Semantic Web 0 (0) 1
IOS Press

Making Sense of Social Media Streams through Semantics: a Survey

Kalina Bontcheva [a,*] Dominic Rout [a]
[a] *Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield, United Kingdom*
E-mail: Initial.Surname@dcs.shef.ac.uk

Abstract.
Using semantic technologies for mining and intelligent information access to social media is a challenging, emerging research area. Traditional search methods are no longer able to address the more complex information seeking behaviour in media streams, which has evolved towards sense making, learning, investigation, and social search. Unlike carefully authored news text and longer web context, social media streams pose a number of new challenges, due to their large-scale, short, noisy, context-dependent, and dynamic nature.

This paper defines five key research questions in this new application area, examined through a survey of state-of-the-art approaches to mining semantics from social media streams; user, network, and behaviour modelling; and intelligent, semantic-based information access. The survey includes key methods not just from the Semantic Web research field, but also from the related areas of natural language processing and user modelling. In conclusion, key outstanding challenges are discussed and new directions for research are proposed.

Keywords: semantic annotation, semantic-based user modelling, semantic search, information visualisation, social media streams

1. Introduction

The widespread adoption of social media is based on tapping into the social nature of human interactions, by making it possible for people to voice their opinion, become part of a virtual community and collaborate remotely. If we take micro-blogging as an example, Twitter has 100 million active users, posting over 230 million tweets a day[1].

Engaging actively with such high-value, high-volume, brief life-span media streams has now become a daily challenge for both organisations and ordinary people. Automating this process through intelligent, semantic-based information access methods is therefore increasingly needed. This is an emerging research area, combining methods from many fields, in addition to semantic technologies, e.g. speech and language processing, social science, machine learning, personalisation, and information retrieval.

Traditional search methods are no longer able to address the more complex information seeking behaviour in social media, which has evolved towards sense making, learning and investigation, and social search [**?**]. Semantic technologies have the potential to help people cope better with social media-induced information overload. Automatic semantic-based methods that adapt to individual's information seeking goals and summarise briefly the relevant social media, could ultimately support information interpretation and decision making over large-scale, dynamic media streams.

Unlike carefully authored news and other textual web content, social media streams pose a number of new challenges for semantic technologies, due to their large-scale, noisy, irregular, and social nature. In this paper we discuss the following key research ques-

*Corresponding author. E-mail: K.Bontcheva@dcs.shef.ac.uk.
[1] http://www.guardian.co.uk/technology/pda/2011/sep/08/twitter-active-users (Visited May 7, 2012)

tions, examined through a survey of state-of-the-art approaches:

1. What ontologies and Web of Data resources can be used to represent and reason about the semantics of social media streams?
2. How can semantic annotation methods capture the rich semantics implicit in social media?
3. How can we extract reliable information from these noisy, dynamic content streams?
4. How can we model the users' digital identity and social media activities?
5. What semantic-based information access methods can help address the complex information seeking behaviour in social media?

To the best of our knowledge, this is the first comprehensive meta-review of semantic technology for mining and intelligent information access, where the focus is on current limitations and outstanding challenges, specifically arising in the context of social media streams.

The paper is structured as follows: section 2 provides background on social media, their different characteristics, and the corresponding technological challenges. Section 3 focuses on ontologies which model different kinds of social media, user profiles and networks, information sharing, and other typical social media activities (research question 1). Section 4 discusses methods for semantic annotation of social media streams, in particular the ways in which they capture the rich implicit semantics (research question 2) and deal with the noisy, streaming nature of this type of content (research question 3). Section 5 investigates in depth research question 4, i.e. how are users, networks, and activities modelled semantically and how can this knowledge be used to personalise information access. Next, section 6 analyses state-of-the-art in intelligent information access for social media streams, in the context of research question 5. In conclusion, section 7 defines outstanding challenges and provides directions for future work.

2. Social Media Streams: Characteristics, Challenges and Opportunities

Social media sites allow users to connect with each other for the purpose of sharing content (e.g. web links, photos, videos), experiences, professional information, and online socialising with friends. Users create posts or status updates and social media sites circulate these to the user's social network. The key difference from traditional web pages is that users are not just information consumers, but many are also prolific content creators.

Social media can be categorised on a spectrum, based on the type of connection between users, how the information is shared, and how users interact with the media streams:

- Interest-graph media [103] encourage users to form connections with others based on shared interests, regardless of whether they know the other person in real life. Shared information comes in the form of a stream of messages in reverse chronological order.
- Social networking sites (SNS) encourage users to connect with people they have real-life relationships with. Facebook, for example, provides a way for people to share information, as well as comment on each other's posts. Typically, short contributions are shared, outlining current events in users' lives or linking to something on the internet that users think their friends might enjoy. These status updates are combined into a time-ordered stream for each user to read.
- Professional Networking Services (PNS), such as LinkedIn, aim to provide an introductions service in the context of work, where connecting to a person implies that you vouch for that person to a certain extent, and would recommend them as a work contact for others. Typically, professional information is shared and PNS tend to attract older professionals [117].
- Content sharing and discussion services, such as blogs, video sharing (e.g. YouTube, Vimeo), slide sharing (e.g. SlideShare), and user discussion/review forums (e.g. CNET). Blogs usually contain longer contributions. Readers might comment on these contributions, and some blog sites create a time stream of blog articles for followers to read. Many blog sites also advertise automatically new blog posts through their users' Facebook and Twitter accounts.

This spectrum of social media shows how each service is defined by: the levels and types of connections; the characteristics of the information exchanged; and the type of uptake and usage. Each service has different forms of connection, and these socio-technical forms are influential in shaping what information access methods are applicable, as well as what additional implicit semantics can be exploited.

2.1. Key Social Media Sites

Twitter, a microblogging service in which users share short status updates (140 characters), is the most widely used **interest-graph social media service**. The "following" relationship between users is often one-way. This can be seen in the way an interesting person can attract large numbers of followers who are interested in what he/she has to say. Followers can choose to spread an interesting message to their own followers via a re-tweet (similar to forwarding an email message) or to post a reply back to the message originator. Tweets can also contain user-supplied hashtags, which are any word or acronym preceded by a #. A unique aspect of Twitter is that the majority of posted messages are public, which coupled with Twitter's extensive API, has made it the focus of most research on semantic methods for social media streams.

With respect to message content, Naaman *et al* [84] found over 40% of their sample of tweets were "me now" messages, that is, posts by a user describing what they are currently doing. Next most common were statements and random thoughts, opinions and complaints and information sharing such as links, each taking over 20% of the total. Less common tweet themes were self-promotion, questions to followers, presence maintenance e.g. "I'm back!", anecdotes about oneself and anecdotes about others. Messages posted from mobile devices are more likely to be "me now" messages (51%). Females post more "me now" messages than males. A relatively small number of people undertake information sharing as a major activity; users can be grouped into *informers* and *meformers*, where meformers mostly share information about themselves. Informers and meformers differ in various ways. Informers tend to be more conversational and have more contacts.

Social and professional networking media such as Facebook and LinkedIn enforce relationship reciprocity by requiring that both parties consent to being linked. Due to the more personal nature of information shared through SNS and PNS, privacy is of paramount concern. Hoadly *et al* [58] investigate the reaction among users when Facebook introduced the news feed, in which information users previously would have had to seek out by going to each others' pages is now aggregated into a time-ordered stream and placed front and centre on the site. "Perceived control" and "ease of information access" were determined to be factors in how comfortable a person feels with privacy aspects of using Facebook.

In their review of social and professional media in the workplace context, Skeels and Grudin [117] find tension around privacy in social media use. However, they also report rapid adoption, as numerous benefits are found. In general, LinkedIn usage seems limited to the professional context, whereas Twitter seems broadly undifferentiated in that regard, and an increasing amount of Facebook usage is becoming professional and public too (for example, Facebook allows businesses to create pages on which they may publicise themselves).

Although Facebook is the most popular social media site, having over twice as many users as Twitter, its nearest competitor, the number of connections is somewhat limited to real-life acquaintances, at least for many users, and therefore has a practical upper limit of around a few hundred, with 90% of users having fewer than 500 Facebook friends [45]. Interest-graph media users tend to follow many more people, sometimes running into thousands (e.g. journalists following politicians, sportsmen, and celebrities; companies following customers). This makes microblogging more of a focus of research when it comes to intelligent information access and semantic technologies.

There are, of course, many other popular social media sites, but in terms of information volumes, user base, different user behaviour, and perceived commercial importance, the above three social media platforms are arguably the most influential. Nevertheless, as we will discuss in the rest of this paper, there are complex interactions between these three kinds of social media streams and the more discrete, enduring, and longer online news, blogs, forums, and other textual web documents.

2.2. Why is Social Media Content a Challenge?

State-of-the-art automatic semantic annotation, browsing, and search algorithms have been developed primarily on news articles and other carefully written, long web content [23]. In contrast, most social media streams (e.g. tweets, Facebook messages) are strongly inter-connected, temporal, noisy, short, and full of slang, leading to severely degraded results[2].

These challenging social media characteristics are also opportunities for the development of new semantic technology approaches, which are better suited to media streams:

[2]For instance, named entity recognition methods typically have 85-90% accuracy on news but only 30-50% on tweets [69,104].

Short messages (microtexts) : Twitter and most Facebook messages are very short (140 characters for tweets). Many semantic-based methods reviewed below supplement these with extra information and context coming from embedded URLs and hashtags[3]. For instance, Abel *et al* [2] augment tweets by linking them to contemporaneous news articles, whereas Mendes *et al* exploit online hashtag glossaries to augment tweets [77].

Noisy content : social media content often has unusual spelling (e.g. 2moro), irregular capitalisation (e.g. all capital or all lowercase letters), emoticons (e.g. :-P), and idiosyncratic abbreviations (e.g. ROFL, ZOMG). Spelling and capitalisation normalisation methods have been developed [55], coupled with studies of location-based linguistic variations in shortening styles in microtexts [52]. Emoticons are used as strong sentiment indicators in opinion mining algorithms (see Section 4.5).

Temporal : in addition to linguistic analysis, social media content lends itself to analysis along temporal lines, which is a relatively under-researched problem. Addressing the temporal dimension of social media is a pre-requisite for much-needed models of conflicting and consensual information, as well as for modelling change in user interests. Moreover, temporal modelling can be combined with opinion mining, to examine the volatility of attitudes towards topics over time (e.g. gay marriage).

Social context is crucial for the correct interpretation of social media content. Semantic-based methods need to make use of social context (e.g. who is the user connected to, how frequently they interact), in order to derive automatically semantic models of social networks, measure user authority, cluster similar users into groups, as well as model trust and strength of connection.

User-generated : since users produce, as well as consume social media content, there is a rich source of explicit and implicit information about the user, e.g. demographics (gender, location, age, etc.), interests, opinions. The challenge here is that in some cases, user-generated content is relatively small, so corpus-based statistical methods cannot be applied successfully.

Multilingual : Social media content is strongly multilingual. For instance, less than 50% of tweets are in English, with Japanese, Spanish, Portuguese, and German also featuring prominently [26]. Unfortunately, semantic technology methods have so far mostly focused on English, while low-overhead adaptation to new languages still remains an open issue. Automatic language identification [26,12] is an important first step, allowing applications to first separate social media in language clusters, which can then be processed using different algorithms.

The rest of this paper discusses which of these challenges have been addressed by semantic technologies and how.

3. Ontologies for Representing Social Media Semantics

Ontologies are the corner stone of semantic technology applications. In this section we focus specifically on ontologies created to model different kinds of social media, user profiles, sharing, tagging, liking, and other common user behaviour in social media.

The two most widely used ontologies, which originated from research on the related topic of Social Semantic Web, are FOAF and SIOC. Friend-of-a-Friend[4] (FOAF) is a vocabulary for describing people, including names, contact information, and a generic `knows` relation. FOAF also supports limited modelling of interests by modelling them as pages on the topics of interest. As acknowledged in the FOAF documentation itself, such an ontological model of interests is somewhat limited.

The Semantically-Interlinked Online Communities[5] (SIOC) ontology models social community sites (e.g. blogs, wikis, online forums). Key concepts are forums, sites, posts, user accounts, user groups, and tags. SIOC supports modelling of user interests through the `sioc:topic property`, which has a URI as a value (posts and user groups can also have topics).

The MOAT (Meaning-Of-A-Tag) ontology [92] allows users to define the semantic meaning of a tag through Linking Open Data and ultimately, to create manually semantic annotations of social media. The ontology defines two kinds of tags: global (across

[3] A recently study of 1.1 million tweets has found that 26% of English tweets contain a URL, 16.6% – a hashtag, and 54.8% contain a user name mention [26].

[4] http://xmlns.com/foaf/0.1/

[5] http://sioc-project.org/

all content) and local (particular tag on a given resource). More recently, MOAT was extended towards modelling microblogs [91], through the new concept of `MicroblogPost`, a `sioc:follows` property (representing follower/followee relationships on Twitter), and a `sioc:addressed_to` property for posts that mention a specific user name.

Bottari [27] is an ontology, which has been developed specifically to model relationships in Twitter, especially linking tweets, locations, and user sentiment (positive, negative, neutral), as extensions to the SOIC (Socially-Interlinked Online Communities) ontology. A new TwitterUser class is introduced, coupled with separate *follower* and *following* properties. The Tweet class is a type of sioc:Post and the ontology also distinguishes retweets and replies. Locations (points-of-interest) are represented using the W3C Geo vocabulary[6], which enables location-based reasoning.

DLPO (The LivePost Ontology) provides a comprehensive model of social media posts, going beyond Twitter [111]. It is strongly grounded in fundamental ontologies, such as FOAF, SKOS, and SOIC. It models personal and social knowledge discovered from social media, as well as linking posts across personal social networks. The ontology captures six main types of knowledge: online posts, different kinds of posts (e.g. retweets), microposts, online presence, physical presence, and online sharing practices (e.g. liking, favouriting). However, while topics, entities, events, and time are well covered, user behaviour roles and individual traits are not addressed as comprehensively as in the SWUM ontology [98] discussed below.

User modelling ontologies are key to the representation, aggregation, and sharing of information about users and their social media interactions. The General User Modelling Ontology (GUMO), for instance, aims to cover a wide range of user-related information, such as demographics, contact information, personality, etc. However, it falls short of representing user interests, which makes it unsuitable for social media, where this is key.

