Forum	Content	Metadata	Content +M'data	Forum	Content	Metadata	Content +M'data
Musicians	*0.973*	0.911	0.981	Books	0.804	*0.836*	0.877
Photography	*0.922*	0.844	0.953	Photography	*0.785*	0.728	0.842
Soccer	0.805	*0.902*	0.945	Games	*0.772*	0.675	0.830
Martial Arts	0.788	*0.881*	0.917	Movies	0.718	*0.777*	0.827
Motors	0.740	*0.869*	0.911	Sports	*0.744*	0.563	0.781
Movies	0.825	*0.845*	0.881	Politics	*0.685*	0.499	0.733
Politics	*0.791*	0.776	0.846				
Poker	0.646	*0.757*	0.823				
Atheism	*0.756*	0.732	0.821				
Television	0.559	*0.664*	0.716				
Macro-Avgd	0.781	0.818	0.879	Macro-Avgd	0.751	0.680	0.815

Table 5 shows the detailed results for each category, for Content, Metadata and Content+Metadata (using the bag-of-words weighting with the best performance). There is a large variation in classification results for different categories. For post classification based on Content, *Forum* results vary from 0.973 down to 0.559 and *Twitter* results vary from 0.804 down to 0.685. Despite the variation between categories, Content+Metadata always results in the best performance. For the two single source representations, some categories obtain better results using Content and others using Metadata. The higher result between these two representations is highlighted with italics.

Table 6 shows the gains in accuracy achieved by performing classification based on different types of metadata from Wikipedia articles and YouTube videos, for the *Forum* dataset. We limit our analysis to these object types because they have consistently good coverage across all of the forums, apart from Musicians which we excluded from this analysis. These results are based only on the posts with links to objects that have non-empty content for every metadata type and amount to 1,623 posts for Wikipedia articles and 2,027 posts for YouTube videos. We compare the results against Content (without URLs), because Wikipedia URLs contain article titles and our aim is to measure the

Table 6. Micro-averaged F_1 for classification based on selected metadata types in *Forum* (± 90% Confidence Interval)

Metadata Type	Content (w/o URLs)	Metadata Only	Content+Metadata
Wikipedia Articles			
Category	0.761 ± 0.014	0.811 ± 0.012	0.851 ± 0.009
Description		0.798 ± 0.016	0.850 ± 0.009
Title		0.685 ± 0.016	0.809 ± 0.011
YouTube Videos			
Tag	0.709 ±0.011	0.838 ± 0.019	0.864 ± 0.012
Title		0.773 ± 0.015	0.824 ± 0.013
Description		0.752 ± 0.010	0.810 ± 0.013
Category		0.514 ± 0.017	0.753 ± 0.014

effects of the inclusion of titles and other metadata. Table 6 shows that the results for different metadata types vary considerably. For posts containing links to Wikipedia articles, the article categories alone result in a better classification of the post's topic than the original post content, with an F_1 of 0.811 compared to 0.761. Likewise, for posts that contain links to YouTube videos, the video tags provide a much better indicator of the post topic than the actual post text. The Content+Metadata column shows results where each metadata type was combined with post content (without URLs), using a bag-of-words representation with 0.5:0.5 weightings. Every metadata type examined improved post classification relative to the post content alone. However some metadata types improve the results significantly more than others, with Content+Category achieving the best scores for articles, and Content+Tags achieving the best scores for videos.

7 Discussion

The usage of external information from hyperlinks for categorisation or retrieval on the Web is a well-established technique. Our experiments show that categorisation of social media posts can be improved by making use of semantically-rich data sources where the most relevant data items can be experimentally identified. Both datasets showed similar patterns, although the *Twitter* scores are consistently lower. It may be that the Twitter hashtags are not as accurate descriptors of topic as the forum categories. Also, for *Forum* the external metadata is a better indicator of the category than the post content while for *Twitter* the reverse is true. This may be partially due to the fact that the distribution of domains linked to in each dataset is different and some domains may provide more useful information than others, either within URLs or within metadata.

We also observe that results vary considerably depending on the topic that is under discussion. For example in *Forum*, classification of a post in the Musicians forum is trivial, since almost all posts that feature a link to MySpace belong here. In contrast, the classification of a Television forum post is much more challenging, because this forum mentions a wide variety of topics which are televised. We also note that some topics achieve better classification results

using only external metadata but others have better results with the original content. In the case of the Musicians and Photography forums, the good results for Content may be due to the fact that links to MySpace are highly indicative of the Musicians forum, and links to Flickr are usually from the Photography forum. The Politics and Atheism forums also achieve better results based on post content - this may be because they have a high percentage of links to Wikipedia articles, whose URLs include title information. We can conclude for posts whose hyperlinks contain such useful indicators, the addition of external metadata may give only a slight improvement, but for posts whose URLs do not give such explicit clues, the addition of external metadata can be an important advantage for topic classification.

A major benefit of using structured data rather than HTML documents is that it becomes possible to compare the improvements gained by integrating different metadata types. Our results show that the effect of the addition of different metadata types varies greatly, *e.g.*, Wikipedia categories and descriptions are much more useful than article titles. The benefit of different metadata types is not consistent across sites - Wikipedia's rich categories are far more useful than YouTube's limited categories. Often particular metadata types from hyperlinked objects in a post can be a better descriptor of the post topic than the post itself, for example YouTube tags, titles and descriptions. In these cases the structure of the data could be exploited to highly weight the most relevant metadata types. Thus, even classification on unstructured Web content can immediately benefit from semantically-rich data, provided that there are hyperlinks to some of the many websites that do provide structured data. While this paper focused on commonly-available metadata types, our approach could be applied to arbitrary metadata types from unknown sources, where machine-learning techniques would be employed to automatically select and weight the most useful metadata.

