What people study when they study Twitter: classifying Twitter related academic papers

What people study when they study Twitter

Classifying Twitter related academic papers

Structured Abstract

Purpose
Since its introduction in 2006, messages posted to the microblogging system Twitter have provided a rich dataset for researchers, leading to the publication of over a thousand academic papers. This paper aims to identify this published work and to classify it in order to understand Twitter based research.

Design/methodology/approach
Firstly the papers on Twitter were identified. Secondly, following a review of the literature, a classification of the dimensions of microblogging research was established. Thirdly, papers were qualitatively classified using open coded content analysis, based on the paper's title and abstract, in order to analyze method, subject, and approach.

Findings
The majority of published work relating to Twitter concentrates on aspects of the messages sent and details of the users. A variety of methodological approaches are used across a range of identified domains.

Research Limitations
This work reviewed the abstracts of all papers available via database search on the term "Twitter" and this has two major implications: 1) the full papers are not considered and so works may be misclassified if their abstract is not clear, 2) publications not indexed by the databases, such as book chapters, are not included. The study is focussed on microblogging, the applicability of the approach to other media is not considered.

Originality/value
To date there has not been an overarching study to look at the methods and purpose of those using Twitter as a research subject. Our major contribution is to scope out papers published on Twitter until the close of 2011. The classification derived here will provide a framework within which researchers studying Twitter related topics will be able to position and ground their work.

Keywords
Twitter, Microblogging, Abstracts, Papers, Classification, Social Network Systems

Paper type

Research paper

Introduction

A number of social networking services (SNS) exist (boyd and Ellison, 2007) which have a range of features that allow users to share and exchange messages, fitting into the broader terrain of social network theory (Merchant, 2011). SNS are sometimes referred to as online social network services (OSN) (Ellison et al., 2007) and they can be divided into a number of sub-areas depending on functionality and practice. With the growing availability of easily accessible and low cost mobile technology, a niche area has developed known generically as microblogging. The use of microblogs has become a means of real time commenting on, responding to, and amplifying the impact of current events. The term "microblogging" was initially used in the early 2000s across a number of websites, and later started to appear in academic papers (Erickson, 2007, Java et al., 2007, Krishnamurthy et al., 2008). With the introduction of applications such as Twitter and Jaiku (Java et al., 2007) microblogging became more popular. By 2008 Twitter had become mainstream (Zhao and Rosson, 2009) and continues to be by far the most widely used platform.

Twitter allows users to rapidly communicate information in up to 140 characters on a one-to-one, specified group or global basis. The ease of use and essentially instantaneous nature of Twitter has made it a media for sharing news, or reports about events, ranging from the mundane (*what I had for breakfast*) through emerging information about politics (*the Arab spring*) to helping dealing with emergencies (*Japanese earthquake*) (Muralidharan et al., 2011). Events that were once closed become open to a much larger community: this has advantages such as increasing the audience for the message, mobilizing people into action, and enabling those unable to attend an event to share in the community (Dork et al., 2010). However, Twitter also brings about some interesting social issues linked to etiquette and potential misuse (Ross et al., 2011).

The openness and availability of messages posted to Twitter has provided a rich dataset for academic researchers from a variety of disciplines to study. Research ranges from the statistical through to the anthropological. This paper seeks to classify academic research on Twitter related topics based on an analysis of the abstracts of over a thousand papers published between 2007 and 2011 on the topic. Search techniques for papers related to Twitter were considered and a corpus of papers were identified, then a grounded research approach was used to identifying classifications of the work presented.

Literature Review

The literature review has been used as an integral part of the research process providing an initial foundation for a new research topic.

Microblogging and Twitter

Much of the published academic work on microblogging has focussed on the Twitter platform, with only a relatively small percentage of academic papers on Twitter using any variant of the term microblog (see Table 1).

Table 1 Numbers of Academic Papers relating to Microblogging and Twitter published between 2007 and 2011

Search Term	Databases	Search area	Items returned
micro-blogging OR micro-blog OR microblogging OR microblog	Scopus (http://www.info.sciverse.com/scopus)	Article Title, Abstracts, Keywords	436
twitter OR tweet	Scopus	Article Title, Abstracts, Keywords	1428
overlap	Scopus	Article Title, Abstracts, Keywords	276
micro-blogging OR micro-blog OR microblogging OR microblog	Web of Science (Part of the Web of Knowledge http://wok.mimas.ac.uk/ based on the Science Citation Index, the Social Sciences Citation Index and the Arts and Humanities Citation Index)	Topic	137
twitter OR tweet	Web of Science	Topic	529
overlap	Web of Science	Topic	81
micro-blogging OR micro-blog OR microblogging OR microblog	Google Scholar (http://scholar.google.com)	No control over search fields	About 10,400
twitter OR tweet	Google Scholar	No control over search fields	About 230,000
overlap	Google Scholar	No control over search fields	About 8,490

A small number of the Twitter papers returned by Scopus and Web of Science are not about the microblogging system, for example (Atencio et al., 2007) addresses vocal communication in owl monkeys: they "twitter". Google Scholar does not allow the search to be limited to specific fields and so returned a lot of papers which were not related to the microblogging system, including several where the author had the surname "Tweet", and lower down in the results returned a large number of web pages where frames surrounding an article had

links to Twitter. All academic papers found published prior to 2007 did not relate to microblogging, so Table 1 is limited to papers published between 2007, the year the first academic papers on microblogging (and Twitter) appeared, and 2011, the last full calendar year before this paper was written.

Definitions

Ross et al. (2011) have conducted an extensive literature review of published work on microblogging and Twitter, giving this definition of microblogging:

> "Microblogging is a variant of blogging which allows users to quickly post short updates, providing an innovative communication method that can be seen as a hybrid of blogging, instant messaging, social networking and status notifications. The word's origin suggests that it shares the majority of elements with blogging, therefore it can potentially be described using blogging's three key concepts (Karger and Quan, 2005): the contents are short postings, these postings are kept together by a common content author who controls publication, and individual blog entries can be easily aggregated together."

As well as incorporating characteristics of blogging, microblogging sites (such as Twitter) have elements of SNS (boyd and Ellison, 2007), with users able to construct profiles (Hughes et al., 2011) and establish and share connections with other users (Gonçalves et al., 2011). The short updates posted on microblogging sites are of limited lengths. Twitter posts are limited to 140 characters because of the original limits on short messages on mobile phones (Weller, 2011); in addition to this they sometimes have other features, with the microblogging systems Mycrocosm allowing users to share simple statistical graphs (Assogba and Donath, 2009).

User practices have had an impact on the functionality available in microblogging sites. Cormod et al. (2010) express user generated changes in the way Twitter is used:

> "What about Twitter, the minimalist site based on micro–content sharing — … the usage of the service has evolved more complex structures: follower/following relationships, targeted replies, hashtags to group tweets, re–tweeting and more. The disparate modes of access (Web, various smartphone apps, SMS) further complicate the model."

Wenger et al. (2009) report that the use of the @ symbol in front of a Twitter user name to direct a post to an individual (while still appearing in the public stream) began in a conference setting in 2007 and was immediately picked up by the developers and incorporated into a replies page. The use of hashtags were adopted by users as a way of grouping messages (Weller, 2011). A retweet button was introduced in to Twitter following users having developed a practice of amplifying messages of others by re-posting the message (boyd et al., 2010).