Based on an analysis of 17 social web applications, Plumbaum *et al* [98] have derived a number of user model dimensions required for a social web user modelling ontology. Their taxonomy of dimensions includes demographics, interests and preferences, needs and goals, mental and physical state, knowledge and background, user behaviour, context, and individual traits (e.g. cognitive style, personality). Based on these, they have created the SWUM (Social Web User Model) ontology. A key shortcoming of SWUM, however, is its lack of grounding in other ontologies. For instance, user location attributes, such as Country and City, are coded as strings, which severely limits their usefulness for reasoning (e.g. it is hard to find all users based in South West England, based on their cities). A more general approach would have been to define these through URIs, grounded in commonly used Linked Data resources, such as DBPedia and Freebase.

Lastly, the User Behaviour Ontology [7] models user interactions in online communities. It has been used to model user behaviour in online forums [7] and also Twitter discussions [106]. It has classes that model the impact of posts (replies, comments, etc), user behaviour, user roles (e.g. popular initiator, supporter, ignored), temporal context (time frame), and other interaction information. Addressing the temporal dimension of social media is of particular important, especially when modelling changes over time (e.g. in user interests or opinions).

To summarise, there are a number of specialised ontologies, aimed at representing and reasoning with automatically derived semantic information from social media. However, none of these ontologies is comprehensive enough to subsume all others, so many applications adopt or extend more than one, in order to meet their requirements.

4. Semantic Annotation of Social Media

The process of tying semantic models and natural language together is referred to as *semantic annotation*. It may be characterised as the dynamic creation of interrelationships between *ontologies* and unstructured and semi-structured documents in a bidirectional manner. From a technological perspective, semantic annotation is about annotating in texts all mentions of concepts from the ontology (i.e., classes, instances, properties, and relations), through metadata referring to their URIs in the ontology. Approaches which enhance the ontology with new instances derived from texts are typically referred to as *ontology population*. For an in-depth introduction to ontology-based semantic annotation from textual documents see [23].

Semantic annotation can be performed manually, automatically, or semi-automatically, i.e., first an automatic system creates some annotations and these are then post-edited and corrected by human annotators.

[6] http://www.w3.org/2003/01/geo/

In the context of social media, the Semantic Microblogging (SMOB) framework has been proposed [91], in order to allow users to add manually machine-readable semantics to messages. SMOB supports also interlinking with the LOD cloud, through hashtags. Hepp [57] proposes a different manual semantic annotation syntax for tweet messages, which is then mapped to RDF statements. The syntax supports relationships between tags (including sameAs), properties from ontologies such as FOAF, and multiple RDF statements in the same tweet.

However, while such manual semantic annotation efforts are valuable, automatic semantic annotation methods are required, in order to make sense of the millions of messages posted daily on Facebook, Twitter, LinkedIn, etc. Consequently, in this section we focus primarily on automatic approaches.

Information Extraction (IE), a form of natural language analysis, is becoming a central technology for bridging the gap between unstructured text and formal knowledge expressed in ontologies. Ontology-Based IE (OBIE) is IE which is adapted specifically for the semantic annotation task. One of the important differences between traditional IE and OBIE is in the use of a formal ontology as one of the system's inputs and as the target output. Some researchers (e.g., [73]) call ontology-based any system which specifies its outputs with respect to an ontology, however, in our view, if a system only has a mapping between the IE outputs and the ontology, this is not sufficient and therefore, such systems should be referred as *ontology-oriented*.

Another distinguishing characteristic of the ontology-based IE process is that it not only finds the (most specific) class of the extracted entity, but also identifies it, by linking it to its semantic description in the target knowledge base, typically via a URI. This allows entities to be traced across documents and their descriptions to be enriched during the IE process. In practical terms, this requires automatic recognition of named entities, terms, and relations and also co-reference resolution both within and across documents. These more complex algorithms are typically preceded by some shallow linguistic pre-processing (tokenisation, Part-Of-Speech (POS) tagging, etc.)

Linking Open Data resources, especially DBpedia, YAGO and Freebase, have become key sources of ontological knowledge for semantic annotation, as well as being used as target entity knowledge bases for disambiguation. These offer: (i) cross-referenced domain-independent hierarchies with thousands of classes and relations and millions of instances; (ii) an inter-linked and complementary set of resources with synonymous lexicalisations; (iii) grounding of their concepts and instances in Wikipedia entries and other external data. The rich class hierarchies are used for fine-grained classification of named entities, while the knowledge about millions of instances and their links to Wikipedia entries are used as features in the OBIE algorithms.

The rest of this section focuses specifically on methods for semantic annotation of social media streams.

4.1. Keyword Extraction

Automatically selected keywords are useful in representing the topic of a document or collection of documents, and less effective in delivering arguments or full statements contained therein. Keyword extraction can therefore be considered as a form of shallow knowledge extraction, giving a topical overview. Keywords can also be used in the context of semantic annotation and retrieval, as a means of dimensionality reduction and allowing systems to deal with smaller sets of important terms rather than whole documents.

Some keyword extraction approaches exploit term co-occurrence; forming a graph of terms with edges derived from the distance between occurrences of a pair of terms and assigning weights to vertices [80]. This class of keyword extraction was found to perform favourably on Twitter data compared to methods which relied on text models [130].

These graph-based approaches to extracting keywords from Twitter perhaps perform well because the domain contains a great deal of redundancy [133]. While this property of Twitter and other social media is somewhat beneficial when producing keyword summaries, another less helpful trait is the sheer variety of topics discussed. In cases when documents discuss more than one topic, it can be more difficult to extract a coherent and faithful set of keywords from it.

Personal Twitter timelines, when treated as single documents, present this problem. Users are generally capable of posting on multiple topics. While [130] use TextRank on the whole of a user's stream, they do not attempt to model or address topic variation, unlike [131], who incorporated topic modelling into their approach. Theirs is not the only application of Topic Modelling to Twitter data, as it is similar to [102]. However in the latter work topics are discovered but never summarised.

4.1.1. Global topics

For many, one of the most exciting aspects of Twitter as a messaging platform is that at any point a suitably public conversation can become "global" - that is, the discussion transcends the social group of those that started it and is addressed by the community at large. These wider discussion topics are generally marked with hashtags; those that grow large enough and quickly enough can be referred to as "trending topics" and the Twitter online service shows a selection of the most popular, encouraging users to join in.

These trending topics have the potential to "go viral"; that is, they can become extremely popular through discussion and sharing. Celebrity users have some capacity to manipulate these trends, introducing popular threads of discussion, but the topics must also be engaged with by their followers in order to spread more widely [9]. Trending topics are not necessarily the result of celebrity mentions then wider spreading, as external events such as television broadcasts can prompt independent discussion by many users within the same thread.

Trending topics are considered to be something of an indicator of public mood. Consistent with this assumption is the work of [11] and [126], which showed the predictive power of message volume alone in terms of votes or sales. Political outrage, reaction to major news events and mocking of public figures have all occupied trending topics in the past, indicating their usefulness in areas such as public relations and policy creation. However, by their very nature globally trending topics contain many tweets and can be difficult to read and interpret.

Larger threads like trending topics tend to contain a great deal of redundancy thanks to retweeting of messages and copying-and-pasting. [114] extracted keyphrases for trending topics by exploiting textual redundancy and selecting common sequences of words. Their short phrases are similar to the pithy, manually generated summaries created by users of services like WhatTheTrend.

The recognition of trending topics in media streams can be useful for user interest modelling (see Section 5.1.2), as well as for semantic-based browsing and visualisation (see Section 6). More specifically, user interests could be separated into "global" ones (based on the user's tweets on trending topics) versus "user-specific" (topics which are of more personal interest, e.g. work, hobby, friends).

4.1.2. Automatic Tagging through Keyphrases

Social tagging and bookmarking services such as Flickr, Delicious, and Bibsonomy, are hugely popular. The user created tags and folksonomies are a kind of crowdsourced, informal semantic resource. One particular semantic annotation task is the automatic tagging of new documents with folksonomy tags.

One of the early approaches is the AutoTag system [83], which assigns tags to blog posts. First, it finds similar pre-indexed blog posts using standard information retrieval methods, using the new blog post as the query. Then it composes a ranked list of tags, derived from the top most relevant posts, boosted with with information about tags used previously by the given blogger.

More recent approaches use keyphrase extraction from blog content, in order to suggest new tags. For instance, [101] generate candidate keyphrases from n-grams, based on their POS tags, then filter these using a supervised, logistic regression classifier. The keyphrase-based method can be combined with information from the folksonomy [118], in order to generate tag signatures (i.e. associate each tag in the folksonomy with weighted, semantically related terms). These are then compared and ranked against the new blog post, in order to suggest the most relevant set of tags.

4.2. Ontology-Based Entity Recognition in Social Media

Ontology-based entity recognition is often broken down into two main phases: entity annotation (or candidate selection) and entity linking (also called reference disambiguation or entity resolution). Ontology-based entity annotation is concerned with identifying all mentions in the text of classes and instances from the ontology (e.g. DBPedia). The entity linking step then uses contextual information from the text, as well as knowledge from the ontology to choose the correct URI. However, it must be noted that not all methods carry out both steps, i.e. some only identify mentions of entities in the text and their class [67].

4.3. Wikipedia-based Approaches

Most recent work on entity recognition and linking has used Wikipedia as a large, freely available human-annotated training corpus. The target knowledge bases are typically DBPedia [75] or YAGO [115], due to being derived from Wikipedia and thus offering

a straightforward mapping between an entity URI and its corresponding Wikipedia page. These more recent, ontology-based approaches have their roots in methods that enrich documents with links to Wikipedia articles (e.g. [82]).

Ontology-based entity disambiguation methods typically collect a dictionary of labels for each entity URI, using the Wikipedia entity pages, redirects (used for synonyms and abbreviations), disambiguation pages (for multiple entities with the same name), and anchor text used when linking to a Wikipedia page. This dictionary is used for identifying all candidate entity URIs for a given text mention. Next is the disambiguation stage, where all candidate URIs are ranked and a confidence score is assigned. If there is no matching entity in the target knowledge base, a NIL value is returned. Text mentions can be disambiguated either independently of each other, or jointly across the entire document (e.g. [82]).

Typically methods use Wikipedia corpus statistics coupled with techniques (e.g. TF/IDF) which match the context of the ambiguous mention in the text against the Wikipedia pages for each candidate entity (e.g. [75]). Michelson *et al* [79] demonstrate how such an approach can be used to derive from a user's tweets, her/his topic profile, which is based on Wikipedia categories. The accuracy of these algorithms has so far been evaluated primarily on Wikipedia articles and news datasets, which are in nature very different from the shorter messages in social media streams.

One widely used Wikipedia-based semantic annotation system is *DBPedia Spotlight* [75]. It is a freely available and customisable web-based system, which annotates text documents with DBPedia URIs. It targets the DBPedia ontology, which has more than 30 top level classes and 272 classes overall. It is possible to restrict which classes (and their sub-classes) are used for named entity recognition, either by listing them explicitly or through a SPARQL query. The algorithm first selects entity candidates through lookup against a Wikipedia-derived dictionary of URI lexicalisations, followed by a URI ranking stage using a vector space model. Each DBPedia resource is associated with a document, constructed from all paragraphs mentioning that concept in Wikipedia. The method has been shown to out-perform OpenCalais and Zemanta (see Section 4.3.2) on a small gold-standard of newspaper articles [75].

Figure 1 shows several tweets annotated with DBPedia Spotlight. The results clearly demonstrate the need for tweet spelling normalisation, as well as the difficulties Spotlight has with recognising URLs. As exemplified here, by default the algorithm is designed to maximise recall (i.e. annotate as many entities as possible, using the millions of instances from DBPedia). Given the short, noisy nature of tweets, this may lead to low accuracy results. Further formal evaluation on a shared, large dataset of short social media messages is required, in order to establish the best values for the various DBPedia Spotlight parameters (e.g. confidence, support).

The LINDEN [115] framework makes use of the richer semantic information in YAGO (semantic similarity), in addition to Wikipedia-based information (using link structure for semantic associativity). The method is heavily dependent on the Wikipedia-Miner[7] toolkit [82], which is used to analyse the context of the ambiguous entity mention and detect the Wikipedia concepts that appear there. Evaluation on the TAC-KBP2009 dataset showed LINDEN outperforming the highest ranked Wikipedia-only systems, which participated in the original TAC evaluation. Unfortunately, LINDEN has not been compared directly to DBPedia Spotlight on a shared evaluation dataset.

4.3.1. Social Media Oriented Approaches

Named entity recognition methods, which are typically trained on longer, more regular texts (e.g. news articles), have been shown to perform poorly on shorter and noisier social media content [104]. However, while each post in isolation provides insufficient linguistic context, additional information can be derived from the user profiles, social networks, and interlinked posts (e.g. replies to a tweet message). This section discusses what we call *social media oriented* semantic annotation approaches, which integrate both linguistic and social media-specific features.

Ritter *et al* [104] address the problem of named entity classification (but not disambiguation) by using Freebase as the source of large number of known entities. The straightforward entity lookup and type assignment baseline, without considering context, achieves only 38% f-score (35% of entities are ambiguous and have more than one type, whereas 30% of entities in the tweets do not appear in Freebase). NE classification performance improves to 66% through the use of labelled topic models, which take into account the context of occurrence and the distribution over Freebase types for each entity string (e.g. Amazon can be either a company or a location).

