Improving Categorisation in Social Media using Hyperlinks to Structured Data Sources

Sheila Kinsella, Mengjiao Wang, John Breslin, Conor Hayes

Introduction

- Topic categorisation:
  - Given a set of classes (topics), we seek to determine which class a given document (post) belongs to
  - Usually based on document content
- Applications in social media :
  - Categorise existing posts for enhanced browsing
    - *e.g.*, to find Twitter posts on a certain topic
  - Suggest categories for new posts on creation
    - *e.g.*, on a message board site where posts are often incorrectly placed

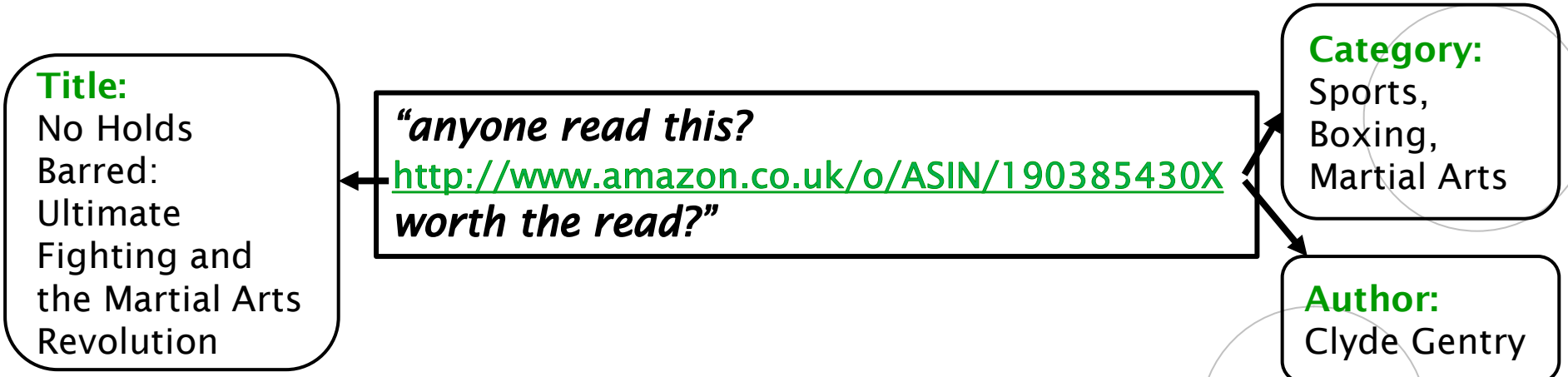
Challenges

- Social media posts are:
  - Short and informal
  - Often reliant on external hyperlinks for context

> ***"anyone read this?***
> ***http://www.amazon.co.uk/o/ASIN/190385430X***
> ***worth the read?"***

Challenges

- Social media posts are:
  - Short and informal
  - Often reliant on external hyperlinks for context



Motivation: Why hyperlinks?

- Hyperlinks often point to objects which are the topic of the post
  - *e.g.*, a movie on IMDB, a Wikipedia article
- Hyperlinks often contain novel and useful information
  - In our forum dataset, of posts that linked to books,
    - 65% mention neither complete title nor complete author
    - at least 11% do not contain even a partial title or author
- Even though hyperlinks only occur in a subset of posts (in our forum dataset, 4%-14%), these posts are more likely to be a source of information rather than just chat

Motivation: Why structured data?

- People don't just link to documents with text, but to objects with metadata
  - Structure enables us to extract only the most relevant data
  - We can experimentally identify most useful metadata types
- Recently the amount of such data is rapidly growing
  - In our forum dataset, 23% of posts with hyperlinks link to an object with structured data available from APIs or Linked Data

Approach

- Identify sources of structured data from hyperlinks
  - Based on domains, *e.g.*, `wikipedia.org`
- Retrieve structured data for these hyperlinks
  - From Linked Data/APIs, *e.g.*, `dbpedia.org`
  - We also retrieve the HTML representation, for comparison
- Perform text classification
  - We use a Naïve Bayes classifier
  - Requires set of already categorised posts for training
  - Post content and external metadata as sources of textual features
  - Compare accuracy achieved by different metadata types
- Related to IR studies that classify documents based on fielded text from hyperlinked pages, but they consider *structural* rather than *semantic* fields

Datasets

	Forum	*Twitter*
Data source	message board	microblogging site
Ground truth topics	forums	#hashtags
# classes (topics)	10	6
# posts	6,626	2,415