Classifications

Cormod et al. (2010) and Cheong and Ray (2011) classify research on Twitter and other microblogging platforms as having two central objects: the user domain (the sender of the tweet) and the message domain ("the tweet itself").

Cheong and Lee (2010) identify that the majority of Twitter-based research is within the message domain. Cormod et al. (2010) further divides research into the "first studies in Twitter" and the "next set of papers". The early work is seen as characterizing Twitter focusing on the properties relating to the domains of user and message, including quantitative studies of: the number of tweets; the number of followers and followings; times of postings; and location of posts. The next set includes linguistic and semantic analysis of tweets and identifiable conversations.

Barnes and Bohringer (2011) classify previous research on Twitter and microblogging into two broad areas: 1) understanding microblogging; 2) microblogging in special use cases. These areas are further sub-divided as:

- 1 a) Descriptive and statistical research about Twitter, including: the initial works (Erickson, 2007, Java et al., 2007, Krishnamurthy et al., 2008); studies of usage practices such as @ replies (Honeycutt and Herring, 2009) and retweeting (boyd et al., 2010).
- 1 b) Model building, for example Erickson (2008)
- 2 a) Enterprise Microblogging, based largely around round case studies (Barnes et al., 2010, Zhang et al., 2010).
- 2 b) Computer Science-oriented research, based around the technologies supporting microblogging (Passant et al., 2008, Assogba and Donath, 2009).

Dann (2010) highlights that there are a number of research papers relating to applications of Twitter in areas such as: health community, politics and government, business, education and learning , journalism, and eyewitness accounts of news stories., Examples of such papers includes work that: predicts flu trends (Achrekar et al., 2011); studies communication within government agencies (Wigand, 2010); investigates the different use by engaged and less engaged companies (Wigley and Lewis, 2012); researches detection and reaction to disasters (Muralidharan et al., 2011, Sakak et al., 2010); and experiments with the use of microblogging in higher education (Ebner et al., 2010). Work presented varies in the size, depth and length of studies. Zhao and Rosson (2009) investigated the use of microblogging in informal communication at work by using semi-structured telephone interviews with eleven subjects over four months, Erickson (2008) studying social translucence used a data set consisting of "total posts (N=1145) produced by ten Twitter subjects over a four-week period" personally interviewing subjects, while Dodds et al. (2011) investigating happiness used a data set consisting of: "over 46 billion words contained in nearly 4.6 billion expressions posted over a 33 month span by over 63 million unique users" using Amazon's Mechanical Turk (http://www.mturk.com) human intelligence work force to conduct the analysis. Collecting data has provided challenges reported in a number of papers, some papers present tools (Whitelaw et al., 2011) or repositories designed to help other researchers (Petrovi et al., 2010, Naveed et al., 2011). However Twitter's terms and conditions have limited access to such resources, such as Twapper Keeper (http://twapperkeeper.com) which is no longer freely available. Many researchers have followed advice from various sources (Russell, 2011b, Russell, 2011a) and devised their own scripts for collecting data from the Twitter API.

Non-Twitter based research still had challenges collecting data but were often able to have direct contact with the data owners (Barnes et al., 2010).

There are a number of papers in academic publications that do not fit into the areas considered above, these are papers that are general introductions or discussions. For example DeVoe (2009) explains how microblogging can be used in libraries, while McFedries (2007) - one of the earliest papers on microblogging - explains what it is and how it may be used. There are a number of papers in widely respected publications that consider the potential of microblogging and Twitter, for example in articles such as "Spies to use Twitter as crystal ball" considering the espionage use of social media (Weinberger, 2011), "Trial by Twitter" which addresses reputation issues for authors of academic papers (Mandavilli, 2011) and "Twitter thou doeth?" discussing the potential minefield for litigation arising from the use of Twitter (Kierkegaard, 2010).

Our Classification

Based on our review of the literature we have identified that microblogging has four aspects that researchers consider, which are presented below with a simple example of each:

1. Message: the text that the user enters and associated metadata identifying such things as the time sent (Cormod et al., 2010, Cheong and Ray, 2011, Barnes and Bohringer, 2011).
   An example would be a researcher considering occurrences of a particular set of words across a random sample of tweets.
2. User: aspects of the user's digital identity exposed by the microblogging system, which may include details of who the user follows, and their profile (Cormod et al., 2010, Cheong and Ray, 2011, Barnes and Bohringer, 2011, Hughes et al., 2011).
   An example would be a study of the number of followers who were also following a particular individual.
3. Technology: ranging through the underlying hardware used to implement the system through any APIs to the software the user interacts with to send messages (Barnes and Bohringer, 2011, Passant et al., 2008, Assogba and Donath, 2009).
   An example would be a researcher who had developed and trialled a new way of interfacing with Twitter.
4. Concept: encompassing introductory overviews, discussion pieces through to reviews, for example McFedries (2007), Mandavilli (2011), (Cheong and Ray, 2011). This paper would be classified as a Concept paper, as would a review of how Twitter could be used in a particular setting such as a library.

In addition researchers consider:

- The domain: Studies are undertaken from a number of different standpoints and often within a domain or a group of domains (Dann, 2010).

- The data: the size, depth and length of studies (Dodds et al., 2011, Erickson, 2008, Zhao and Rosson, 2009) impact on data collection, as does the way in which it is collected (Russell, 2011a).
- The method for their research, ranging from the use of coders to prepare data for content analysis (Waters and Jamal, 2011), through details of algorithm development (Avello, 2011) to papers predominantly on other topics but with an element of review of Twitter such as a study of accessibility of SNS that focus on Facebook (Buzzi et al., 2010).

Thus for our study we attempted to classify the aspect of an academic paper as predominantly one of these:

- Message
- User
- Technology
- Concept

With three free format fields:

- Domain
- Data
- Method

Plus an indictor as to whether the paper has: a focus on microblogging topics such as Twitter; includes mention of the topic; or is another topic but has a matching keyword.
These dimensions have similarities to conceptual models of information science which identify axes and parameters of specialisms (Hjørland, 2002, Tennis, 2003, Robinson, 2009). However here there is no attempt to define domain other than to use what Tennis (2003) describes as "common-sense parameters".

Method

Data collection

Researchers normally identify papers to consider by a number of methods such as searching in electronic databases, and chaining from existing papers. Ellis (1989) defined six characteristics of search by academic social scientists: "starting, chaining, browsing, differentiating, monitoring, and extracting", later extending the work to other disciplines, including engineering (Ellis and Haugan, 1997). Green (2000) reports humanities scholars often find resources "by following bibliographic references from documents already known to them or to their colleagues". The use of electronic databases is known to vary within domains (Talja and Maula, 2003, Tenopir et al., 2009). A number of authors have compared different databases and their use, primarily concentrating on the utility of Web of Science, Scopus and Google Scholar (Levine-Clark and Gil, 2009, Jacso, 2005), which are the most widely used.