[7] http://wikipedia-miner.cms.waikato.ac.nz/

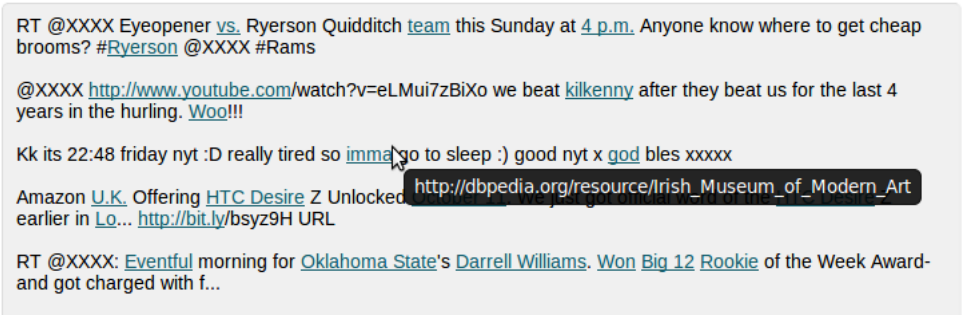

Fig. 1. DBPedia Spotlight results on tweets

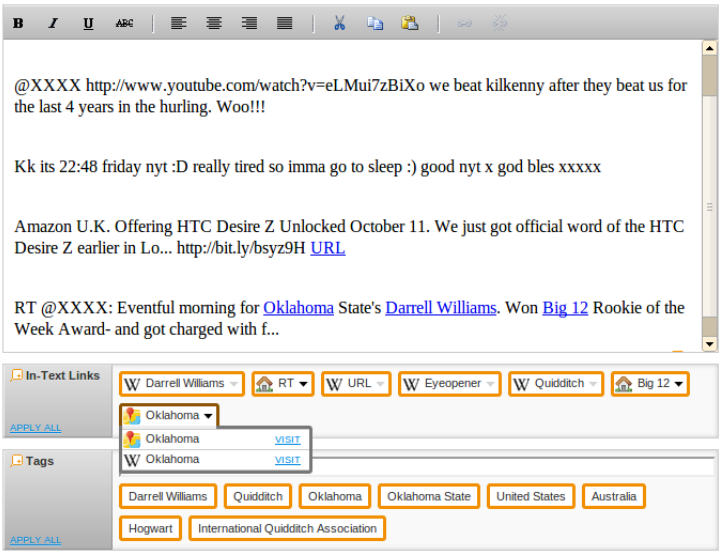

Fig. 2. Zemanta's online tagging interface

Ireson *et al* [61] study the problem of location disambiguation (toponym resolution) of name tags in Flickr. The approach is based on the Yahoo! GeoPlanet semantic database, which provides a URI for each location instance, as well as a taxonomy of related locations (e.g. neighbouring locations). The tag disambiguation approach makes use of all other tags assigned to the photo, the user context (all tags assigned by this user to all their photos), and the extended user context, which takes into account the tags of the user contacts. The use of this wider, social network-based context was shown to improve significantly the overall disambiguation accuracy.

Another source of additional, implicit semantics are hashtags in Twitter messages, which have evolved as means for users to follow conversations on a given topic. Laniado and Mika [66] investigate hashtag semantics in 369 million messages, using four metrics: frequency of use, specificity (use of the hashtag vs use of the word itself), consistency of usage, and stability over time. These measures are then used to determine which hashtags can be used as identifiers and linked to Freebase URIs (most of them are named entities). Hashtags have also been used as an additional source of semantic information about tweets, by adding textual hashtag definitions from crowdsourced online glossaries [77]. Next semantic annotation is carried out, through a simple entity lookup against DBPedia entities and categories without further disambiguation. User-related attributes and social connections are coded in FOAF, whereas semantic annotations are coded through the MOAT ontology (see Section 3).

Wikipedia-based entity linking approaches (see Section 4.3) benefit significantly from the larger linguistic context of news articles and web pages. Evalua-

tion of DBPedia Spotlight [75] and the Milne and Witten method [82] on a tweet dataset has shown significantly poorer performance [74]. Meij *et al* [74] propose a Twitter-specific approach for linking such short, noisy messages to Wikipedia articles. The first step uses n-grams to generate a list of candidate Wikipedia concepts, then supervised learning is used to classify each concept as relevant or not (given the tweet and the user who wrote it). The method uses features derived from the n-grams (e.g. number of Wikipedia articles containing this n-gram), Wikipedia article features (e.g. number of articles linking to the given page), and tweet-specific features (e.g. using hashtag definitions and linked web pages).

Gruhl *et al.* [53] focus in particular on the disambiguation element of semantic annotation and examine the problem of dealing with highly ambiguous cases, as is the case with song and music album titles. Their approach first restricts the part of the MusicBrainz ontology used for producing the candidates (in this case by filtering out all information about music artists not mentioned in the given text). Secondly, they apply shallow language processing, such as POS tagging and NP chunking, and then use this information as input to a support vector machine classifier, which disambiguates on the basis of this information. The approach was tested on a corpus of MySpace posts for three artists. While the ontology is very large (thus generating a lot of ambiguity), the texts are quite focused, which allows the system to achieve good performance. As discussed by the authors themselves, the processing of less focused texts, e.g., Twitter messages or news articles is likely to prove much more challenging.

4.3.2. Commercial Entity Recognition Services

There are a number of commercial online entity recognition services which annotate documents with entities and assign Linked Data URIs to them. The NERD online tool [105] allows their easy comparison on user-uploaded datasets. It also unifies their results and maps them to the Linking Open Data cloud. Here we focus only on the services used by research methods surveyed here (e.g. [108,2,107]).

Zemanta (http://www.zemanta.com) is an online semantic annotation tool, originally developed for blog and email content to help users insert tags and links through recommendations. Figure 2 shows an example text and the recommended tags, potential in-text link targets (e.g., the W3C Wikipedia article and the W3C home page), and other relevant articles. It is then for the user to decide which of the tags should apply and which in-text link targets they wish to add. In this example, in-text links have been added for the terms highlighted in orange, all pointing to the Wikipedia articles on the respective topics.

Open Calais is another commercial web service for semantic annotation, which has been used by some researchers on social media. For instance, Abel *et al* [2] harness OpenCalais to recognise named entities in news-related tweets[8]. The target entities are mostly locations, companies, people, addresses, contact numbers, products, movies, etc. The events and facts extracted are those involving the above entities, e.g., acquisition, alliance, company competitor. Figure 3 shows an example text annotated with some entities.

The entity annotations include URIs, which allow access via HTTP to obtain further information on that entity via Linked Data. Currently OpenCalais links to eight Linked Data sets, including its own knowledge base, DBPedia, Wikipedia, IMDB, Shopping.com. These broadly correspond to the entity types covered by the ontology.

The main limitation of Calais comes from its proprietary nature, i.e., users send documents to be annotated by the web service and receive results back, but they do not have the means to give Calais a different ontology to annotate with or to customise the way in which the entity extraction works.

4.4. Events Detection

Much as trending topics can be used to monitor global opinions and reactions, social media streams can be used as a discussion backchannel to real world events [41], and even to discover and report upon such events, almost as soon as they occur. While it may at first appear that trending topics alone are sufficient for this task, there are a few reasons that they are unsatisfactory:

- *Generality*: trending topics may discuss events, but may also refer to celebrities, products or online memes.
- *Scale*: only the topics with which a huge margin of Twitter users engage can appear as trending topics.
- *Censorship*: it is believed by many that the trending topics displayed by the official Twitter service are censored for political and language content.

[8] Unfortunately they do not evaluate the named entity recognition accuracy of OpenCalais on their dataset.

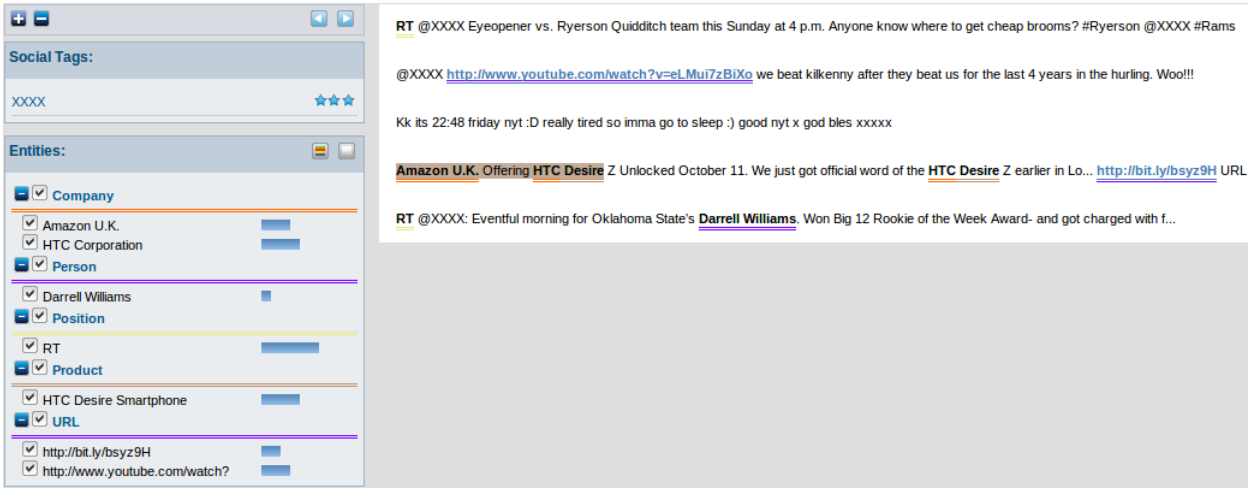

Fig. 3. Calais results on part of the Semantic Web Wikipedia entry

– *Algorithm*: the method used to select trending topics is not published anywhere and is generally not understood.

Automatic event detection therefore presents an interesting task for social media streams. While it is possible to have access to an enormous quantity of tweets, enough to reveal global trends and events, the problem of developing and evaluating scalable event detection algorithms which can handle such magnitudes of streaming text remains.

4.4.1. Event Detection through Clustering

Clustering documents together according to textual content may be useful in identifying those which belong to the same event as one another. This is the intuition behind the method of [95], who attempted to create a scalable, fast method for text clustering based on locality sensitive hashing. They argue that their algorithm is suitable for use on the vast, real time streams from Twitter's firehose API.

Though they did not address performance in the same way, [16] also perform document clustering. However, since just clustering tweets together is only helpful if all tweets are known to discuss events, they filter candidate clusters using a classifier over temporal, topical and social features [18].

Document clusters can still be somewhat difficult for humans to read, containing a great many tweets. For summarisation of these document sets, [17] evaluate methods based on centroid, degree centrality and LexRank, finding that centroid works the best in some cases, but without statistically significant differences.

Clustering and centroid similarity have also been exploited in previous efforts to detect events in full length blogs [110].

4.4.2. Event Modelling and Signal Processing

[56] produced two types of model which they believe could be used to predict relevance of a tweet within summarisation. They describe a graphical, generative model of the structure of an event within Twitter, where vertices are the mentions of an event and edges are formed by relationships between these mentions. They also model user behaviour with chains of interactions such as replies and retweets, though their experimentation reveals that the best summaries were produced using the event structures as a model of relevance.

Once an event is detected in social media streams, the next problem is how to generate useful thematic/topical descriptors for this event. Point-wise mutual information has been coupled with user geolocation and temporal information, in order to derive n-gram event descriptors from tweets. By making the algorithm sensitive to the originating location, it is possible to see what people from a given location are saying about an event (e.g. those in the US), as well as how this differs from tweets elsewhere (e.g. those from India). Similarly, the temporal information results in different text descriptors being extracted on different days, as the event unfolds.

Another class of event detection takes inspiration from signal processing, analysing tweets as sensor data. [109] used such an approach to detect earthquakes in Japan on the basis of Tweets with geolo-

cation information attached to them. Similarly, individual words have been treated as wavelet signals and analysed as such in order to discover temporally significant clusters of terms [129].

4.4.3. Detecting Sub-Events

It could be argued that many events, once successfully extracted from social media, remain too coarse to be used for informative aggregation and presentation. Although [17] investigate the effectiveness of various traditional summarisation methods when applied to clusters of tweets belonging to events, the granularity of an event is somewhat unpredictable; while the TDT initiative defines an event as "Something interesting which occurs at a specific time and place" [5], in practise this constraint is difficult to enforce.

Collections of events in a larger sequence could be referred to as sagas; they may be perfectly legitimate events in their own right, or their individual constituents might similarly be coherent on their own. Citing the example of an academic conference, [107] point out that tweets may refer to the conference as a whole, or to specific sub-events such as presentations at a specific time and place. Using semantic information about the conference event and its sub-events from the Web of Data, tweets are aligned to these sub-events automatically, using machine learning. The method includes a concept enrichment phase, which uses Zemanta to annotate each tweet with DBPedia concepts.

While an academic conference can be considered a "planned event" in that all the sub-events are known beforehand, not all events are so structured. For American football matches, [28] decompose collections of tweets using Hidden Markov Models, capturing play-by-play information such as touchdowns and penalties. Similarly, temporal peaks in tweet frequency have been used to detect important sub-events as they arise (e.g. identifying goals, penalties in football games) [70]. Frequent keywords from around the sub-event are extracted and used as short textual descriptors of each sub-event.

A semantic, entity-based approach to sub-event detection has been proposed by [34], who use manually created background knowledge about the event (e.g. team and player names for cricket games), coupled with domain-specific knowledge from Wikipedia (e.g. cricket-related sub-events like getting out). In addition to annotating the tweets with this semantic information, the method utilises tweet volume (similarly to [70]) and re-tweet frequency as sub-event indicators. The limitation of this approach, however, comes from the need for manual intervention, which is not always feasible outside of limited application domains.

4.5. Sentiment Detection and Opinion Mining

The existence and popularity of websites dedicated to reviews and feedback on products and services is something of a homage to the human urge to post what they feel and think online. When the most common type of message on Twitter is about 'me now' [84], it is to be expected that users talk often about their own moods and opinions. Bollen *et al* [21] argue that users express both their own mood in tweets about themselves and more generally in messages about other subjects. Another study [62] estimates that 19% of microblog messages mention a brand and from those that do, around 20% contain brand sentiment.

The potential value of these thoughts and opinions is enormous. For instance, mass analysis could provide a clear picture of overall mood, exploring reactions to ongoing public events [21] or feedback to a particular individual, government, product or service [65]. The resulting information could be used to improve services, shape public policy or make a profit on the stock market.

The user activities on social networking sites are often triggered by specific events and related entities (e.g. sports events, celebrations, crises, news articles, persons, locations) and topics (e.g. global warming, financial crisis, swine flu). In order to include this information, semantically- and social network-aware approaches are needed.

There are many challenges inherent in applying typical opinion mining and sentiment analysis techniques to social media. Microposts are, arguably, the most challenging text type for opinion mining, since they do not contain much contextual information and assume much implicit knowledge. Ambiguity is a particular problem since we cannot easily make use of coreference information: unlike in blog posts and comments, tweets do not typically follow a conversation thread, and appear much more in isolation from other tweets. They also exhibit much more language variation, tend to be less grammatical than longer posts, contain unorthodox capitalisation, and make frequent use of emoticons, abbreviations and hashtags, which can form an important part of the meaning. Typically, they also contain extensive use of irony and sarcasm, which are particularly difficult for a machine to detect. On the other hand, their terseness can also be beneficial in focusing the topics more explicitly: it is very

rare for a single tweet to be related to more than one topic, which can thus aid disambiguation by emphasising situational relatedness.