External structured data sources

amazon.com

DBTune.org

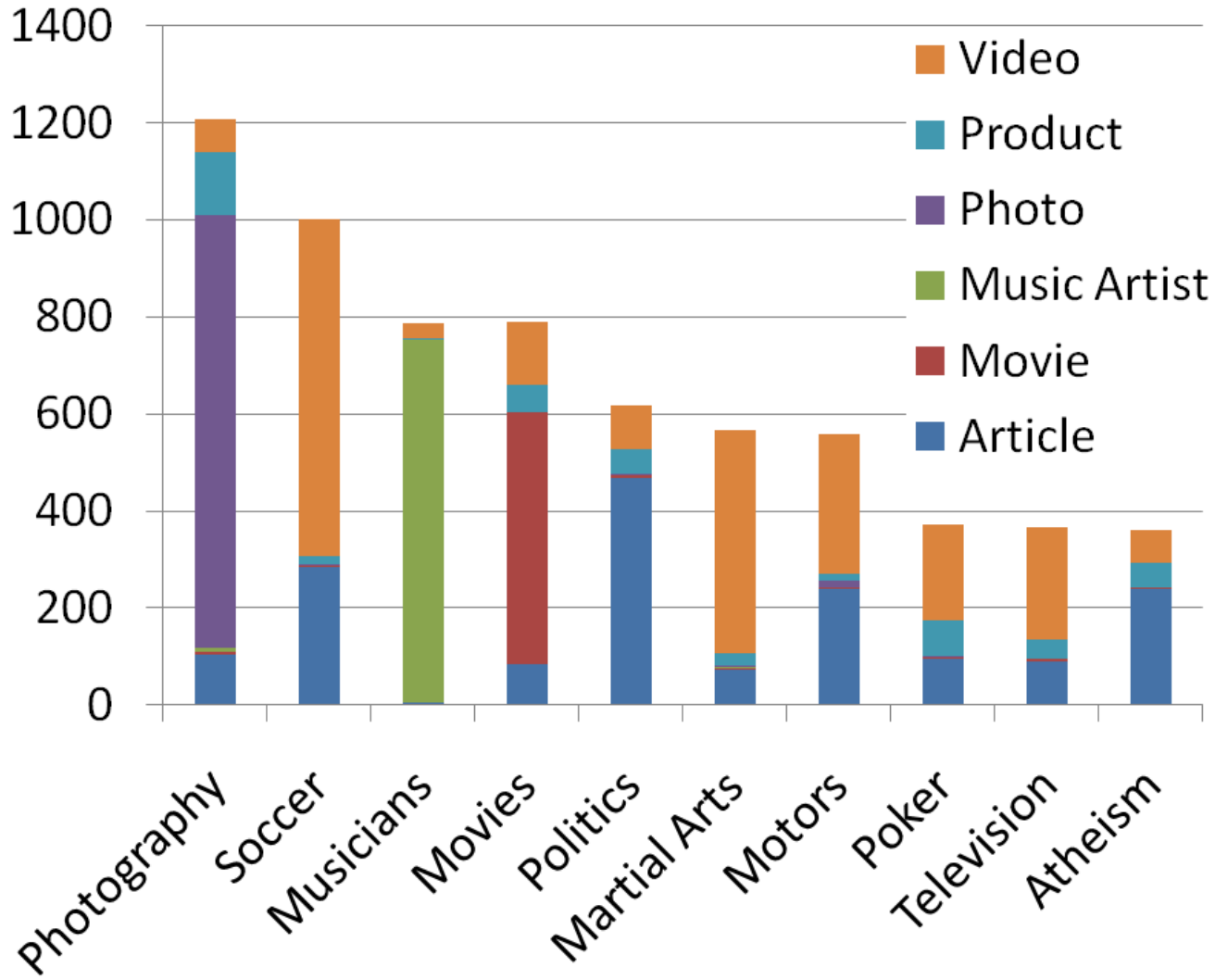
flickr from YAHOO!