The aim of this study was to locate academic papers on Twitter according to the classification above and identify characteristics within these classes. To ensure that the study was replicable it was decided to base it on database searches, for the period 2007 (when the first papers appeared on Twitter) to 2011 (the last complete year). There are known difficulties in social sciences and the

humanities that although books and monographs play an important role in research communication they are not indexed in major databases (Kousha and Thelwall, 2009), so it was decided to limit this study to journal articles and conference papers to ensure complete coverage of a particular format. Initial use of Google Scholar had produced many results where Twitter was mentioned on the web page, such as "Share this on Twitter" while the paper indexed itself was nothing to do with Twitter. Therefore this study was based on searches using the search word "Twitter" of Scopus and Web of Science, via our university library access, in both cases the search was based on abstract, keyword and title. Web of Science returned 384 items and Scopus 1132. Data cleansing was used to remove obvious duplicates, and items with missing data, leaving a total of 1161 items. The data cleansing was performed within an Excel spreadsheet; sorting on: year, first author name, other authors, paper title, abstract and then publication; adjacent identical items were treated as duplicates; and verified with EndNote (http://www.endnote.com/) to allow automatic detection of duplicates.

Data Classification

Papers were qualitatively classified using open coded content analysis, based on the paper's title and abstract, a technique used by Miller et al. (1996) in a similar study looking at literature relating to educational resources. Open coded analysis was selected as it facilitates delineation of concepts (Corbin and Strauss, 2008), this approach is adapted from that used in grounded theory (Glaser and Strauss, 1967) where line by line coding produces label variables from within the data itself, allowing large amounts of data to be synthesized (Glaser and Strauss, 1967). This adapted approach has been successfully used in classification of Twitter data (Ross et al., 2011).

Each paper's title and abstract was read and re-read and classified according to the schema shown in Table 2.

Table 2 Schema for classifying titles and abstracts of papers related to Twitter

Classification	Format	Details
Topic	Fixed	F = focussed on microblogging/Twitter; P = partially; N = not on topic
Message	Number	1 = mainly on this, 2 – secondly on this
User	Number	1 = mainly on this, 2 – secondly on this
Technology	Number	1 = mainly on this, 2 – secondly on this
Concept	Number	1 = mainly on this, 2 – secondly on this
Domain	Free	Semicolon separated list of domain, such as health, software development
Data	Free	Indicator of type of data and size
Method	Free	Methodological approach to research indicated.

Where a paper was partially on topic, the other classifications were based on the proportion of work relating to microblogging, not the full study.

Through our analysis, we were able to derive and develop categories from the corpus data, for both domain and method. These categories are specific to the Twitter paper corpus: they were decided upon through close examination of the corpus content. It is important to note that the stated goal of the coding was to hypothesize on the categorization of the paper, rather than to provide a descriptive evaluation of it.

Findings

Focus

Of the original 1161 papers reviewed 575 were found to have a focus on Twitter and related microblogging work; 550 included mention of the topic but it was not the focus, for example the paper entitled "Twittering on about social networking and babyfeeding matters" (Guy et al., 2010) was a cross social network investigation of potential for increasing traffic to websites related to babyfeeding, Twitter was considered alongside Facebook and Bebo. "Content is liberated!" (Goldstein and Romero, 2009) is an article about the publication IEEE Spectrum and its revamped online presence. Of the remaining papers: in 27 the reference to the term twitter was not related to microblogging but to other topics such as the sound monkeys and tractor engines make, the other 9 had identical titles and abstracts but had not been identified as duplicates in the original data cleansing due to differences in other fields, for example a conference paper also published in the employer's technical report series. The full list of papers considered is listed in CentAUR (the University of Reading's institutional repository - http://centaur.reading.ac.uk/28909/), separated into: 1) those papers that are Twitter-focussed, 2) those that mention Twitter, but do not focus on it, 3) those using the word twitter but are not related to microblogging.

The remainder of this paper concentrates on the 575 papers that focussed on Twitter and related microblogging research, below we use the term "Twitter-focussed" to refer to this group.

Year published

The first Twitter-focussed papers published appeared in 2007, when a total of 3 papers were identified in this study, this number did not increase significantly in 2008 and 2009 where 8 and 36 papers respectively were identified. There was a significant increase with 210 identified in 2010 and 320 in 2011. This matches Cormod et al. (2010) grouping of "first studies in Twitter" and the "next set of papers". As the number of papers published increases we are reaching a point where individual researchers will not be able to be familiar with all the literature published. The aim of this paper, then, is to contribute to our understanding of approach and method in studying twitter by classifying the research in this corpus.

Methods

From the abstracts, some thirty-three different research methods were initially identified as used in the published research. A number of abstracts reported using more than one method and hence the total of methods exceeds the number

of papers. Studies of methods as a source for information retrieval have indicated that it would be very useful for documents to classified by methods (Szostak, 2011), however this information is sometimes missing or presented differently according to the domain (Szostak, 2008, Hjørland, 2008). Additionally we found while reviewing the abstracts that some authors provided much more detail of their methods than others, and that one abstract may only refer to undertaking analysis while another may specify that the researchers undertook content analysis and sentiment analysis on their corpus. Therefore an overarching set of four methods were defined embracing a set of approaches.

1. Analytic
   Where the researchers had performed some type of analysis, such as content analysis [1,2], data analysis [3], semantic analysis [4], social network analysis [4]; with a quantitative or qualitative approach.
2. Design and Development
   Where systems are proposed or built [5,6], which may be exploratory, including experimental [7] or a demonstrator [8]; a model [9,10] or simulation; a full design and implementation.
3. Examination
   Where the authors had undertaken review and survey type [11] works, embracing approaches such as: biography, case study [12], essay, ethnography, evaluation, interview [10], investigation and longitudinal studies.
4. Knowledge Discovery
   In which existing techniques from artificial intelligence [2], mathematics and statistics have been applied, for the purposes of data mining, text mining and natural language processing. In addition, embracing the development of new algorithmic [13] approaches to the above.

Across the group of 575 papers spread of methodological approaches is shown in Table 3.

Table 3 Use of methods across Twitter-focussed papers in total, with an additional indication of where only one set of methods were used

Method	Total	solely
Analytic	153	97
Design and Development	267	211
Examination	139	103
Knowledge Discovery	127	59

Note there are a number of abstracts where the methods used span two or three of these methodological approaches, none spread across all four, the column "Solely" indicates the number of times a single methodological approach was used. Most of the combinations of methods happened a relatively few times, the most noteworthy were:

- Knowledge Discovery methods were used in:
  - 24 papers alongside Analytic methods [2]
  - 28 papers alongside Design and Development
  - 7 papers with both Analytic and Design and Development
- Examination methods were used in:

- 15 papers alongside Analytic
- 11 papers alongside Design and Development [10]

Almost half the abstracts indicated that the work had an element which involved the Design and Development of a system, ranging from proposals, through experiments to full implementations. While Knowledge Discovery, incorporating existing techniques from artificial intelligence, mathematics and statistics, was most frequently combined with the other methodological approaches.
Earlier work has not attempted to quantify the methods used in Twitter-focussed work and so here we have shown for the first time the diversity of approaches and the spread of their usage.

Aspects

Of the 575 Twitter-focussed papers the spread over the aspects identified are shown in Table 4. Note the diagonals indicate that there was no secondary aspect and no papers were identified as having more than two aspects.