[90] present a wide-ranging and detailed review of traditional automatic sentiment detection techniques, including many sub-components, which we shall not repeat here. In general, sentiment detection techniques can be roughly divided into lexicon-based methods (e.g. [112,121]) and machine-learning methods, e.g. [20]. Lexicon-based methods rely on a sentiment lexicon, a collection of known and pre-compiled sentiment terms. Machine learning approaches make use of syntactic and/or linguistic features [89,50], and hybrid approaches are very common, with sentiment lexicons playing a key role in the majority of methods, e.g. [38].

4.5.1. Polarity classification

One of the most common tasks undertaken when attempting to detect sentiment is that of polarity classification. This is so common that the term 'Sentiment classification' is sometimes used to refer to the detection of sentiment polarity, even though it could equally refer to subjectivity classification [90]. The problem is one of assigning to a given document or textual unit either 'positive' or 'negative' sentiment, or some value on a scale between the two.

In the context of a product review, or a post about ones own life, a positive sentiment will usually refer to an endorsement of the subject or a positive feeling about events. However, [90] argue that in other problems the polarity can be more abstract, giving the example of classifying good news from bad.

[89] aimed to classify arbitrary tweets on the basis of positive, negative and neutral sentiment (here neutral presumably means that the author has no opinion to express). They constructed a simple binary classifier which used n-gram and POS features, and trained it on instances which had been annotated according to the existence of positive (':)') and negative (':(') emoticons. Their approach has a lot in common with an earlier sentiment classifier constructed by [50], which also used unigrams, bigrams and POS tags, though the former demonstrated through analysis that the distribution of certain POS tags varies between positive and negative posts. The use of such shallow linguistic information leads to a data sparsity problem. Saif *et al* [108] demonstrate that by using semantic concepts, instead of entities such as iPhone, polarity classification is improved. The approach uses AlchemyAPI for the semantic annotation and the results are evaluated on the Stanford Twitter Sentiment Dataset[9].

[65] tackles a somewhat different sentiment analysis task. Tweets relating to president Obama are analysed and a daily overall "strong sentiment" is calculated. This figure is given as the ratio of the count strongly positive tweets over the strongly negative ones. The strength and polarity of Tweets in the dataset is calculated according to learned lexicons, which are lists of keywords which in general correspond to either positive or negative sentiment.

[38] also made use of a sentiment lexicon to initially annotate positive and negative sentiment in tweets related to political events. They performed supervised learning with manually annotated examples to train a binary classifier of political opinion, using this second classifier when the former failed to make a classification. They only report the overall sentiment from a collection of Tweets during a specific time-window, and their system will refrain from reporting sentiment when no consensus appears to be reached for that period.

One problem faced by many search based approaches to sentiment analysis is that the topic of the retrieved document is not necessarily the object of the sentiment held therein. Tweets retrieved using a keyword search like that of [65] may contain multiple distinct sentiments about different objects, or they might not really be on-topic at all. One possible approach is to use semantic annotation to discover entities mentioned in the Tweets themselves. [71] identify people, opinions and political parties in tweets using rule-based grammars, and analyse these patterns to generate triples representing opinions and voter intentions. They still capture positive and negative sentiments, but in a more precise way which sacrifices some recall.

The simpler bag-of-words sentiment classifiers have the weakness that they do not handle negation well; the difference between the phrases 'not good' and 'good' is somewhat ignored in a unigram model, though they carry completely different meanings. A possible solution is to incorporate longer range features such as higher order n-grams or dependency structures, which would help capture more complete, subtle patterns, such as in the sentence "Surprisingly, the build quality is well above par, considering the rest of the features" in which the term 'surprisingly' should partially negate the positive overall sentiment [90].

[9] http://twittersentiment.appspot.com/

Another way to deal with negation, avoiding the need for dependency parsing, is to capture simple patterns such as 'isn't helpful' or 'not exciting' by inserting unigrams like 'NOT-helpful' and 'NOT-exciting' respectively [35]. This work-around was implemented for tweets by [88]. [71] made use of similar simple patterns to negate their extracted sentiment judgements.

Once sentiment has been successfully discovered for each tweet in a dataset, there remains the secondary task of deciding out to display the resulting annotations. While it is possible to simply show these per-tweet classifications in some existing order, overall sentiment can also be aggregated, using graphical displays such as pie-charts or coloured time-lines [38,89]. More complex analysis of aggregate sentiment can enable novel uses of Twitter, such as predicting future stock performance [22] or election results [86].

4.5.2. Identifying Subjectivity

A common problem in classical sentiment analysis is that of classifying the subjectivity of a posting or other unit of text. In a dataset of tweets it is likely that many will contain no sentiment whatsoever (for example those which are simply updates or those which are intended only to maintain the author's presence). The task of opinion or subjectivity detection is quite distinct to that of discovering polarity, in that it is concerned solely with whether or not one exists.

To some extent this problem has been ignored in the literature on sentiment detection from social media streams, though [65] attempt to use the same measure required for identifying polarity to also characterise the strength of a sentiment, and [127] train a classifier of subjectivity as a way to assist in the broader task of discovering 'situational awareness' tweets in an emergency.

Many others appear to either rely on datasets of tweets which can be confidently labelled as containing sentiment thanks to the existence of emoticons or other typographical features [50,51], or they collect tweets according to a specific topic such as politics or films and simply assumed to be suitable for sentiment detection [11]. Other approaches train polarity classifiers which are able to choose not to classify in some instances [38], or which are able to produce a 'neutral' classification [88].

It is important to note that although subjectivity detection can be a precursor to polarity identification, the two are in some sense separate tasks and subjectivity detection can be useful as a component in other systems. The level of subjectivity within a tweet may go some way towards classifying its intent; an intuition exploited by [127] for Tweets which were intended to provide information to others and aid the emergency services.

4.5.3. Beyond Subjectivity and Polarity

The work on sentiment detection we have discussed so far has been in the problem area of the classification of subjectivity and polarity. However, sentiment analysis is far more general, covering many other kinds of non-factual data. One particularly challenging task is the detection of sarcasm on Twitter [125], though other areas of investigation include political attitudes and general mood.

Sentiment analysis can be used to characterise and compare emotional state, such as in the Twitter timelines of political figures [126]. [21] automatically summarise of the posts of all users of Twitter on a given day, according to a Profile Of Mood State (POMS) which considers emotions such as tension, depression and anger. Two case studies are shown in which the authors believe these overall POMS scores reflect changes in public mood caused by global factors such as political elections or public holidays. Mood profiling in this way has also been attempted with other kinds of text [68,6].

Sentiment summarisation is the problem of presenting a concise view of user sentiment on a given topic/entity. Polarity-based sentiment lends itself trivially to summarisation, typically expressed as percentages (e.g. 63% positive vs 37% negative opinions) and visualised as bar charts, pie charts or colour coding (see Section 6.3). However, while showing the overall sentiment, such summaries are of fairly limited utility in cases when users wish to see more concrete details, e.g. what are the key complaints made in these negative opinions. Concise pros and cons summaries, including lists of key opinion points made (e.g. short battery life), have been studied in the context of product reviews [48]. The approach generates n-grams, starting from some high frequency words derived from across all reviews. Each candidate n-gram is tested for representativeness (does it cover the major opinions from the original texts) and readability (be grammatical), in order to generate a non-redundant list of such phrases. Although the algorithm has so far been tested only on product reviews, it should be well suited also to Twitter and Facebook messages.

4.6. Cross-Media Linking

The short nature of Twitter and Facebook messages, coupled with their frequent grounding in real world events, means that often short posts cannot be understood without reference to external context. While some posts already contain URLs, the majority do not. Therefore automatic methods for cross-media linking and enrichment are required.

Abel *et al* [2] link tweets to current news stories in order to improve the accuracy of semantic annotation of tweets. Several linkage strategies are explored: utilising URLs contained in the tweet, TF-IDF similarity between tweet and news article, hashtags, and entity-based similarity (semantic entities and topics are recognised by OpenCalais), with the entity-based one being the best one for tweets without URLs. The approach bears similarities with the keyphrase-based linking strategy for aligning news video segments with online news pages [42]. [60] go one step further by aggregating social media content on climate change from Twitter, YouTube, and Facebook with online news, although details of the cross-media linking algorithm are not supplied in the paper.

An in-depth study comparing Twitter and New York Times news [135] has identified three types of topics: event-oriented, entity-oriented, and long-standing topics. Topics are also classified into categories, based on their subject area. Nine of the categories are those used by NYT (e.g. arts, world, business) plus two Twitter-specific ones (Family&Life and Twitter). Family&Life is the most predominant category on Twitter (called 'me now' by [84]), both in terms of number of tweets and number of users. Automatic topic-based comparison showed that tweets abound with entity- oriented topics, which are much less covered by traditional news media.

Going beyond news and tweets, future research on cross-media linking is required. For instance, some users push their tweets into their Facebook profiles, where they attract comments, separate from any tweet replies and retweets. Similarly, comments within a blog page could be aggregated with tweets discussing it, in order to get a more complete overall view.

4.7. Discussion

Even though some inroads have been made already, current methods for semantic annotation of social media streams have many limitations. Firstly, most methods address the more shallow problems of keyword and topic extraction, while ontology-based entity and event recognition do not reach the significantly higher precision and recall results obtained on longer text documents. One way to improve the currently poor automatic performance is through crowdsourcing. The ZenCrowd system [37], for instance, combines algorithms for large-scale entity linking with human input through micro-tasks on Amazon Mechanical Turk. In this way, textual mentions that can be linked automatically and with high confidence to instances in the LOD cloud, are not shown to the human annotators. The latter are only consulted on hard to solve cases, which not only significantly improves the quality of the results, but also limits the amount of manual intervention required.

Another way to improve semantic annotation of social media is to make better use of the vast knowledge available on the Web of Data. Currently this is limited mostly to Wikipedia and resources derived from it (e.g. DBPedia and YAGO). One of the challenges here is ambiguity. For instance, song and album titles in MusicBrainz are highly ambiguous and include common words (e.g. Yesterday), as well as stop words (The, If) [53]. Consequently, an automatic domain categorisation step might be required, in order to ensure that domain-specific LOD resources, such as MusicBrainz, are used to annotate only social media content from the corresponding domain. The other major challenges are robustness and scalability. Firstly, the semantic annotation algorithms need to be robust in the face of noisy knowledge in the LOD resources, as well as being robust with respect to dealing with the noisy, syntactically irregular language of social media. Secondly, given the size of the Web of Data, designing ontology-based algorithms which can load and query efficiently these large knowledge bases, while maintaining high computational throughput is far from trivial.

The last obstacle to making better use of Web of Data resources, lies in the fairly limited lexical information. With the exception of resources grounded in Wikipedia, lexical information in the rest is mostly limited to RDF labels. This in turn limits their usefulness as a knowledge source for ontology-based information extraction and semantic annotation. Recent work on linguistically grounded ontologies [24] has recognised this shortcoming and proposed a more expressive model for associating linguistic information to ontology elements. While this is a step in the right direction, nevertheless further work is still required, especially with respect to building multilingual semantic annotation systems.

In addition, it is axiomatic that semantic annotation methods are only as good as their training and evaluation data. Algorithm training on social media gold standard datasets is currently very limited. For example, there are currently fewer than 10,000 tweets annotated with named entity types and events. Bigger, shared evaluation corpora from different social media genres are therefore badly needed. Creating these through traditional manual text annotation methodologies is unaffordable, if a significant mass is to be reached. Research on crowdsourcing evaluation gold standards has been limited, primarily with focus on using Amazon Mechanical Turk to acquire small datasets (e.g. tweets with named entity types) [47]. We will revisit this challenge again in Section 7.

In the area of sentiment analysis, researchers have investigated the problems of sentiment polarity detection, subjectivity classification, prediction through social media and user mood profiling. Some of the efforts discussed here could potentially prove useful as components in actual applications, while others are interesting in their own right.

Topical information retrieval, especially for Twitter, is still in its infancy, and keyword search can be imprecise. Little effort has yet been invested in the problems of automatically detection that tweets are on topic, actually contain a sentiment and address that the sentiment addresses the expected object.

Moreover, evaluation of opinion mining is particularly difficult for a number of methodological reasons (in addition to the lack of shared evaluation resources discussed above). First, opinions are often subjective, and it is not always clear what was intended by the author. For example, a person cannot necessarily tell if a comment such as "I love Baroness Warsi", in the absence of further context, expresses a genuine positive sentiment or is being used sarcastically. Inter-annotator agreement performed on manually annotated data therefore tends to be low, which affects the reliability of any gold standard data produced.

Lastly, social media streams impose a number of further outstanding challenges on opinion and sentiment mining methods:

- *Relevance*: In social media discussions and comment threads can rapidly diverge into unrelated topics, as opposed to product reviews which rarely stray from the topic at hand.
- *Target identification*: There is often a mismatch between the topic of the social media post, which is not necessarily the object of the sentiment held therein. For example, the day after Whitney Houston's death, TwitterSentiment and similar sites all showed an overwhelming majority of tweets about Whitney Houston to be negative; however, almost all these tweets were negative only in that people were sad about her death, and not because they disliked her.
- *Volatility over time*: More specifically, opinions can change radically over time, from positive to negative and vice versa. To address this problem, the different types of possible opinions can be associated as ontological properties with the classes describing entities, facts and events, discovered through semantic annotation techniques, similar to those in [72] which aimed at managing the evolution of entities over time. The extracted opinions and sentiments can be time-stamped and stored in a knowledge base, which is enriched continuously, as new content and opinions come in. A particularly challenging question is how to detect emerging new opinions, rather than adding the new information to an existing opinion for the given entity. Contradictions and changes also need to be captured and used to track trends over time, in particular through opinion aggregation.
- *Opinion aggregation*: Another challenge is the type of aggregation that can be applied to opinions. In entity-based semantic annotation, this can be applied to the extracted information in a straightforward way: data can be merged if there are no inconsistencies, e.g. on the properties of an entity. Opinions behave differently here, however: multiple opinions can be attached to an entity and need to be modelled separately, for which we advocate populating a knowledge base. An important question is whether one should just store the mean of opinions detected within a specific interval of time (as current opinion visualisation methods do), or if more detailed approaches are preferable, such as modelling the sources and strength of conflicting opinions and how they change over time. A second important question in this context involves finding clusterings of the opinions expressed in social media, according to influential groups, demographics and geographical and social cliques. Consequently, the social, graph-based nature of the interactions requires new methods for opinion aggregation.

However, even though state-of-the-art methods have a large scope for improvement, semantic annotation re-

sults are already being used by methods that derive automatically models of users and social networks, from the information implicit in social media streams. This is where we turn to next.