YouTube

External structured data sources

Linked Data	Web APIs
DBpedia	amazon.com
DBTune.org	flickr from YAHOO!
Linked Movie DataBase (preview)	YouTube

Objects linked to in *Forum*



Experiments

- Multinomial Naïve Bayes classifier (WEKA)
  - tf-idf and document length normalisation
  - Lower-cased, removed stopwords and non-alphabetic chars
- 10-fold cross-validation
- Compared classification accuracy for different post representations based on post content, hyperlinked HTML pages and hyperlinked object metadata
- Experimented to find optimal ways of combining feature vectors (*e.g.*, weightings)

Post Representations

- Content (no URLs)
- Content (with URLs)
- External HTML pages
- External Metadata
  - title
  - description/abstract
  - tags
  - categories/genre
  - author/director
- Content + HTML
- Content + Metadata

Results – single sources

Data Source	*Forum*	*Twitter*
Content (no URLs)	0.745	0.722
Content (with URLs)	0.811	**0.759**
HTML	0.730	0.645
Metadata	**0.835**	0.683

(micro-averaged F_1)

Results – all sources

Data Source	*Forum*	*Twitter*
Content (no URLs)	0.745	0.722
Content (with URLs)	0.811	**0.759**
HTML	0.730	0.645
Metadata	**0.835**	0.683
Content + HTML	0.832	0.784
Content + Metadata	**0.899**	**0.820**

(micro-averaged F_1)

Results per topic in *Forum*

Topic	Content	Content + Metadata	change
Musicians	0.973	0.981	+0.008
Photography	0.922	0.953	+0.031
Soccer	0.805	0.945	+0.140
Martial Arts	0.788	0.917	+0.129
Motors	0.740	0.911	+0.171
Movies	0.825	0.881	+0.056
Politics	0.791	0.846	+0.055
Poker	0.646	0.823	+0.177
Atheism	0.756	0.821	+0.065
Television	0.559	0.716	+0.157

Comparing metadata types

- We identified posts in *Forum* which link to Wikipedia articles
  - 1.6k posts
- Classified based only on Wikipedia metadata

Metadata type	Content (no URLs)	Metadata only	Content+ Metadata
Category	0.761	**0.811**	**0.851**
Description		0.798	0.850
Title		0.685	0.809

Comparing metadata types

- We identified posts in *Forum* which link to YouTube videos
  - 2k posts
- Classified based only on YouTube metadata

Metadata type	Content (no URLs)	Metadata only	Content+ Metadata
Tag	0.709	**0.838**	**0.864**
Title		0.773	0.824
Description		0.752	0.810
Category		0.514	0.753

Metadata types on different sites

Wikipedia

Metadata type	Content (no URLs)	Metadata only	Content+ Metadata
Category	0.761	0.811	0.851
Description		0.798	0.850
Title		0.685	0.809

YouTube

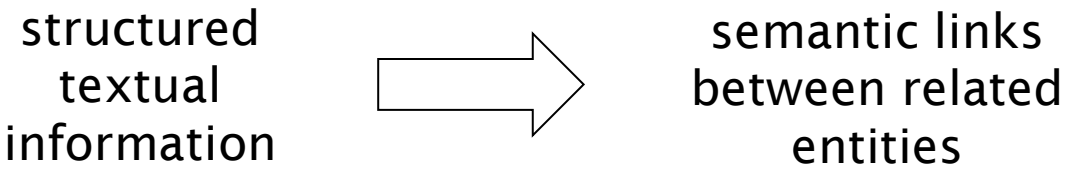
Metadata type	Content (no URLs)	Metadata only	Content+ Metadata
Tag	0.709	0.838	0.864
Title		0.773	0.824
Description		0.752	0.810
Category		0.514	0.753

Summary of results

- For both datasets, a combination of post content and metadata gives best classification results
- Improvement varies by category - depends on characteristics of URLs
- The most useful metadata types can be found experimentally, but for different objects, the usefulness of metadata types varies