Table 4 Combinations of Primary and Secondary aspects across the Twitter-focussed papers, note the highlighted diagonals indicate there were no Secondary aspects

	Message	**User**	**Technology**	**Concept**	**Total Secondary**
Message	266	66	12	0	78
User	80	55	2	0	82
Technology	3	0	45	1	4
Concept	1	0	0	44	1
Total Primary	**350**	**121**	**59**	**45**	

As we can clearly see the most studied topic is the Message [1,2,5] indicating that most research is done about the content of messages exchanged in Twitter. The second most studied topic is the User [8] with work relating to user profiles including lists of followings. Some 146 papers jointly considering the Message and the User (80 primary the Message [7] and 66 primarily the User [3,10]), linking investigations of content of messages with details of the tweeter and potential readers. While the Concept [11] is the least studied it should be noted that it is likely that the majority of Twitter-focussed papers will have a literature review section that discusses conceptual issues, our classification is based on the features of the work highlighted in the title and abstract. There is a relatively small proportion of work studying the Technologies [6,13] and developing them further, this maybe in part due to the proprietary nature of Twitter and the limited access developers now have to its API.

Our results are in line with the work of Cheong and Lee (2010) who identified that the majority of Twitter-based research around the message. As with Cormod et al. (2010) and Cheong and Ray (2011) we identified a second central area of

user, quantifying that a large proportion of authors address both the Message and the user: what people are saying, combined with who these people are. Other authors have not identified that there are a number of papers that do not concentrate on the Message or the User, but rather are relating to Technology and Concept. Figure 1 summarises the division of primary aspects across all the Twitter-focused papers.

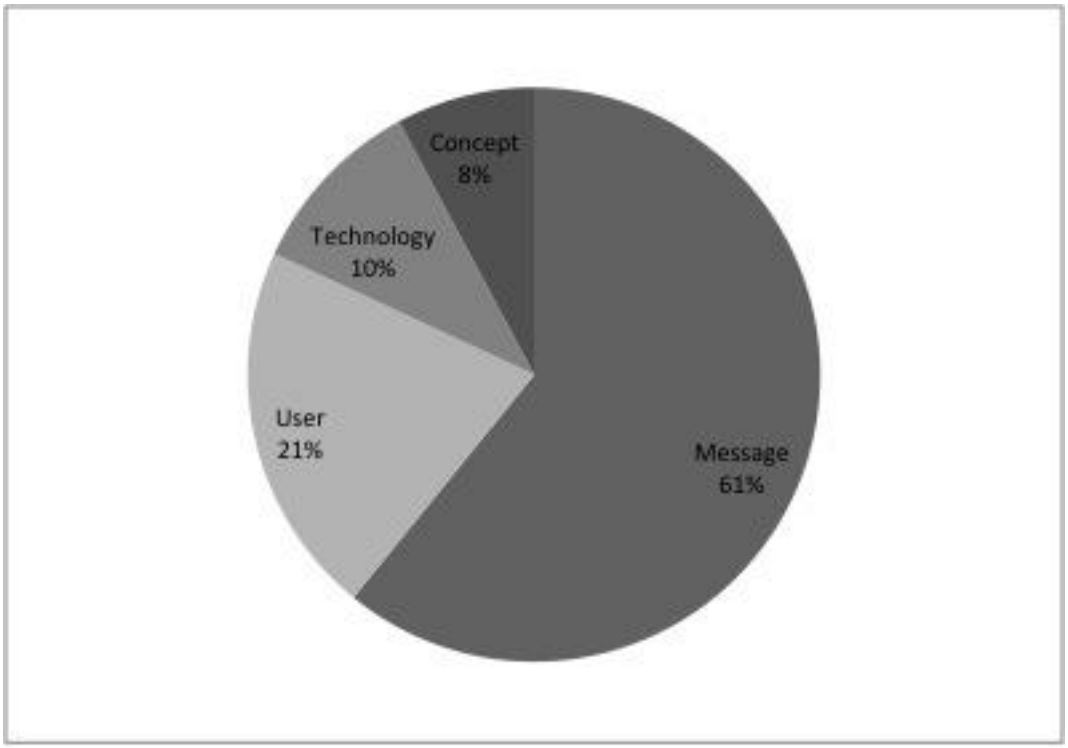

Figure 1 Pie chart summarising the division of primary aspects across all the Twitter-focused papers

Methods and Aspects

The research methods used in papers that concentrate on different aspects were investigated and are summarised in Table 5, against the broad headings of methods previously identified and the aspects: Message, User, Technology and Concept.

Table 5 methods used in Twitter-focussed papers Investigating particular aspects

Primary Aspect	**Message**	**User**	**Technology**	**Concept**	**Total**
Methods					
Analytic	120	30	3		153
Design and Development	154	58	50	4	267
Examination	60	30	8	41	140
Knowledge Discovery	94	29	4		127

The majority of the Technology papers took a Design and Development methodological approach [6], with a number of authors presenting conference

papers on systems that they have developed, and trialled. In comparison, the majority of Concept papers were based on Examination methods [11], including reviews of systems. The majority of Message oriented papers took a Design and Development approach [5].

Data

The majority of the Twitter-focussed abstracts (over 80%) did not provide any quantitative information of the data that was used in the study nor how it was collected. Phrases such as "large scale" could not be interpreted in comparison to the small number of studies which indicated orders of magnitude [5] or those giving precise details [15]. So within this study we are unable to report on results relating to the size and scope of data used in studies. This analysis therefore shows that those writing abstracts do not tend to elaborate enough on scope or method: the size of a corpus should be central to their research description.

Domain

The initial classification of domains produced over 280 categories, many of which where only used a few times, the top categories are shown in Table 6.

Table 6 The twenty most frequently used terms following the Initial Classification of domains

Domain	total
location	43
communication	29
health	29
search	29
spam	27
classification	25
education	23
politics	23
visualisation	20
sentiment	19
disaster	17
recommender	16
business	14
clustering	14
intelligence	14
libraries	13
marketing	13
semantic	12
influence	11
network	11
hashtag	10
Japan	10

The domains were therefore re-stratified into 13 broader categories, from this initial sift, to understand patterns in the data. Consolidation in this manner is a normal approach when an emergent coding approach is undertaken within content analysis (Stemler, 2001). This resulted in the following categories:

1. Business
   covering all commercial topics including public relations and marketing [16].
2. Classification
   encompassing papers that identify any patterns and clusters, including intelligence [13].
3. Communication
   ranging from communications between individuals to influencing others [3], to media such as TV and radio [1].
4. Education
   use in an educational context ranging from a formal university setting [12] to general public awareness.
5. Emergency
   covering unexpected circumstances [9], including disasters related to earthquakes and flooding.
6. Geography
   embracing place, named countries, culture and political aspects; along with the location of the user [9].
7. Health
   all health and medical issues [7].
8. Libraries
   including archives [11] and repositories.
9. Linguistics
   including syntax, semantics and sentiment, cultural protocol [4], and use in multilingual communities.
10. Search
    including recommenders, and trend recognition as well as manual and automated searches [17].
11. Security
    including SPAM, the use of automated tweeters (bot), as well as credentials, aspects of trust [8] and identity [10].
12. Technical
    embracing areas including the use of visualisation [6], networks and Twitter specifics such as hashtags.
13. Other
    all things not fitting in the above [5], including papers not grounded in a specific domain.