5. Semantic-Based User Modelling

A *User Model* (UM) is a knowledge resource containing *explicit* semantic information about various aspects of the user, which the system has *a priori* (e.g. by importing a Facebook profile) or has inferred them from user behaviour, user-generated content, social networks or other sources. Some important characteristics of user models are:

- UM is a *distinct knowledge resource* within the overall system;
- semantic information is represented *explicitly*. Implicit information disclosed in social media is used to derive this explicit knowledge.
- *abstraction*, i.e., representation of types of users, roles and groups, as well as of individual users.
- *multi-purpose* – the semantically encoded user model can be used in different ways, e.g. personalised content recommendation, filtering.
- *reasoning* – the representation should allow for reasoning *about* the knowledge, as well as reasoning *with* it.
- *interconnected* – a user model is more than a collection of attributes. Usually, there are also complex relations between them, as well as relations to other types of knowledge (e.g. posts made by the user).

Ontology-based user models have been used extensively on content other than social media streams, especially in the context of Personal Information Management (PIM). PIM work originated in research on the social semantic desktop [36], where information from the user's computer (e.g. email, documents) is used to derive models of the user. For a detailed overview of user modelling for the semantic web see [10].

In this paper, we focus on the extension of this work towards social media streams, as well as mention sensor-based information where relevant (e.g. GPS coordinates in tweets). As discussed in Section 2, the social and user-generated nature of these streams make it possible to derive rich semantic user models. More specifically, we examine the application of semantic annotation for user model construction. Consequently, we consider outside the scope of this paper research, which is focused purely on social network analysis (e.g. [81]) and/or uses purely quantitative user and post characteristics (e.g. number of threads/posts, number of replies/re-tweets [106]) and/or post metadata only (e.g. the re-tweet and in-reply-to JSON fields).

5.1. Constructing Social Semantic User Models from Semantic Annotations

Among the various kinds of social media, folksonomies have probably received most attention from researchers studying how semantic models of user interactions and interests can be derived from user-generated content. Many approaches focused on exploring the social and interaction graphs, using techniques from social network analysis (e.g. [81]). In this section, however, we are concerned with methods that discover and exploit the semantics of textual tags instead. This section also includes semantic-based user modelling research on online forums, blogs, and Twitter.

Based on the kinds of semantic information used, methods can be classified as follows:

- Bag of words ([31]);
- Semantically disambiguated entities: mentioned by user (e.g. [2,63]) or from a linked longer Web document (e.g. [2]);
- Topics: Wikipedia categories (e.g. [2,119]), latent topics (e.g. [134]), or tag hierarchies (e.g. [136]).

This is typically supplemented with more quantitative social network information (e.g. how many connections/followers does a user have [7]) and interaction information (e.g. post frequency [106], average number of posts per thread [7]).

The rest of the section discusses in more detail the kinds of user information that has been extracted from semantically annotated social media and concludes with a discussion of open issues.

5.1.1. Discovering User Demographics

Every Twitter user has a profile which reveals some details of their identity. The profile is semi-structured, including a textual bio field, a full name, the user's location, a profile picture, a time zone and a homepage URL (most of these are optional and often empty). The user's attributes can be related to the content of their posts, for example their physical location can determine to a degree the language they use [33] or the events on which they comment [132].

There have been efforts to discover user demographics information, when it is not available in the fields in their profile. [25] classify users as male or female based on the text of their tweets, their description fields and their names. They report better-than-human accuracy, compared to a set of annotators on Mechanical Turk. [93] present a general framework for user classification which can learn to automatically discover political alignment, ethnicity and fans of a particular business.

Twitter users may share the location from which they are tweeting by posting from a mobile device and allowing it to attach a reading from its GPS receiver to the message or by setting their own location in a field of their profile. However, [33] found that only around 36% of users actually filled in their location field in their profile with a valid location as specific as their nearest city. Furthermore, when we analysed a dataset of over 30,000 tweets discussing the 2011 London Riots, less than 1% of messages contained any GPS information.

There have been attempts to automatically locate Twitter users. [33] exploit the content of user tweets, discovering phrases and terms that seem to be restricted to a single area and using those to identify location. Additionally, [44] made use of a generative model of tweet content, assuming that the content of tweets is in some way produced according to a random process distorted by a distribution conditioned on where they live.

5.1.2. Deriving User Interests from Semantic Annotations

Abel *et al* [2] propose simple entity-based and topic-based user profiles, built from the user's tweets. The entity-based profile for a given user is modelled as a set of weighted entities, where the weight each entity *e* is computed based either on the number of user tweets that mention *e*, or based on frequency of entity occurrences in the tweets, combined with the related news articles (which are identified in an earlier, linking step). Topic-based profiles are defined in a similar fashion, but represent higher level Wikipedia categories (e.g. sports, politics). Both entities and topics are identified using OpenCalais (see Section 4.3.2). Abel *et al* have also demonstrated that hashtags are not a useful indicator of user interests – a finding which is also supported by [79]. A major limitation of the method is that it depends heavily on the news linking, which the authors have shown applies successfully to only 15% of tweets.

In a subsequent paper [3], Abel *et al* refine their approach to modelling user interests in a topic, to take also into account re-tweets, as well as changes over time (when do users become interested in a topic, for how long, and which concepts are relevant to which topic). Evaluation is based on global topics (i.e. the Egyptian revolution). Their findings demonstrate that a time-dependent topic weighting function produces user interest models, which are better for tweet recommendation purposes. They also identify different groups of users, based on the duration of their interest in a given topic: *long-term adopters* who join early for longer vs. *short-term adopters* who join global discussions later and are influenced by public trends.

Kapanipathi *et al* [63] similarly use semantic annotations to derive user interests (entities or concepts from DBPedia), weighted by strength (calculated on the basis of frequency of occurrence). They also demonstrate how interests can be merged based on information from different social media (LinkedIn, Facebook and Twitter). Facebook likes and explicitly stated interests in LinkedIn and Facebook are combined with the implicit interest information from the tweets. The Open Provenance Model[10] is used to keep track of interest provenance.

A similar entity- and topic-based approach to modelling user interests is proposed by Michelson and Macskassy [79] (called Twopics). All capitalised, non-stop words in a tweet are considered as entity candidates and looked up against Wikipedia (page titles and article content). A disambiguation step then identifies the Wikipedia entity which matches best the candidate entity from the tweet, given the tweet content as context. For each disambiguated entity, the sub-tree of Wikipedia categories is obtained. In a subsequent, topic-assignment step, all category sub-trees are analysed to discover the most frequently occurring categories, which are then assigned as user interests in the topic-based profile. The authors also argue that such more generic topics, generated by leveraging the Wikipedia category taxonomy, are more appropriate for clustering and searching for users, than the term-based topic models derived using bag-of-words or LDA methods.

[134] propose a probabilistic method for identifying user interests from folksonomy tags (Del.icio.us). The first step is to induce hierarchies of latent topics from a set of tags in an unsupervised manner. This approach,

[10] http://openprovenance.org

based on Hierarchical Dirichlet Process, models topics as probability distributions over the tag space, rather than clustering the tags themselves [136]. Next, user interest hierarchies are induced via log-likelihood and hierarchy comparison methods. Zavitsanos *et al* however stop short of assigning explicit semantics to the topics through URIs.

In order to ground user interest models semantically, researchers have used Wikipedia as a multi-domain model [119]. They also propose a method for consolidation of user profiles across social networking sites. Tags from different sites are filtered based on WordNet synonymy and correlated to Wikipedia pages. Subsequently, Wikipedia categories are used, in order to select representative higher-level topics of interest for the user. The approach is very similar to Twopics [79].

5.1.3. Capturing User Behaviour

Categorising user behaviour is key to understanding interactions in social media. Here we focus primarily on approaches which utilise automatic semantic annotation, in order to classify users into roles.

In the case of online forums, the following user behaviour roles have been identified [29]: *elitist*, *grunt*, *joining conversationalist*, *popular initiator*, *popular participant*, *supporter*, *taciturn*, and *ignored*. In Twitter, the most common role distinction is between *meformers* (80% of users) and *informers* (20% of users) [84].

In order to assign behaviour roles in online forums automatically, Angeletou *et al* [7] create skeleton rules in SPARQL, that map semantic features of user interaction to a level of behaviour (high, medium, and low). These levels are constructed dynamically from user exchanges and can be altered over time, as the communities evolve. User roles, contexts, and interactions are modelled semantically through the User Behaviour Ontology (see Section 3) and are used ultimately to predict the health of a given online forum.

The problem of characterising Twitter user behaviour, based on the content of their posts has yet to be fully explored. [131] generated keyphrases for users with the aid of topic modelling and a PageRank method. Similarly, [130] use a combination of POS filtering and TextRank to discover tags for users. It should also be noted that while [84] went some way towards categorising user behaviour and tweet intention, their method is not automatic and it remains unclear whether or not similar categories could be assigned by a classifier.

5.2. Discussion

As demonstrated by our survey, a key research challenge for semantic user modelling lies in addressing the diverse, dynamic, temporal nature of user behaviour. An essential part of that is the ability to represent and reason with conflicting personal views, as well as to model change in user behaviour, interests, and knowledge over time. For instance, in the context of blogs, Cheng *et al* [32] have proposed an interest forgetting function for short-term and long-term interest modelling. Anegeletou *et al* [7] recently developed time-contextualised models of user behaviour and demonstrated how these could be used to predict changes in user participation in online forums.

With respect to capturing user interests from tweets, further work is required on distinguishing globally interesting topics (e.g. trending news) from interests specific to the given user (e.g. work-related, hobby, gossip from a friend, etc.). What is interesting to a user also ties in with user behaviour roles (see Section 5.1.3). In turn, this requires more sophisticated methods for automatic assignment of user roles, based on the semantics of posts, in addition to the current methods based primarily on quantitative interaction patterns.

Since many users now participate in more than one social network, the issue of merging user modelling information across different sources arises, coupled with the challenge of modelling and making use of provenance information. [63] have carried out some preliminary work, but more sophisticated models, as well as detailed quantitative and user-based evaluations, are still required.

Lastly, another challenging question is how to go beyond interest-based models and interaction-based social networks. For instance, Gentile *et al* [49] have demonstrated how people's expertise could be captured from their email exchanges and used to build dynamic user profiles. These are then compared to each other, in order to derive automatically an expertise-based user network, rather than one based on social interactions. Such approach could be extended and adapted to blogs (e.g. for discovery and recommendation of blogs), as well as to information sharing posts in Twitter and LinkedIn streams.

6. Semantic-based Information Access over Media Streams

Semantic annotations enable users to find documents that mention one or more concepts from the on-

tology and, optionally, their relations. Depending on the methods used, search queries can often mix free-text keywords with restrictions over semantic annotations. Search tools often provide also browsing functionality, as well as search refinement capabilities. Due to the fact that social media streams are high volume and change over time, semantic search and browsing is a very challenging task.

In general, semantic-based search and retrieval over social media streams differ from traditional information retrieval, due to the additionally available ontological knowledge. On the other hand, they also differ from semantic web search engines, such as Swoogle [40], due to their focus on semantic annotations and using those to retrieve documents, rather than forming queries against ontologies to obtain sets of machine-readable triples.

This section discusses methods specifically developed for social media streams.

6.1. Semantic Search over Social Media Streams

Searching social media streams differs significantly from web searches [123] in a number of important ways. Firstly, users search message streams, such as Twitter, for temporally relevant information and are mostly interested in people. Secondly, searches are used to monitor Twitter content over time and can be saved as part of user profiles. Thirdly, Twitter search queries are significantly shorter and results include more social chatter, whereas web searches look for facts. Coupled with the short message length, noisy nature, and additional information hidden in URLs and hashtags, these differences make traditional keyword-based search methods sub-optimal on media streams. Here we focus on recent work on semantic search, addressing these challenges.

The TREC 2011 Microblog track[11] has given impetus to research by providing a set of query topics, a time point, and a corpus of 16 million tweets, a subset of which was hand-annotated for relevance as a gold standard. In addition to the widely used keyword-based and tweet syntax features (e.g. whether it contains a hashtag), Tao *et al* [122] experimented with entity-based semantic features produced by DBPedia Spotlight, which provided significantly better results.

The Twarql system [76] generates RDF triples from tweets, based on metadata from the tweets themselves, as well as entity mentions, hashtags, and URLs [77]. These are encoded using standard Open Data vocabularies (FOAF, SIOC) (see Section 3) and can be searched through SPARQL queries. It is also possible to subscribe to a stream of tweets matching a complex semantic query, e.g. what competitors are mentioned with my product (Apple iPad in their use case). At the time of writing, Twarql has not been evaluated formally, so its effectiveness and accuracy are yet to be established.

Abel *et al* propose an adaptive faceted search framework for social media streams [1]. It uses semantic entity annotations by OpenCalais, coupled with a user model (see Section 5.1.2), in order to create and rank facets semantically. Keyword search and hashtag-based facets are used as the two baselines. The best results are achieved when facets are personalised, i.e. ranked according to which entities are interesting for the given user (as coded in their entity-based user model). Facet ranking also needs to be made sensitive to the temporal context (essentially the difference between query time and post timestamp).

6.2. Filtering and Recommendations for Social Media Streams

The unprecedented rise in the volume and perceived importance of social media content has resulted in individuals starting to experience information overload. In the context of Internet use, research on information overload has shown already that high levels of information can lead to ineffectiveness, as "a person cannot process all communication and informational inputs" [15]. Consequently, researchers are studying information filtering and content recommendation, in order to help alleviate information overload, arising from social media streams. Since Facebook streams are predominantly private, the bulk of work has so far focused on Twitter.

As discussed in [31], social media streams are particularly challenging for recommender methods, and different from other types of documents/web content. Firstly, relevance is tightly correlated with recency, i.e. content stops being interesting after just a few days. Secondly, users are active consumers and generators of social content, as well as being highly connected with each other. Thirdly, recommenders need to strike a balance between filtering out noise and supporting serendipity/knowledge discovery. Lastly, interests and preferences vary significantly from user to user, depending on the volume of their personal stream; what

[11] http://sites.google.com/site/trecmicroblogtrack/

and how they use social media for (see Section 5.1.3 on user roles); and user context (e.g. mobile vs tablet, work vs home).

Chen *et al* [31] and Abel *et al* [3] focused on recommending URLs to Twitter users, since it is a common information sharing task. The approach of Chen *et al* is based on a bag-of-words model of user interests, based on the user tweets, what is trending globally, and the user's social network. URL topics are modelled similarly as a word vector and tweet recommendations are computed using cosine similarity.