In our experiments, we used the structure of the external data to identify which types of metadata provide the most useful texts for improving classification. In addition to providing metadata, the Linked Data sources are also part of a rich interconnected graph with semantic links between related entities. We have shown that the textual information associated with resources can improve categorisation, and it would be interesting to also make use of the semantic links between concepts. For example, imagine a Television post contains links to the series `dbpedia:Fawlty_Towers`. A later post that links to the series `dbpedia:Mr_Bean` could be classified under the same category, due to the fact that the concepts are linked in several ways, including their genres and the fact that they are both produced by British television channels. Just as we used machine-learning techniques to identify the most beneficial metadata types, we could also identify the most useful properties between entities.

Potential applications for our approach include categorisation of either new or existing post items. For example, on a multi-forum site (*i.e.*, one that contains a hierarchy of individual forums categorised according to topic area), a user may not know the best forum where they should make their post, or where it is most likely to receive comments that are useful to the user. This can be the case where

the content is relevant to not just one forum topic but to multiple topic areas. On post creation, the system could use previous metadata-augmented posts and any links if present in the new post to suggest potential categories for this post. Similarly, posts that have already been created but are not receiving many comments could be compared against existing augmented posts to determine if they should be located in a different topic area than they are already in.

This approach also has potential usage across different platforms. While it may be difficult to use augmented posts from Twitter to aid with categorisation of posts on forums due to the differing natures of microblogs and discussion forums, there could be use cases where augmented posts from discussion forums, news groups or mailing lists (*e.g.*, as provided via Google Groups) could be used to help categorisations across these heterogeneous, yet similar, platforms. Also, the categories from augmented discussion forum posts could be used to recommend tags or topics for new blog content at post creation time.

8 Conclusion

In this work, we have investigated the potential of using metadata from hyperlinked objects for classifying the topic of posts in online forums and microblogs. The approach could also be applied to other types of social media. Our experiments show that post categorisation based on a combination of content and object metadata gives significantly better results than categorisation based on either content alone or content and hyperlinked HTML documents. We observed that the significance of the improvement obtained from including external metadata varies by topic, depending on the properties of the URLs that tend to occur within that category. We also found that different metadata types vary in their usefulness for post classification, and some types of object metadata are even more useful for topic classification than the actual content of the post. We conclude that for posts that contain hyperlinks to structured data sources, the semantically-rich descriptions of entities can be a valuable resource for post classification. The enriched structured representation of a post as content plus object metadata also has potential for improving search in social media.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: 1st Int'l Conference on Web Search and Data Mining, WSDM 2008. ACM, New York (2008)
2. Angelova, R., Weikum, G.: Graph-based text classification: Learn from your neighbors. In: 29th Int'l SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2006. ACM, New York (2006)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A nucleus for a web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)

4. Berendt, B., Hanser, C.: Tags are not metadata, but "just more content"–to some people. In: 5th Int'l Conference on Weblogs and Social Media, ICWSM 2007 (2007)
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The story so far. International Journal on Semantic Web and Information Systems 5(3) (2009)
6. Breslin, J.G., Harth, A., Bojars, U., Decker, S.: Towards semantically-interlinked online communities. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 500–514. Springer, Heidelberg (2005)
7. Cha, M., Pérez, J., Haddadi, H.: Flash Floods and Ripples: The spread of media content through the blogosphere. In: 3rd Int'l Conference on Weblogs and Social Media, ICWSM 2009 (2009)
8. Figueiredo, F., Belém, F., Pinto, H., Almeida, J., Gonçalves, M., Fernandes, D., Moura, E., Cristo, M.: Evidence of quality of textual features on the Web 2.0. In: 18th Conference on Information and Knowledge Management, CIKM 2009. ACM, New York (2009)
9. Garcia Esparza, S., O'Mahony, M.P., Smyth, B.: Towards tagging and categorization for micro-blogs. In: 21st National Conference on Artificial Intelligence and Cognitive Science, AICS 2010 (2010)
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. ACM SIGKDD Exp. 11(1) (2009)
11. Irani, D., Webb, S., Pu, C., Li, K.: Study of trend-stuffing on Twitter through text classification. In: 7th Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, CEAS 2010 (2010)
12. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. 60(11) (2009)
13. Kinsella, S., Passant, A., Breslin, J.G.: Using hyperlinks to enrich message board content with Linked Data. In: 6th Int'l Conference on Semantic Systems, I-SEMANTICS 2010. ACM, New York (2010)
14. Kinsella, S., Passant, A., Breslin, J.G.: Topic classification in social media using metadata from hyperlinked objects. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Murdock, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 201–206. Springer, Heidelberg (2011)
15. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: Can we trust what we RT? In: 1st Workshop on Social Media Analytics, SOMA 2010. ACM, New York (2010)
16. Qi, X., Davison, B.: Classifiers without borders: Incorporating fielded text from neighboring web pages. In: 31st Int'l SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008. ACM, New York (2008)
17. Sergey, B., Lawrence, P.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7), 107–117 (1998)
18. Stankovic, M., Rowe, M., Laublet, P.: Mapping tweets to conference talks: a goldmine for semantics. In: 3rd Int'l Workshop on Social Data on the Web, SDoW 2010 (2010), CEUR-WS.org
19. Sun, A., Suryanto, M.A., Liu, Y.: Blog classification using tags: An empirical study. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 307–316. Springer, Heidelberg (2007)
20. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Fourth Int'l Conference on Web Search and Data Mining, WSDM 2011. ACM, New York (2011)
21. Yin, Z., Li, R., Mei, Q., Han, J.: Exploring social tagging graph for web object classification. In: 15th SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, KDD 2009. ACM, New York (2009)