The Twitter-focused papers were then reallocated to these domains, where there was an apparent predominant domain that was chosen. In thirty-two cases there were two domains allocated, for example abstracts that were related to the education of health professions were classified as: Education; Health. It was not necessary to allocate more than two domains, and there were no particular pairs

of domains that were predominant and so these pairings are not considered in detail unless interesting data was observed. Figure 2 shows the number of papers allocated to each domain.

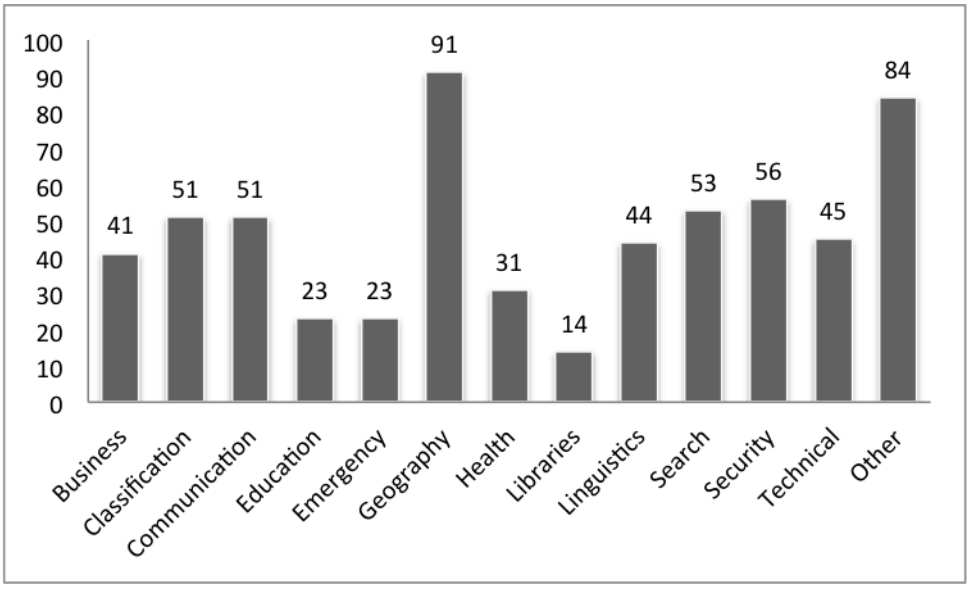

Figure 2 The stratified domains and the number Twitter-focussed Papers allocated to each

As can be seen Geography was the dominant domain with 91 of the 575 papers being related to place including named countries, the culture of the place and its politics; along with the physical location of the user. Eleven of the papers were joint with other domains, four of which were Emergency with papers addressing a particular incident in a place, and the researchers unable to identify whether the incident or place was dominant, other Emergency papers were clearly more about the incident and so were not allocated to Geography. "Other" was composed of varied areas including: tweeting pets and clothes, celebrity, and legal aspects, as within the abstract many appeared general and not in an identifiable specific domain.
These domains are in line with those identified by other researchers (Dann, 2010), however other stratifications could be chosen dividing larger categories and linking smaller ones, as is the nature of content analysis. We believe our stratification reflects the general categories people focus on when carrying out studies of Twitter based communication, based on the titles and domains of the publications in which the papers appear.

Domain, Methods and Aspects

In Table 7 we summarise for each domain the percentages of the Twitter-focused papers that used each set of methods and concentrated on each aspect.

Table 7 The methods Used and aspects Considered for each domain, expressed as percentages. Darker shading reflects larger percentage.

	Method				Aspect			
	Analytic	Design and Development	Examination	Knowledge Discovery	Message	User	Technology	Concept
Business	32%	24%	37%	15%	56%	17%	2%	24%
Classification	27%	51%	12%	29%	75%	18%	8%	0%
Communication	29%	39%	18%	27%	59%	24%	8%	10%
Education	22%	57%	43%	9%	52%	30%	9%	9%
Emergency	26%	30%	30%	22%	91%	0%	4%	4%
Geography	30%	43%	15%	26%	68%	21%	9%	2%
Health	45%	23%	42%	23%	61%	16%	3%	19%
Libraries	7%	14%	86%	7%	21%	0%	7%	64%
Linguistics	45%	45%	16%	27%	80%	14%	7%	0%
Search	21%	55%	28%	25%	62%	26%	8%	4%
Security	27%	55%	18%	18%	55%	32%	13%	0%
Technical	22%	58%	16%	18%	51%	16%	31%	2%
Other	13%	54%	32%	19%	48%	26%	14%	12%
Across all domains	**27%**	**45%**	**25%**	**22%**	**61%**	**21%**	**10%**	**8%**

Note that because more than one method is identified as used in some papers the total for methods is more than 100% within single domains. Rounding the percentages to whole numbers also introduces minor inaccuracies to the table.

The shading in the table can be used to identify anomalies, for example in the Technology aspect column most cells are lightly shaded, the darkest at 31% is Technical. This can be seen as an indication that researchers in the Technical domain having a greater proportional interest in the Technology aspect, these researchers less interested in the use of Twitter but more in how underlying tools are designed and can be improved.

There are considerable differences with the choice of methods within the various domains compared to the average across all domains. Of particular note studies within the domain of Libraries, twelve of the fourteen studies use an Examination methodological approach, with little use of other methods. While in the domain of Health only seven of the thirty-one studied adopted a Design and Development method compared to 45% overall, there was a similar lack of selection of Design and Development methods within the domain of Business (ten from forty-one), perhaps reflecting within these domains that researchers are less likely to build experimental systems or simulations than in the other domains. Studies from both the Health and the Linguistics domains were based largely on Analytic methods with respectively fourteen out of thirty-one and twenty out of forty-four compared with an average of 27%, perhaps reflecting within both domains researchers frequently want to undertake quantitative and qualitative analysis of both data and content.

When looking at the aspects the domain of Libraries is again an outlier with ten of the fourteen studies concentrating on the Concept compared with an average of only 8%. The Emergency domain concentrates on the Message with twenty-one out of twenty-three compared to the average of 61%, possibly reflecting that in emergency situation Twitter is able to provide information when conventional news services are not fast enough or may not even be available.

A Pearson correlation is a statistical measure of association between two variables: calculated values of Pearson correlation always lie between +1 and -1, a positive value indicating the two variables increase together, a negative value indicating one increases as the other decreases. The closer the Pearson value is to 1 (or -1) the stronger the association. Considering the correlation between methods and aspects across domains give Pearson values as shown in Table 8.

Table 8 The Correlation between methods and aspects across domains calculated as Pearson Values

	Message	**User**	**Technology**	**Concept**
Analytic	0.86	0.55	0.25	-0.32
Design and Development	0.82	0.92	0.79	-0.21
Examination	0.23	0.58	0.30	0.68
Knowledge Discovery	0.97	0.76	0.48	-0.22

We see there is a particularly strong correlation between the use of Knowledge Discovery methods and studying the Message. Of course a correlation does not mean that there is a causal relationship, but it would be reasonable to suppose that the Knowledge Discovery methods are suited to handling large amounts of information and that Messages are source of large quantities of information. Likewise there is a strong correlation between User and the Design and Development methods. Figure 3 presents the correlation information data in a different form mapping the number of papers in each domain that use Knowledge Discovery methods against the number of papers focusing on the Message as the first series; the second series is a similar comparison of number of papers in each domain using Design and Development methods compared to the number focusing on the User aspect.