Abel *et al* [3] improve on this approach by deriving semantic-based user interest models (see Section 5.1.2), which are richer and more generic. They also capture more information through hashtag semantics, replies, and, crucially, by modelling temporal dynamics of user interests.

Recently, Chen *et al* [30] extended their work towards recommending interesting conversations, i.e. threads of multiple messages. The rationale comes from the widespread use of Facebook and Twitter for social conversations [84], coupled with the difficulties that users experience with following these conversations over time, in Twitter in particular. Conversations are rated based on thread length, topic (using bag-of-words as above) and tie-strength (higher priority for content from tightly connected users). Tie strength is modelled for bi-directionally connected users only, using the existence of direct communication, its frequency, and the tie strengths of their mutual friends. Results showed that different recommendation strategies are appropriate for different types of Twitter users, i.e. those who use it for social purposes prefer conversations from closely tied friends, whereas for information seekers, the social connections are much less important.

In the context of Facebook, researchers from Microsoft [87] have trained SVM classifiers to predict, for a given user, the importance of Facebook posts within their news feed, as well as the overall importance of their friends. They also demonstrate a correlation between the two, i.e. the overall importance of a friend influences significantly the importance of posts. In terms of semantic information, the method utilises the Linguistic Inquiry and Word Count (LIWC) dictionary and its 80 topic categories [94]. One of the key findings was the empirical validation for the need of filtering and recommendation of user posts, going beyond reverse chronological order. A second very important, but less strongly substantiated, finding is the need for personalisation (i.e. the same post could be very important for one user, while marked as non-relevant by another).

The issue has recently been recognised by Facebook, who have started to filter the posts shown in the user's news feed, according to the system's proprietary EdgeRank model of importance [64]. EdgeRank takes into account the tie strength (affinity) between the posting user and the viewing user, the type of post (comment, like, etc), and a time decay factor. However, the full details of the algorithm are currently unknown, neither is its evaluation. Anecdotally, in 2010 50% of all users were still clicking on the reverse chronological timeline of their feeds. This feature has since been removed and the EdgeRank algorithm refined further. However, it is still not yet possible for the users themselves to train the system, by marking explicitly which posts they consider important.

6.3. Stream Browsing and Visualisation

The main challenge in browsing and visualisation of high-volume stream media is in providing a suitably aggregated, high-level overview. Timestamp-based list interfaces that show the entire, continuously updating stream (e.g. the Twitter timeline-based web interface) are often impractical, especially for analysing high-volume, bursty events. For instance, during the royal wedding in 2011, tweets during the event exceeded 1 million. Similarly, monitoring long running events, such as presidential election campaigns, across different media and geographical locations is equally complex.

One of the simplest and widely used visualisations is word clouds. These generally use single word terms, which can be somewhat difficult to interpret without extra context. Word clouds have been used to assist users in browsing social media streams, including blog content [13] and tweets [113,85]. For instance, Phelan *et al* [96] use word clouds to present the results of a Twitter based recommendation system. The Eddi system [19] uses topic clouds, showing higher-level themes in the user's tweet stream. These are combined with topic lists, which show who tweeted on which topic, as well as a set of interesting tweets for the highest ranked topics. The Twitris system (see Figure 4) derives even more detailed, contextualised phrases, by using 3-grams, instead of uni-grams [85]. More recently, the concept has been extended towards image clouds [41].

The main drawback of cloud-based visualisations is their static nature. Therefore, they are often com-

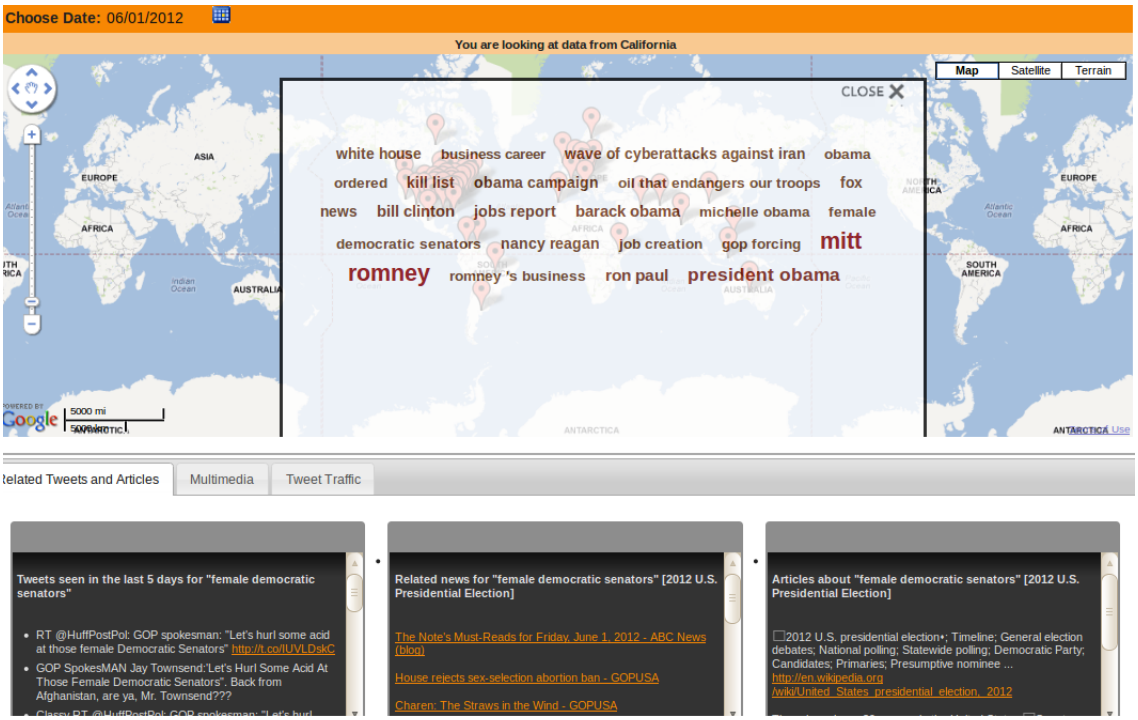

Fig. 4. The Twitris Social Media Event Monitoring Portal (http://twitris.knoesis.org)

bined with timelines showing keyword/topic frequencies over time [4,19,60,129], as well as methods for discovery of unusual popularity bursts [13]. [38] use a timeline which is synchronised with a transcript of a political broadcast, allowing navigation to key points in a video of the event, and displaying tweets from that time period. Overall sentiment is shown on a timeline at each point in the video, using simple colour segments. Similarly, TwitInfo (see Figure 6 [70]) uses a timeline to display tweet activity during a real-world event (e.g. a football game), coupled with some example tweets, colour-coded for sentiment. Some of these visualisations are dynamic, i.e. update as new content comes in (e.g. topic streams [41], falling keyword bars [60] and dynamic information landscapes [60]).

Fig. 7. Different Topics Extracted by Twitris for Great Britain

In addition, some visualisations try to capture the semantic relatedness between topics in the media streams. For instance, BlogScope [13] calculates keyword correlations, by approximating mutual information for a pair of keywords using a random sample of documents. Another example is the information landscape visualisation, which convey topic similarity through spatial proximity [60] (see Figure 5). Topic-document relationships can be shown also through force-directed, graph-based visualisations [43]. Lastly, Archambault *et al* [8] propose multi-level tag clouds, in order to capture hierarchical relations.

Another important dimension of user-generated content is its place of origin. For instance, some tweets are geo-tagged with latitude/longitude information, while many user profiles on Facebook, Twitter, and blogs specify a user location. Consequently, map-based visualisations of topics have also been explored [78,70,60,85] (see also Figures 5 and 6). For instance, Twitris [85] allows users to select a particular state from the Google map and it shows the topics discussed in social media from this state only. Figure 4 shows the Twitris US 2012 Presidential elections monitor, where we have chosen to see the related topics discussed in social media originating from California. Clicking on the topic "female democratic senators" displays

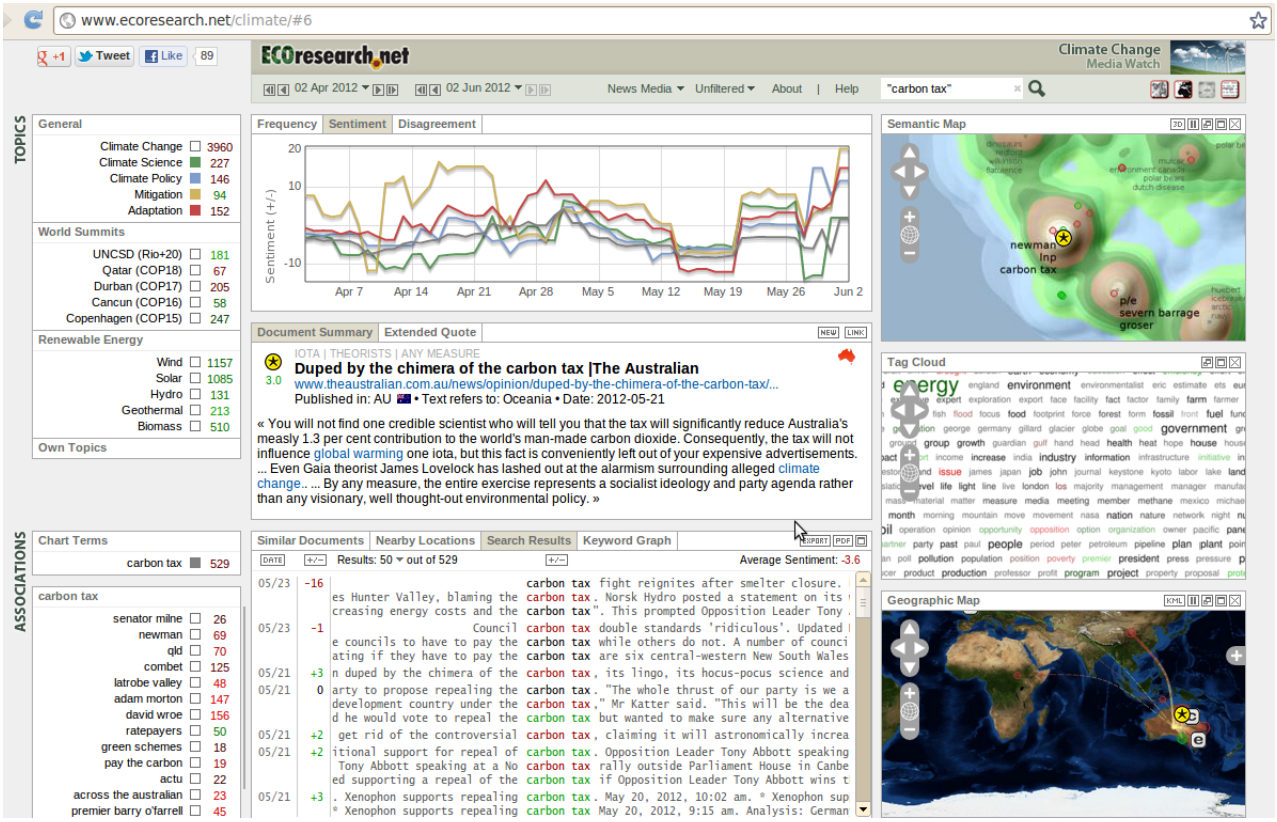

Fig. 5. Media Watch on Climate Change Portal (http://www.ecoresearch.net/climate)

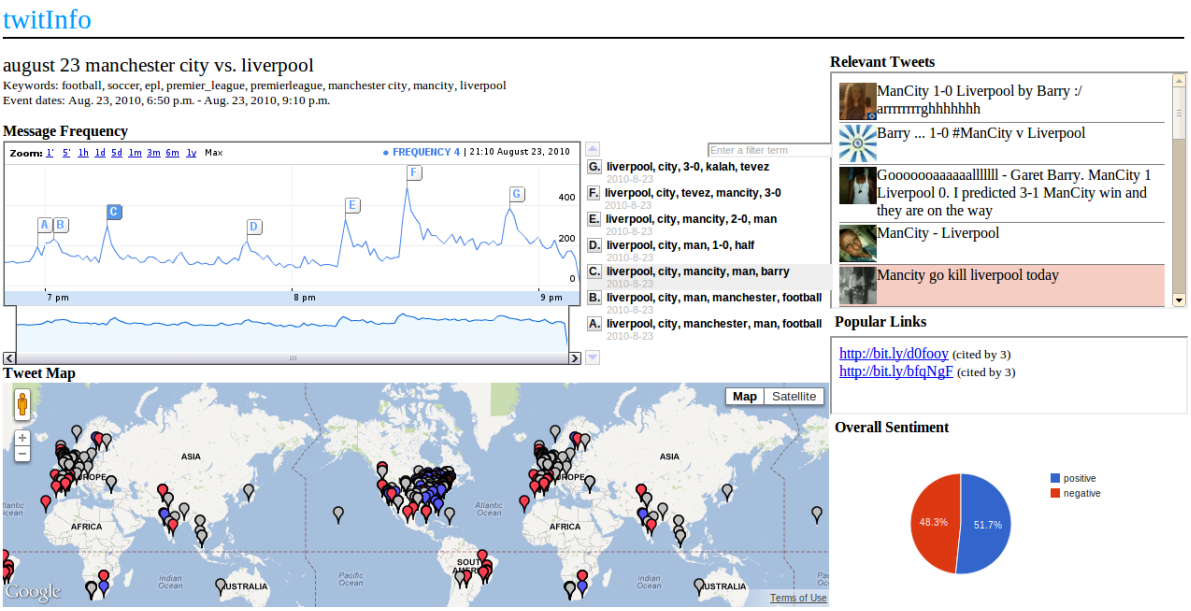

Fig. 6. TwitInfo tracks a football game (http://twitinfo.csail.mit.edu/)

the relevant tweets, news, and Wikipedia articles. For comparison, Figure 7 shows the most discussed topics related to the election, extracted from social media originating from Great Britain. While there is significant topic overlap between the two locations, the differences become also clearly visible.

Opinions and sentiment also feature frequently in visual analytics interfaces. For instance, Media Watch (Figure 5 [60]) combines word clouds with aggregated sentiment polarity, where each word is coloured in a shade of red (predominantly negative sentiment), green (predominantly positive), or black (neutral/no sentiment). Search results snippets and faceted browsing terms are also sentiment coloured. Others have combined sentiment-based colour coding with event timelines [4], lists of tweets (Figure 6 [70]), and mood maps [4]. Aggregated sentiment is typically presented using pie charts [129] and, in the case of TwitInfo, the overall statistics are normalised for recall (Figure 6 [70]).