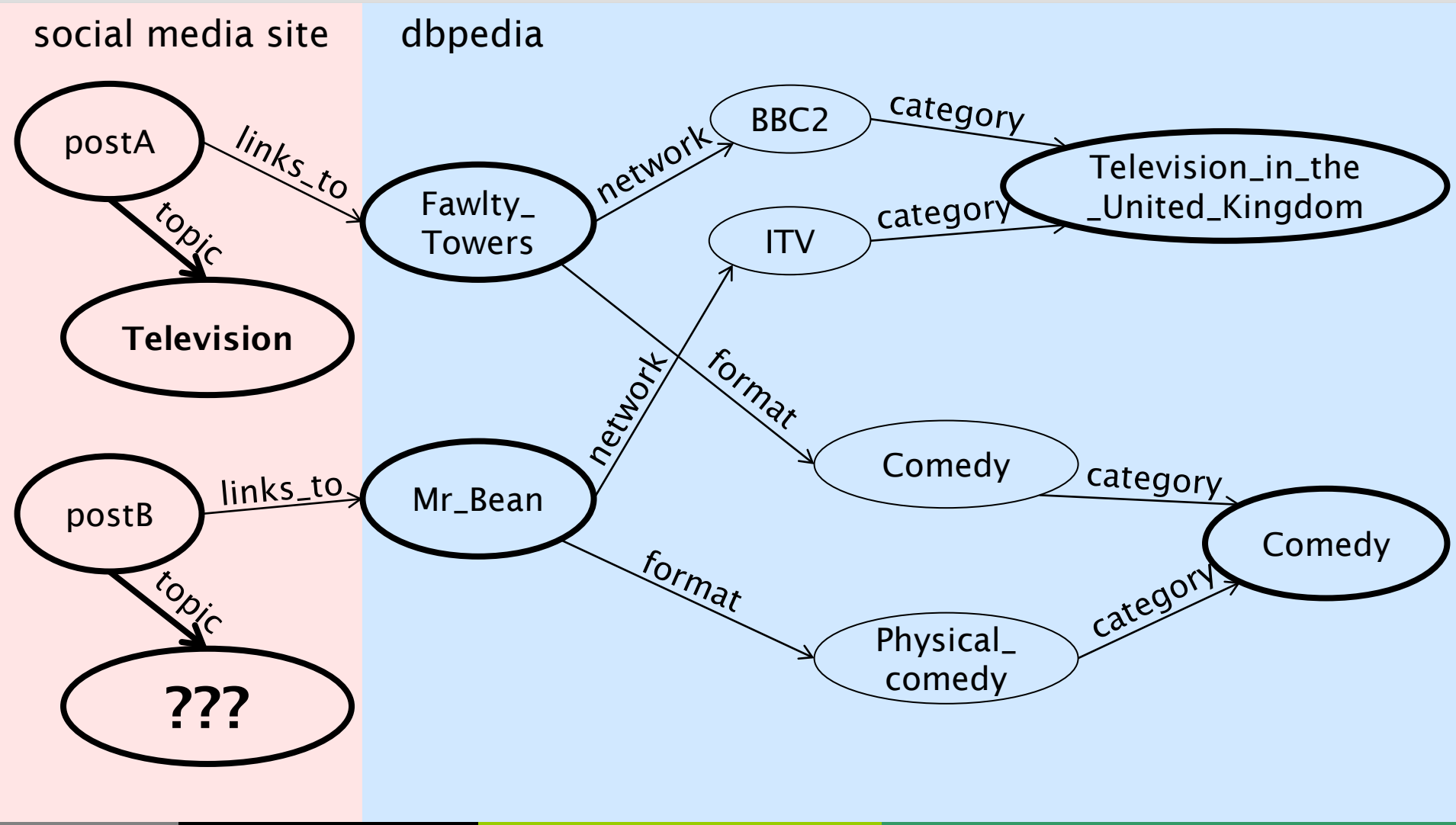
Future Directions

- We have shown that standard IR techniques can by improved by making use of semantic data sources
- Next, we want to make more use of semantic links



- Similar to experimentally identifying the most useful metadata types, we could identify the most beneficial property paths between entities
- With the large amounts of RDFa now available (*e.g.*, from Facebook Open Graph Protocol), we could greatly increase the coverage of our approach
- Not limited to social media, but any unstructured data which has links to sources of structured data

Example of using semantic links



Conclusions

- Topic classification in social media can be improved by making use of structured metadata from hyperlinked objects
- This shows how even unstructured Web content can benefit from more structured data on the Web
- Linked Data sources have the potential to give even more improvements, by considering semantic links as well as textual metadata

Additional Slides

Novelty of metadata in *Forum*

- Percentage of tokens from metadata which do not occur in the text of the original post:
  - After lower-casing, removing stopwords and non-alphabetic chars

Object	Website	Title	Category /Genre	Description /Abstract	Tags	Author/ Director
Article	Wikipedia	0%	79%	68%	-	-
Movie	IMDB	17%	76%	-	-	43%
Music Artist	Myspace	10%	85%	-	-	-
Photo	Flickr	73%	-	50%	75%	-
Product	Amazon	40%	81%	-	51%	32%
Video	YouTube	62%	96%	78%	74%	-

#hashtag → category mappings

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

Results per topic in *Forum*

Topic	Content (with URLs)	Metadata
Musicians	**0.973**	0.911
Photography	**0.922**	0.844
Soccer	0.805	**0.902**
Martial Arts	0.788	**0.881**
Motors	0.740	**0.869**
Movies	**0.825**	0.845
Politics	**0.791**	0.776
Poker	0.646	**0.757**
Atheism	**0.756**	0.732
Television	0.559	**0.664**

(F_1)