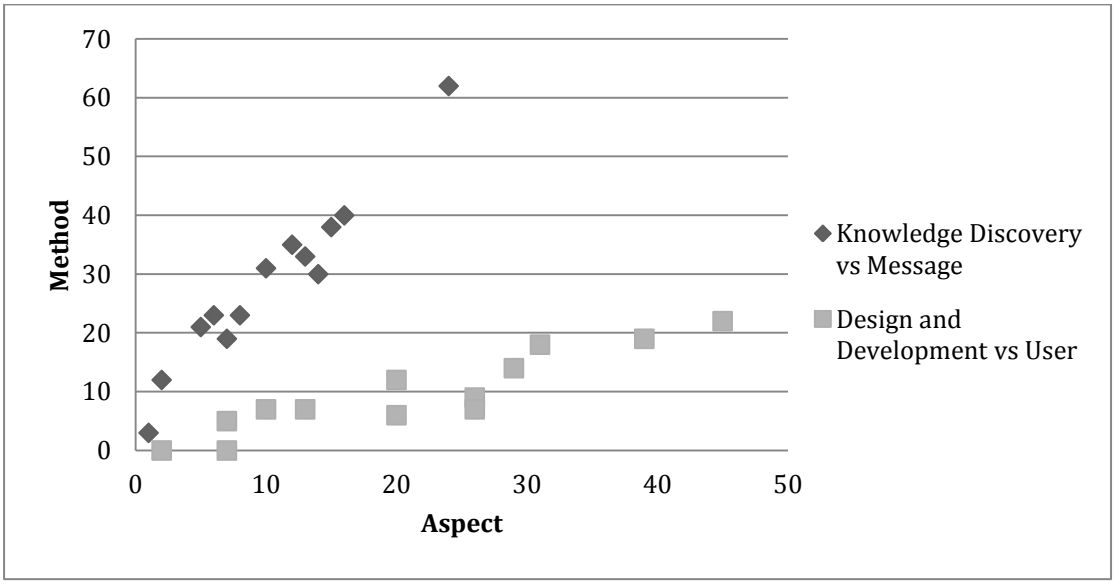

Figure 3 Number of Papers per domain for the Given method vs Number for the Given aspect

Domain Characteristics

We used the text analysis portal TAPoR (http://portal.tapor.ca) and the Voyant (http://voyant-tools.org/) toolset to analyse the text within the abstracts for each of the domains and the full set of Twitter-focussed abstracts. Frequencies of words were calculated for each set, having discounted common words and symbols using stop words from a list Taporware provide by TAPoR.

For all sets the most frequent word was "Twitter", so for the rest of this section we look at the next most frequent words. Table 9 shows the ten most frequent words. Examination of this list shows stemming has not taken place and that there are three variants of use (use, users and using), combining groups that should be stemmed and then selecting the next words gives the revised list in Table 10. Note "network" is now high in the list, it is often used in an abstract with the word "social" in phrases such as: "social network" and "social networking", in several cases these phrases were hyphenated. The list of words is not surprising and extending the list to more words did not reveal more. What was more interesting was the differences in the top ten between the full set and the individual domains. Table 11 lists distinct frequent words in the top ten of

each domain that are not in the top ten for the set of all of the Twitter-focussed abstracts.

Table 9 The Most Frequent Words across the full set of Twitter-focussed Abstracts

Word	frequency
social	711
information	495
users	473
data	376
tweets	339
paper	320
use	274
messages	257
using	257
media	243

Table 10 The Revised list of Most Frequent stemmed Words across the full set of Twitter-focussed Abstracts

Word	frequency
use	1004
social	711
tweet	543
network	498
information	495
data	376
message	323
paper	320
media	243
analysis	233

Table 11 Lists of frequent words in the most frequent top ten each domain, but not in the top ten for the full set

Domain	Distinct Words
Business	Business, Marketing, Study
Classification	Topics, Microblogging
Communication	Influence, News, TV
Education	Students, Learning, Course, Microblogging, Education, Study
Emergency	Earthquake, Event, Public
Geography	Location, Event
Health	Health, Public, Antibiotic
Libraries	Libraries, Access, Microblogging, Reference, Public, Service, New
Linguistics	Sentiment, Approach, Show

Search	Search, Web, Results, Content
Security	Spam, Web, Based, Content
Technical	Based, New
Other	Model, Microblogging

The distinct words can largely be seen to have a obvious relationship to their domain: Students participate in Education, an Earthquake causes an Emergency, the existence of SPAM means Security needs to be considered.

We highlighted in the Literature Review that the word "Microblogging" is not as widely used as the word Twitter, but in four of the domains it is among the ten most frequently used words within the abstracts, suggesting a more prevalent academic use of the term.

The word "New" is used in two domains: Libraries and Technical, where it is associated with new approaches within the discipline, this is different to the use of "News" in Communications where it is associated with current events. Automatic stemming would have occluded this difference.

The domains of Search and Security overlap on Web and Content as well as generic words, reflecting that researchers in these areas are particularly interested in material on the Internet.

Performing text analysis on the abstracts did not reveal any surprising results, rather it validated the stratification of domains and the allocation of abstracts to these. The topic of the domains were reflected by the words used within the abstracts.

Conclusions

This work has undertaken a study of over one thousand papers related to Twitter, it is to the best of our knowledge the largest study of the area. We have established that approximately half the papers that are returned by searching major databases are not focussed on Twitter, instead contributing to wider studies, often in the general area of social networking. A small group of papers (~5%) are not to do with the microblogging system but are using the term "twitter" in other ways such as describing a noise made by animals and machinery.

We have classified the remaining Twitter-focussed papers according to their abstracts across three dimensions:

- Aspect: the aspect of Twitter primarily considered, which can be one of: Message, User, Technology, Concept.
- Method: a grouping of methodological approaches, classified as one or more of: Analytic, Design and Development, Examination, Knowledge Discovery.
- Domain: a stratified list of the researchers standpoint or field interest, made up of one or more of: Business, Classification, Communication,

Education, Emergency, Geography, Health, Libraries, Linguistics, Search, Security, Technical, Other.

A fourth dimension, Data, was identified but there was not enough information provided within the abstracts to be able to attempt a classification of the quantity or quality of the data used in the studies, nor of how it was collected. The lack of this information shows that to many authors the size of the corpus or scope of their studies is not considered of sufficient importance to be included in when summarising their research

We have shown that the majority of papers (some 80%) concentrate their research around the Message and the User, considering the content of tweets and the people communicating. However, we are aware that beyond the abstract most academic papers will include a literature review that in itself we would class as Concept. The Technology aspect is thus the most under-represented in the Twitter-focussed abstracts reviewed – perhaps reflecting the technical barriers to adoption in developing tools for the Twitter API.