Researchers have also investigated specifically the problem of browsing and visualising social media conversations about real-world events, e.g. broadcast events [113], football games (Figure 6 [70]), conferences [41], and news events [78,4]. A key element here is the ability to identify sub-events and combine these with timelines, maps, and topic-based visualisations.

Lastly, given the user-generated and social nature of the media streams, some visualisations have been designed to exploit this information. For instance, the PeopleSpiral visualisation [41] plots Twitter users who have contributed to a topic (e.g. posted using a given hashtag) on a spiral, starting with the most active and 'original' users first. User originality is measured as the ratio between the number of tweets authored by the user versus re-tweets made. OpinionSpace [46] instead clusters and visualizes users in a two-dimensional space, based on the opinions they have expressed on a given set of topics. Each point in the visualisation shows a user and their comment, so the closer two points – the more similar the users and opinions are. However, the purely point-based visualisation was found hard to interpret by some users, since they could not see the textual content until they clicked on a point. ThemeCrowds [8] instead derives hierarchical clusters of Twitter users through agglomerative clustering and provides a summary of the tweets, generated by this user cluster, through multilevel tag clouds (inspired by treemap visualisation). Tweet volumes over time are shown in a timeline-like view, which also allows the selection of a time period.

6.4. Discussion

Most current search, recommendation, and visualisation methods tend to use shallow textual and frequency-based information. For instance, a comparison between TF-IDF weighted topic models and LDA topic modelling has shown the former to be superior [30,102]. However, these can be improved further through integration of semantic information, as suggested by [30]. In the case of personalised recommendations, these could be improved by incorporating user behaviour roles, making better use of the latent semantics and implicit user information, as well as better integration of the temporal dimension in the recommender algorithms.

Browsing and visualisation interfaces can also be improved by taking into account the extra semantic knowledge about the entities mentioned in the media streams. For instance, when entities and topics are annotated with URIs to LOD resources, such as DBPedia, the underlying ontology can underpin hierarchically-based visualisations, including semantic relations. In addition, the exploration of media streams through topic-, entity-, and time-based visualisations can be enriched with ontology-based faceted search and semantic query interfaces. One such example is the KIM semantic platform, which is, however, aimed at largely static document collections [100].

Algorithm scalability and efficiency is particularly important, due to the large-scale, dynamic nature of social media streams. For instance, the interactive Topic Stream visualisation takes 45 seconds to compute on 1 million tweets and 325,000 contributing users, which is too long for most usage scenarios [41]. Similarly, calculating keyword correlations through point-wise mutual information is computationally too expensive on high volume blog posts [13]. A frequently used solution is to introduce a sliding window over the data (e.g. between one week and one year) and thus limit the content used for IDF and other such calculations.

In conclusion, designing effective semantic search, browsing and visualisations interfaces for media streams has proven particularly challenging. Based on our survey of the state-of-the-art, we have derived the following requirements:

- designing meaningful and intuitive visualisations, conveying intuitively the complex, multi-dimensional semantics of user-generated content (e.g. topics, entities, events, user demographics (including geolocation), sentiment, social networks);

- visualising changes over time;
- supporting different levels of granularity, both at the level of semantic content, user clusters, and temporal windows;
- allowing interactive, real-time exploration;
- integration with search, to allow users to select a subset of relevant content;
- exposing the discussion/threaded nature of the social conversations;
- addressing scalability and efficiency.

Amongst the systems surveyed, only Twitris [85] and Media Watch [60] have started to address most of these requirements, but not without limitations. Firstly, their current visualisations are mostly topic- and entity-centric and could benefit from integration of event-based visualisations, such as TwitInfo [70] and Tweetgeist [113]. Secondly, user demographics as means for stream media aggregation and exploration is mostly limited to map-based visualisations. Additional search and browsing capabilities, based around users' age, gender, political views, interests, and other such characteristics is also needed. Thirdly, methods for information aggregation and exploration, based on social networks (e.g hubs and authorities) could be combined with the currently prevailing topic- and content-centric approaches. Lastly, we would like to advocate a more substantial end-user involvement in the design and testing of new intelligent information access systems. In this way, the resulting user interfaces will address the emerging complex information seeking requirements, in terms of better support for sense making, learning and investigation, and social search [97].

7. Outstanding Challenges and Conclusions

This paper set out to explore a number of research questions arising from applications of semantic technologies to social media.

Firstly, we examined existing ontologies in the context of modelling the semantics of social media streams. Our conclusion is that most applications tend to adopt or extend more than one ontology, since they model different aspects. With respect to Web of Data resources, current methods have made most use of Wikipedia-derived resources (namely DBPedia and YAGO) and, to a lesser degree – Geonames, Freebase, and domain-specific ones like MusicBrainz. Better exploiting this wealth of semantic knowledge for semantic annotation of social media remains a challenge, which we discussed in more details in section 4.7.

Next the questions of capturing the implicit semantics and dealing with the noisy, dynamic nature of social media streams, were addressed as part of our analysis of semantic annotation state-of-the-art. We identified the need for more robust and accurate large-scale entity and event recognition methods, as well as finer-grained opinion mining algorithms to address target identification, volatility over time, detecting and modelling conflicting opinions, and opinion aggregation (see section 4.7 for details).

Thirdly, current methods for modelling users' digital identity and social media activities were discussed. Limitations with respect to modelling user interests and integration of temporal dynamics were identified, coupled with emerging need for cross-media user models. A more in-depth discussion appears in section 5.2.

Lastly, semantic-based methods for search, browsing, recommendation, and information visualisation of social media content were reviewed, from the perspective of supporting complex information seeking behaviour. As a result, seven key requirements were identified and limitations of current approaches were discussed in this context.

In conclusion, we discuss three major areas where further research is necessary.

7.1. Cross-Media Aggregation

The majority of methods surveyed here have been developed and evaluated only on one kind of social media (e.g. Twitter or blog posts). Cross-media linking, going beyond connecting tweets to news articles, is a crucial open issue, due to the fact that increasingly users are adopting more than one social media platform, often for different purposes (e.g. personal vs professional use). In addition, as people's lives are becoming increasingly digital, this work will also provide a partial answer to the challenge of inter-linking our personal collections (e.g. emails, photos) with our social media online identities.

The challenge is to build computational models of cross-media content merging, analysis, and visualisation and embed these into algorithms capable of dealing with the large-scale, contradictory and multi-purpose nature of multi-platform social media streams. For example, further work is needed on algorithms for cross-media content clustering, cross-media identity tracking, modelling contradictions between different sources, and inferring change in interests and attitudes over time.

Another related major challenge is multilinguality. Most of the methods surveyed here were developed and tested on English content only.

Lastly, as users are increasingly consuming social media streams on different hardware platforms (desktops, tablets, smart phones), cross-platform and/or platform-independent information access methods need to be developed. This is particularly challenging in the case of information visualisation on small screen devices.

7.2. Scalability and Robustness

In Information Extraction research, large-scale algorithms (also referred to as data-intensive or web-scale natural language processing) are demonstrating increasingly superior results compared to approaches trained on smaller datasets [54]. This is mostly thanks to addressing the data sparseness issue through collection of significantly larger numbers of naturally occurring linguistic examples [54]. The need for and the success of data-driven NLP methods to a large extent mirrors recent trends in other research fields, leading to what is being referred to as "the fourth paradigm of science" [14].

At the same time, semantic annotation and information access algorithms need to be scalable and robust, also in order to cope with the large content volumes, encountered in social media streams. Many use cases require online, near real-time processing, which introduces additional requirements in terms of algorithm complexity. Cloud computing [39] is increasingly being regarded as a key enabler of scalable, on-demand processing, giving researchers everywhere affordable access to computing infrastructures, which allow the deployment of significant compute power on an on-demand basis, and with no upfront costs.

However, developing scalable and parallelisable algorithms for platforms such as Hadoop is far from trivial. Straightforward deployment and sharing of semantic annotation pipelines and algorithm parallelisation are only few of the requirements, which need to be met. Research in this area is still in its infancy, especially around general purpose platforms for scalable semantic processing.

GateCloud.net [120] can be viewed as the first step in this direction. It is a novel cloud-based platform for large-scale text mining research, which also supports ontology-based semantic annotation pipelines. It aims to provide researchers with a platform-as-a-service, which enables them to carry out large-scale NLP experiments by harnessing the vast, on-demand compute power of the Amazon cloud. It also minimises the need to implement specialised parallelisable text processing algorithms. Important infrastructural issues are dealt with by the platform, completely transparently for the researcher: load balancing, efficient data upload and storage, deployment on the virtual machines, security, and fault tolerance.

7.3. Evaluation: Shared Datasets, Repeatability, and Scale

The third major open issue is evaluation. As discussed in all three application areas of semantic technologies for social media streams, lack of shared gold-standard datasets is hampering repeatability and comparative evaluation of algorithms. At the same time, comprehensive user- and task-based evaluation experiments are also required, in order to identify problems with existing search and visualisation methods. Particularly in the area of intelligent information access, most of the papers surveyed either did not report evaluation experiments, or those that did, tended to carry out small-scale, formative studies. Longitudinal evaluation with larger user groups is particularly lacking.

Similarly, algorithm training and adaptation on social media gold standard datasets is currently very limited. For example, no gold standard datasets of Twitter and blog summaries exist and there are fewer the 10,000 tweets annotated with named entities. Creating sufficiently large, vitally needed datasets through traditional expert-based text annotation methodologies is very expensive, both in terms of time and funding required. The latter can vary between USD 0.36 and 1.0 [99], which is unaffordable for corpora consisting of millions of words.

Commercial crowdsourcing marketplaces have been reported to be 33% less expensive than in-house employees on tasks such as tagging and classification [59]. Consequently, in the field of language processing, researchers have started experimenting with Amazon Mechanical Turk and game-based approaches as less expensive alternatives for the creation of annotated corpora. Poesio *et al* [99] estimate that, compared to the cost of expert-based annotation (estimated as $1.000.000), the cost of 1 million annotated tokens could be indeed reduced to less than 50% by using MTurk (i.e., $380.000 - $430.000) and to around 20% (i.e., $217,927) when using a game based approach such as their own PhraseDetectives game. In the Semantic Web field, researchers have explored mostly

crowdsourcing through games with a purpose, primarily for knowledge acquisition [116,124] and LOD improvement [128].

At the same time, researchers have turned to crowdsourcing as means for scaling up human-based evaluation experiments. The main challenge here is in how to define the evaluation task, so that it can be crowdsourced from non-specialists, with high quality results.

To conclude, crowdsourcing has recently emerged as a promising method for creating shared evaluation datasets, as well as for carrying out user-based evaluation experiments. Adapting these efforts to the specifics of semantic annotation and information visualisation, as well as using these to create large-scale resources and repeatable, longitudinal evaluations, are key areas for future work.

Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (grant EP/I004327/1). The authors wish to thank Marta Sabou and Arno Scharl for the discussions on crowdsourcing and its role in semantic technologies research, as well as Diana Maynard for the discussions on opinion mining of Twitter and Facebook messages.

References

[1] F. Abel, I. Celik, G.-J. Houben, and P. Siehndel. Leveraging the semantics of tweets for adaptive faceted search on twitter. In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, pages 1–17, Berlin, Heidelberg, 2011. Springer-Verlag.

[2] F. Abel, Q. Gao, G. J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *ESWC (2)*, pages 375–389, 2011.

[3] F. Abel, Q. Gao, G.J. Houben, and K.ele Tao. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *Proceedings of 3rd International Conference on Web Science (WebSci'11)*, Koblenz, Germany, 2011.

[4] B. Adams, D. Phung, and S. Venkatesh. Eventscapes: visualizing events over time with emotive facets. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1477–1480, 2011.

[5] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45, 1998.

[6] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586, 2005.

[7] S. Angeletou, M. Rowe, and H. Alani. Modelling and analysis of user behaviour in online communities. In *Proceedings of the 10th International Conference on the Semantic Web*, ISWC'11, pages 35–50. Springer-Verlag, 2011.

[8] D. Archambault, D. Greene, P. Cunningham, and N. J. Hurley. Themecrowds: multiresolution summaries of twitter usage. In *Workshop on Search and Mining User-Generated Contents (SMUC)*, pages 77–84, 2011.

[9] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose. Spatio-temporal analysis of topic popularity in twitter. *CoRR*, abs/1111.2904, 2011.

[10] L. Aroyo and G.-J. Houben. User modeling and adaptive semantic web. *Semantic Web*, 1(1,2):105–110, April 2010.

[11] S. Asur and B. A. Huberman. Predicting the Future with Social Media. *CoRR*, abs/1003.5, 2010.

[12] T. Baldwin and M. Lui. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California, June 2010.

[13] N. Bansal and N. Koudas. Blogscope: Spatio-temporal analysis of the blogosphere. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1269–1270, 2007.

[14] Roger Barga, Dennis Gannon, and Daniel Reed. The client and the cloud: Democratizing research computing. *IEEE Internet Computing*, 15(1):72–75, 2011.

[15] C. Beaudoin. Explaining the relationship between internet use and interpersonal trust: Taking into account motivation and information overload. *Journal of Computer Mediated Communication*, 13:550—568, 2008.

[16] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the Third International Conference on Web Search and Web Data Mining*, pages 291–300, 2010.

[17] H. Becker, M. Naaman, and L. Gravano. Selecting Quality Twitter Content for Events. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[18] Hila Becker, Mor Naaman, and Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[19] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd ACM Symposium on User Interface Software and Technology (UIST)*, pages 303–312, 2010.

[20] E. Boiy and M-F. Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558, 2009.

[21] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *http://arxiv.org/abs/0911.1583*, 2009.

[22] Johan Bollen and Huina Mao. Twitter mood as a stock market predictor. *IEEE Computer*, 44(10):91–94, 2011.

[23] K. Bontcheva and H. Cunningham. Semantic annotation and retrieval: Manual, semi-automatic and automatic generation. In J. Domingue, D. Fensel, and J. A. Hendler, editors, *Hand-*

book of Semantic Web Technologies. Springer, 2011.

[24] Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. Towards Linguistically Grounded Ontologies. In *Proceedings of the European Semantic Web Conference (ESWC'09), LNCS 5554*, pages 111—125, 2009.

[25] J. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, 2011.

[26] S. Carter, W. Weerkamp, and E. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation Journal*, Forthcoming.