Earlier work did not identify the research methods used within various Twitter-focussed studies. We have identified that there are a wide variety of methods used, and often one piece of work will use multiple methods. We have grouped these methods into four broad categories of methodological approaches: Analytic, Design and Development, Examination, Knowledge Discovery. The choice of methodological approaches varies within domains, but we note there is a strong correlation between the methodological approaches of the Knowledge Discovery domain and the study of the Message [2]. Also of interest is that the majority of the Technology papers took a Design and Development methodological approach, many of these works were presented at conferences with the authors describing systems that they have developed, and trialled.

A number of areas for future work have been identified, and will be considered further. This study was based on papers published between 2007 and 2011: in future years new papers should be added to the study, and a longitudinal study undertaken of changes that occur in the focus of work, particularly linked to changes in the affordances offered by Twitter and the tools used to access it. More information is needed about the data used in the research studies and how it is collected. However since this information is not widely present in abstracts a more detailed study will be needed within a sub-area: we will investigate the largest domain: Geography and by studying the full papers aim to identify the quantity of data and how it was collected, the more detailed study of this large area will also enable the identifications of sub-domains. Differences within domains have been highlighted and within each domain there are sub-domains which may have different approaches to the study of Twitter. The approach used in this study may be applicable to papers based on other existing and emerging social networking services, academic papers relating to these services will need to be collected and considered.

The classification derived here will provide a framework within which researchers studying development and use of Twitter will be able to position

their work and against which those undertaking comparative studies of research relating to Twitter will be able to ground their work.

Notes

In this section we present examples of papers which are classified according to the dimensions identified above, and provide some explanation in the form of a thumbnail sketch based on the paper's abstract. The papers are selected to demonstrate how classification was achieved.

1. Ferguson and Greer (2011) in a paper entitled "Local Radio and Microblogging: How Radio Stations in the U.S. are Using Twitter" mention in their abstract that they use content analysis methods to understand the use of Twitter by 111 local radio stations. The study was based on examining the contents of messages, the domain was initially identified as media and radio, but following stratification this became Communication.
2. Bollen et al. (2011) present a paper "Twitter mood predicts the stock market" which examines Twitter messages to forecast according to behavioural economics. Their approach uses Analytic methods including text analysis and Knowledge Discovery including those based on artificial intelligence.
3. Khrabrov and Cybenko (2010) in the abstract of their paper "Discovering influence in communication networks using dynamic graph analysis" explain they use data analysis, within the domain of Communication. We identified the analysis is primarily on the user aspect but also the message to allow the researchers to uncover what they describe as "an ecosystem of users".
4. Lindgren and Lundstrom (2011) use both semantic and social network analysis to understand linguistic nuances in their paper "Pirate culture and hacktivist mobilization: The cultural and social protocols of #Wikileaks on Twitter". Their abstract indicates this work is in the domain of discourse later stratified to Linguistics and that they concentrate on the message aspect.
5. Dodds et al. (2011) in the abstract of their paper "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter" describe the use of Analytic methods to examine expressions made in tweets, they use Design and Development methods to construct a system that will measure happiness. Their work focuses on the message aspect, their domain is happiness/hedonemeter which was stratified as Other. This is one of the few abstracts giving details of the data set (including 46 billion words in nearly 4.6 billion expressions) and the length of the study (thirty-three months), it does not detail how the data was collected.
6. Dork et al. (2010) paper "A Visual backchannel for large-scale events" present the design of a system that will visualize Twitter data on what is called the back channel (that is not official) during large scale events such as sporting events and conferences. Their method is classed as Design and Development, their domain is Technical. They are particularly interested

in the Twitter technology which they interact with but also the messages which they display.

7. Sadikov et al. (2011) paper "Correcting for missing data in information cascades" consider the transmission of infectious diseases and the impact of identification due to missing data, they have built experimental tools which they have evaluated against 70 million Twitter nodes. The experimental nature led to classifying as a Design and Development methodological approach, the research was interested primarily in the message but also in the user. Because of the interest in infectious disease this was classed as Health.
8. Yamasaki (2011) in the paper "A trust rating method for information providers over the social web service: A pragmatic protocol for trust among information explorers and information providers" describes a demonstrator system developed for rating trust among IT-engineers based on the number of Twitter followers and other user oriented data. The paper is positioned within the domain of Security, because of the interest in trust, the method is Design and Development as a demonstrator system is described and the primary aspect is user as the interest is in the individual.
9. Gelernter and Mushegian (2011) work "Geo-parsing messages from microtext" is classified in both the domain geography and the domain emergency, with a primary aspect of message, as their work is about the type of locations that occur in disaster-related messages. They report the development of a model and so their method is classified as Design and Development.
10. Marwick and boyd (2010) paper "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience" focuses primarily on the aspect of user but also considers the message. The paper addresses the imagined audience that Twitter users interact with. The domain was initially classed as digital identity but stratified to Security. There initial approach involved talking to users and so the paper was deemed to use an Examination method, but they also develop a model and so used Design and Development methodological approaches.
11. Marshall and Shipman (2011) in their paper "Attitudes about Institutional Archiving of Social Media" report on the results of two surveys, one of which concentrated on respondents attitudes to the archiving and subsequent access of Twitter data. The domain was initially recognized as archiving, but this is not an area in which there are currently many Twitter-focused papers and so it was stratified to Libraries. The methodological approach was based on surveys and so the approach was classed as Examination. The research was generally about Twitter and so the paper was classed as the concept aspect.
12. Ebner et al. (2010) in the paper "Microblogs in Higher Education – A chance to facilitate informal and process-oriented learning?" present a case study of the use of microblogs by a group of students at an Austrian university. The research considers primarily the messages but also the users, the domain is clearly Education and the methodological approach being a case study is classed as Examination.

13. Bernstein et al. (2010) present a Twitter client they have developed in their paper “Eddi: Interactive topic-based browsing of social status streams”. The work is based on a novel algorithm and so classed as using Knowledge Discovery methodological approach. The primary aspect of interest is technology with the message secondary. The domain was initially cast as topic search, but reexamining bought it into the broader strata Classification.
14. Naaman et al. (2010)examine the Tweets of over 350 users in their paper “Is it Really About Me? Message Content in Social Awareness Streams” identifying differences in the types of messages sent. The abstract does not identify the quantity of tweets analysed nor how they were collected.
15. Arakawa et al. (2010) in the abstract for their paper “Relationship Analysis between User's Contexts and Real Input Words through Twitter” specify they examined 421274 tweets collected between two given dates, the data was collected by the then available Twitter streaming and search APIs.
16. Li et al. (2011) examined 22 official brands on the Chinese microblogging site (http://t.sina.com) in their paper “Brand tweets: How to popularize the enterprise Micro-blogs” presenting advice on how microblogging can be used in the domain of Business.
17. Chen et al. (2011) in their paper “TI: An efficient indexing mechanism for real-time search on tweets” consider the difficulties of real-time searching of Twitter data and introduce a new indexing scheme to assist. This technical paper is classified as belonging to the domain Search.

References

Achrekar, H., Gandhe, A., Lazarus, R., Ssu-Hsin, Y. and Benyuan, L. (2011), "Predicting Flu Trends using Twitter data", in *Computer Communications Workshops (INFOCOM WKSHPS)*, IEEE, pp. 702-707.