[27] I. Celino, D. Dell'Aglio, E. Della Valle, Y. Huang, T. Lee, S. Park, and V. Tresp. Making Sense of Location-based Micro-posts Using Stream Reasoning. In *Proceedings of the Making Sense of Microposts Workshop (#MSM2011), collocated with the 8th Extended Semantic Web Conference, Heraklion, Crete, Greece*, 2011.

[28] D. Chakrabarti and K. Punera. Event Summarization Using Tweets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[29] J. Chan, C. Hayes, and E. Daly. Decomposing discussion forums using common user roles. In *Proceedings of WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[30] J. Chen, R. Nairn, and E. Chi. Speak little and well: recommending conversations in online social streams. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, CHI '11, pages 217–226, 2011.

[31] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, CHI '10, pages 1185–1194, 2010.

[32] Y. Cheng, G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, and C. Chen. Model bloggers' interests based on forgetting mechanism. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1129–1130, 2008.

[33] Z. Cheng. You are where you tweet: A content-based approach to geo-locating twitter users. *Proceedings of the 19th ACM Conference*, 2010.

[34] S. Choudhury and J. Breslin. Extracting semantic entities and events from sports tweets. In *Proceedings, 1st Workshop on Making Sense of Microposts (#MSM2011): Big things come in small packages*, pages 22–32, 2011.

[35] S. Das and M. Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, 2001.

[36] S. Decker and M. Frank. The Social Semantic Desktop. Technical report, DERI Technical Report 2004-05-02, 2004.

[37] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st Conference on World Wide Web*, pages 469–478, 2012.

[38] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 115–122, 2010.

[39] Marios D Dikaiakos, Dimitrios Katsaros, Pankaj Mehra, George Pallis, and Athena Vakali. Cloud computing: Distributed internet computing for it and scientific research. *IEEE Internet Computing*, 13(5):10–13, 2009.

[40] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal C. Doshi, and Joel Sachs. Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, 2004.

[41] M. Dork, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, November 2010.

[42] M. Dowman, V. Tablan, H. Cunningham, and B. Popov. Web-assisted annotation, semantic indexing and search of television and radio news. In *Proceedings of the 14th International World Wide Web Conference*, Chiba, Japan, 2005. `http://gate.ac.uk/sale/www05/web-assisted-annotat`

[43] J. Eisenstein, D. H. P. Chau, A. Kittur, and E. Xing. Topicviz: Semantic navigation of document collections. In *CHI Work-in-Progress Paper (Supplemental Proceedings)*, 2012.

[44] J. Eisenstein, B. O'Connor, N.A. Smith, and E.P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[45] Facebook. Statistics. http://www.facebook.com/press/info.php?statistics. Accessed on July 21st, 2011., 2011.

[46] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI)*, pages 1175–1184, 2010.

[47] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, 2010.

[48] K. Ganesan, C. Zhai, and E. Viegas. Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st World Wide Web Conference*, pages 869–878, 2012.

[49] A. L. Gentile, V. Lanfranchi, S. Mazumdar, and F. Ciravegna. Extracting semantic user networks from informal communication exchanges. In *Proceedings of the 10th International Conference on the Semantic Web*, ISWC'11, pages 209–224. Springer-Verlag, 2011.

[50] A. Go, R. Bhayani, , and L. Huang. Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University, 2009.

[51] A. Go, L. Huang, and R. Bhayani. Twitter Sentiment Analysis. Technical report, Stanford University, 2009.

[52] S. Gouws, D. Metzler, C. Cai, and E. Hovy. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 20–29, 2011.

[53] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. Sheth. Context and Domain Knowledge Enhanced Entity Spotting in Informal Text. In *Proceedings of the 8th International Semantic Web Conference (ISWC'2009)*, 2009.

[54] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[55] B. Han and 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*, HLT '11, pages 368–378, 2011.

[56] Sanda Harabagiu and Andrew Hickl. Relevance Modeling for Microblog Summarization. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[57] M. Hepp. Hypertwitter: Collaborative knowledge engineering via twitter messages. In *Knowledge Engineering and Management by the Masses - 17th International Conference EKAW 2010*, pages 451–461, 2010.

[58] Christopher M. Hoadley, Heng Xu, Joey J. Lee, and Mary Beth Rosson. Privacy as information access and illusory control: The case of the facebook news feed privacy outcry. *Electronic Commerce Research and Applications*, 9(1):50 – 60, 2010. Special Issue: Social Networks and Web 2.0.

[59] Leah Hoffmann. Crowd control. *Communications of the ACM*, 52(3):16 –17, 2009.

[60] A. Hubmann-Haidvogel, A. M. P. Brasoveanu, A. Scharl, M. Sabou, and S. Gindl. Visualizing contextual and dynamic features of micropost streams. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[61] N. Ireson and F. Ciravegna. Toponym resolution in social media. In *Proceedings of the 9th International Semantic Web Conference (ISWC)*, pages 370–385, 2010.

[62] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *JASIST*, 60(11):2169–2188, 2009.

[63] P. Kapanipathi, F. Orlandi, A. Sheth, and A. Passant. Personalized Filtering of the Twitter Stream. In *2nd workshop on Semantic Personalized Information Management at ISWC 2011*, 2011.

[64] J. Kincaid. Edgerank: The secret sauce that makes facebook's news feed tick, April 2010.

[65] Patrick Lai. Extracting Strong Sentiment Trends from Twitter. *http://nlp.stanford.edu/courses/cs224n/2011/reports/patlai.pdf*, 2010.

[66] David Laniado and Peter Mika. Making sense of twitter. In *International Semantic Web Conference (1)*, pages 470–485, 2010.

[67] Y. Li, K. Bontcheva, and H. Cunningham. Hierarchical, Perceptron-like Learning for Ontology Based Information Extraction. In *16th International World Wide Web Conference (WWW2007)*, pages 777–786, May 2007.

[68] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 125–132, 2003.

[69] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, 2011.

[70] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 Conference on Human Factors in Computing Systems (CHI)*, pages 227–236, 2011.

[71] D. Maynard and A. Funk. Automatic detection of political opinions in tweets. In *Proceedings of MSM 2011: Making Sense of Microposts Workshop at 8th Extended Semantic Web Conference*, Heraklion,Greece, 2011.

[72] D. Maynard and M. A. Greenwood. Large Scale Semantic Annotation, Indexing and Search at The National Archives. In *Proceedings of LREC 2012*, Turkey, 2012.

[73] L. K. McDowell and M. Cafarella. Ontology-Driven Information Extraction with OntoSyphon. In *5th Internal Semantic Web Conference (ISWC'06)*. Springer, 2006.

[74] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proceedings of the Fifth International Conference on Web Search and Data Mining (WSDM)*, 2012.

[75] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8, 2011.

[76] P. N. Mendes, A. Passant, and P. Kapanipathi. Twarql: tapping into the wisdom of the crowd. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 45:1–45:3, 2010.

[77] P. N. Mendes, A. Passant, P. Kapanipathi, and A. P. Sheth. Linked open social signals. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '10, pages 224–231, Washington, DC, USA, 2010. IEEE Computer Society.

[78] B. Meyer, K. Bryan, Y. Santos, and B. Kim. Twitterreporter: Breaking news detection and visualization through the geotagged twitter network. In *Proceedings of the ISCA 26th International Conference on Computers and Their Applications*, pages 84–89, 2011.

[79] M. Michelson and 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*, AND '10, pages 73–80, 2010.

[80] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411, 2004.

[81] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, 5(1):5–15, 2007.

[82] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM)*, pages 509–518, 2008.

[83] G. Mishne. Autotag: A collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the 15th International Conference on World Wide Web*, pages 953–954, 2006.

[84] M. Naaman, J. Boase, and C. Lai. Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work*, pages 189–192. ACM, 2010.

[85] M. Nagarajan, K. Gomadam, A. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *Web Information Systems Engineering*, pages 539–553, 2009.

[86] B. O'Connor, R. Balasubramanyan, B.R. Routledge, and 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*, 2010.

[87] T. Paek, M. Gamon, S. Counts, D. M. Chickering, and A. Dhesi. Predicting the importance of newsfeed posts and social network friends. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2010.

[88] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*, 2010.

[89] A. Pak and P. Paroubek. Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 436–439, 2010.

[90] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Information Retrieval*, 2(1), 2008.

[91] A. Passant, J. G. Breslin, and S. Decker. Rethinking microblogging: open, distributed, semantic. In *Proceedings of the 10th International Conference on Web Engineering*, pages 263–277, 2010.

[92] A. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the Linked Data on the Web Workshop (LDOW), Beijing, China*, 2008.

[93] M. Pennacchiotti and A.M. Popescu. A Machine Learning Approach to Twitter User Classification. In *Proceedings of ICWSM 2011*, pages 281–288, 2011.

[94] J. D Pennebaker, C. K. Chung, M. Ireland, Gonzales A., and R. J. Booth. The LIWC2007 Application. Technical report, 2007. http://www.liwc.net/liwcdescription.php.

[95] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189, 2010.

[96] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the 2009 ACM Conference on Recommender Systems*, pages 385–388, 2009.

[97] Peter Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *IEEE Computer*, 42(3):33–40, 2009.

[98] T. Plumbaum, S. Wu, E. W. De Luca, and S. Albayrak. User Modeling for the Social Semantic Web. In *2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, in conjunction with ISWC 2011*, 2011.

[99] M. Poesio, U. Kruschwitz, J. Chamberlain, L. Robaldo, and L. Ducceschi. Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. *Transactions on Interactive Intelligent Systems*, 2012. To Appear.

[100] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. KIM – Semantic Annotation Platform. In *2nd International Semantic Web Conference (ISWC2003)*, pages 484–499, Berlin, 2003. Springer.

[101] L. Qu, C. Müller, and I. Gurevych. Using tag semantic network for keyphrase extraction in blogs. In *Proceedings of the 17th Conference on Information and Knowledge Management*, pages 1381–1382, 2008.

[102] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[103] N. Ravikant and A. Rifkin. Why Twitter Is Massively Undervalued Compared To Facebook. *TechCrunch*, 2010. http://techcrunch.com/2010/10/16/why-twitter-is-massively-undervalued-compared-to-facebook/.

[104] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK, 2011.

[105] G. Rizzo, R. Troncy, S. Hellmann, and M. Bruemmer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *5th Workshop on Linked Data on the Web (LDoW)*, Lyon, FRANCE, 2012.

[106] M. Rowe, S. Angeletou, and H. Alani. Predicting discussions on the social semantic web. In *Proceedings of the 8th Extended Semantic Web Conference on the Semanic Web*, ESWC'11, pages 405–420. Springer-Verlag, 2011.

[107] M. Rowe and M. Stankovic. Aligning Tweets with Events : Automation via Semantics. *Semantic Web*, 1, 2009.

[108] H. Saif, Y. He, and H. Alani. Alleviating data sparsity for twitter sentiment analysis. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[109] Takeshi Sakaki, M Okazaki, and Y Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web (WWW)*, pages 851–860. ACM, 2010.

[110] H Sayyadi, M Hurst, and A Maykov. Event Detection and Tracking in Social Streams. In *Proceedings of the Third International ICWSM Conference*, pages 311–314, 2009.

[111] S. Scerri, K. Cortis, I. Rivera, and S. Handschuh. Knowledge Discovery in Distributed Social Web Sharing Activities. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[112] A. Scharl and A. Weichselbraun. An automated approach to investigating the online media coverage of US presidential elections. *Journal of Information Technology and Politics*, 5(1):121–132, 2008.

[113] D.A. Shamma, L. Kennedy, and E.F. Churchill. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events? In *Proceedings of CSCW 2010*, 2010.

[114] B. Sharifi, M. A. Hutton, and J. Kalita. Summarizing Microblogs Automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 685–688, Los Angeles, California, June 2010.

[115] W. Shen, J. Wang, P. Luo, and M. Wang. Linden: Linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st Conference on World Wide Web*, pages 449–458, 2012.

[116] K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. *Intelligent Systems, IEEE*, 23(3):50 –60, may-june 2008.

[117] Meredith M. Skeels and Jonathan Grudin. When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, pages 95–104, New York, NY, USA, 2009. ACM.

[118] G. Solskinnsbakk and J. A. Gulla. Semantic annotation from social data. In *Proceedings of the Fourth International Workshop on Social Data on the Web Workshop*, 2011.

[119] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic modelling of user interests based on cross-

folksonomy analysis. In *Proceedings of the 7th International Conference on The Semantic Web (ISWC)*, pages 632–648. Springer-Verlag, 2008.

[120] V. Tablan, I. Roberts, H. Cunningham, and K. Bontcheva. Gatecloud.net: a platform for large-scale, open-source text processing on the cloud. *Philosophical Transactions of the Royal Society A*, In Press.

[121] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 1(September 2010):1–41, 2011.

[122] K. Tao, F. Abel, C. Hauff, and G.-J. Houben. What makes a tweet relevant to a topic. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[123] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 35–44, 2011.

[124] S. Thaler, K. S. E. Simperl, and C. Hofer. A survey on games for knowledge acquisition. Technical Report Tech. Rep. STI TR 2011-05-01, Semantic Technology Institute, 2011.

[125] O. Tsur, D. Davidov, and A. Rappoport. Icwsm–a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 162–169, 2010.

[126] A Tumasjan, T O Sprenger, P G Sandner, and I M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

[127] S. Verma, S. Vieweg, W. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[128] J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. Whoknows? evaluating linked data heuristics with a quiz that cleans up dbpedia. *Interactive Technology and Smart Education*, 8(4):236–248, 2011.

[129] J. Y. Weng, C. L. Yang, B. N. Chen, Y. K. Wang, and S. D. Lin. IMASS: An Intelligent Microblog Analysis and Summarization System. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 133–138, Portland, Oregon, 2011.

[130] W Wu, B Zhang, and M Ostendorf. Automatic generation of personalized annotation tags for twitter users. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 689–692, 2010.

[131] W. Xin, Z. Jing, J. Jing, H. Yang, S. Palakorn, W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E. P. Lim, and X. Li. Topical keyphrase extraction from Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 379–388, 2011.

[132] S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In *Proceedings of ICWSM*, 2010.

[133] F. Zanzotto, M. Pennaccchiotti, and K. Tsioutsiouliklis. Linguistic Redundancy in Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 659–669, Edinburgh, UK, 2011. Association for Computational Linguistics.

[134] E. Zavitsanos, G. A. Vouros, and G. Paliouras. Classifying users and identifying user interests in folksonomies. In *Proceedings of the 2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation*, 2011.

[135] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR)*, pages 338–349, 2011.

[136] M. Zhou, S. Bao, X. Wu, and Y. Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 680–693. Springer-Verlag, 2007.