Arakawa, Y., Tagashira, S. and Fukuda, A. (2010), "Relationship Analysis between User's Contexts and Real Input Words through Twitter", in *Globecom Workshops*, IEEE, pp. 1751-1755.

Assogba, Y. and Donath, J. (2009), "Mycrocosm: Visual Microblogging", in *42nd Hawaii International Conference on System Sciences (HICSS)*, IEEE Computer Society, pp. CD-ROM 1-10.

Atencio, C. A., Blake, D. T., Strata, F., Cheung, S. W., Merzenich, M. M. and Schreiner, C. E. (2007), "Frequency-modulation encoding in the primary auditory cortex of the awake owl monkey", *J Neurophysiol,* Vol. 98 No. 4, pp. 2182-95.

Avello, D. G. (2011), "All liaisons are dangerous when all your friends are known to us", in *HT '11: Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, ACM, pp. 171-180.

Barnes, S. J. and Bohringer, M. (2011), "Modeling use Continuance Behavior in Microblogging Services: The Case of Twitter", *Journal of Computer Information Systems,* Vol. 51 No. 4, pp. 1-10.

Barnes, S. J., Böhringer, M., Kurze, C. and Stietzel, J. (2010), "Towards an understanding of social software: the case of Arinia", in *43rd Hawaii International Conference on System Sciences (HICSS)*, IEEE Computer Society, pp. CD-ROM 1-9.

Bernstein, M. S., Suh, B., Hong, L., Chen, J., Kairam, S. and Chi, E. H. (2010), "Eddi: Interactive topic-based browsing of social status streams", in *UIST 2010 - 23rd ACM Symposium on User Interface Software and Technology*, pp. 303-312.

Bollen, J., Mao, H. and Zeng, X. (2011), "Twitter mood predicts the stock market", *Journal of Computational Science,* Vol. 2 No. 1, pp. 1-8.

boyd, d. and Ellison, N. B. (2007), "Social Network Sites: Definition, History, and Scholarship", *Journal of Computer-Mediated Communication,* Vol. 13 No. 1, pp. 210-230.

boyd, d., Golder, S. and Lotan, G. (2010), "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter", in *43rd Hawaii International Conference on System Sciences (HICSS)*, IEEE Computer Society, pp. CD-ROM 1-10.

Buzzi, M. C., Buzzi, M., Leporini, B. and Akhter, F. (2010), "Is Facebook really "open" to all?", in *IEEE International Symposium onTechnology and Society (ISTAS)*, pp. 327-336.

Chen, C., Li, F., Ooi, B. C. and Wu, S. (2011), "TI: An efficient indexing mechanism for real-time search on tweets", in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 649-660.

Cheong, M. and Lee, V. (2010), "Dissecting Twitter: A Review on Current Microblogging Research and Lessons from Related Fields". *In:* Memon, N. & Alhajj, R. (eds.) *From Sociology to Computing in Social Networks: Theory, Foundations and Applications.* Springer-Verlag, New York, pp. 343 – 362.

Cheong, M. and Ray, S. (2011), "A Literature Review of Recent Microblogging Developments", report, Clayton School of Information Technology, Monash University, http://www.csse.monash.edu.au/publications/2011/tr-2011-263-full.pdf (accessed 10 July 2012).

Corbin, J. and Strauss, A. (2008) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory,* Sage Publications, , Thousand Oaks, CA.

Cormod, G., Krishnamurthy, B. and Willinger, W. (2010), "A manifesto for modeling and measurement in social media", *First Monday*, Vol. 15 No. 9, available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3072/2601 (accessed 12 July 2012).

Dann, S. (2010), "Twitter content classification", *First Monday*, Vol. 15 No. 12, available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2745/2681 (accessed 12 July 2012).

DeVoe, K. M. (2009), "Bursts of Information: Microblogging", *The Reference Librarian,* Vol. 50 No. 2, pp. 212-214.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. and Danforth, C. M. (2011), "Temporal patterns of happiness and information in a global social network: hedonometrics and twitter", *PLoS One*, Vol. 6 No. 12, available

at:
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0026752 (accessed 12 July 2012).
Dork, M., Gruen, D., Williamson, C. and Carpendale, S. (2010), "A Visual Backchannel for Large-Scale Events", *IEEE Transactions on Visualization and Computer Graphics,* Vol. 16 No. 6, pp. 1129-1138.
Ebner, M., Lienhardt, C., Rohs, M. and Meyer, I. (2010), "Microblogs in Higher Education – A chance to facilitate informal and process-oriented learning?", *Computers & Education,* Vol. 55 No. 1, pp. 92-100.
Ellis, D. (1989), "A Behavioural Approach to Information Retrieval System Design", *Journal of Documentation,* Vol. 45 No. 3, pp. 171-212.
Ellis, D. and Haugan, M. (1997), "Modelling the information seeking patterns of engineers and research scientists in an industrial environment", *Journal of Documentation,* Vol. 53 No. 4, pp. 384-403.
Ellison, N. B., Steinfield, C. and Lampe, C. (2007), "The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites", *Journal of Computer-Mediated Communication,* Vol. 12 No. 4, pp. 1143-1168.
Erickson, I. (2007), "Understanding socio-locative practices", in, pp.
Erickson, I. (2008), " The translucence of Twitter", in *Proceedings of the Ethnographic Praxis in Industry Conference*, American Anthropological Association, pp. 64-78.
Ferguson, D. A. and Greer, C. F. (2011), "Local Radio and Microblogging: How Radio Stations in the U.S. are Using Twitter", *Journal of Radio and Audio Media,* Vol. 18 No. 1, pp. 33-46.
Gelernter, J. and Mushegian, N. (2011), "Geo-parsing messages from microtext", *Transactions in GIS,* Vol. 15 No. 6, pp. 753-773.
Glaser, B. G. and Strauss, A. (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research,* Aldine Publishing, Chicago, IL.
Goldstein, H. and Romero, J. J. (2009), "IEEE Spectrum online: Content is liberated!", *Spectrum, IEEE,* Vol. 46 No. 7, pp. 10-10.
Gonçalves, B., Perra, N. and Vespignani, A. (2011), "Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number", *PLoS ONE*, Vol. 6 No. 8, available at:
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0022656 (accessed 12 July 2012).
Green, R. (2000), "Locating Sources in Humanities Scholarship: The Efficacy of following Bibliographic References", *The Library Quarterly,* Vol. 70 No. 2, pp. 201-229.
Guy, C., Paterson, A., Currie, H., Lee, A. J. and Cumming, G. P. (2010), "Twittering on about social networking and babyfeeding matters", *British Journal of Midwifery,* Vol. 18 No. 10, pp. 620-627.
Hjørland, B. (2002), "Domain analysis in information science: Eleven approaches - traditional as well as innovative", *Journal of Documentation,* Vol. 58 No. 4, pp. 422-462.
Hjørland, B. (2008), "Core classification theory: a reply to Szostak", *Journal of Documentation,* Vol. 64 No. 3, pp. 333-341.

Honeycutt, C. and Herring, S. C. (2009), "Beyond Microblogging: Conversation and Collaboration via Twitter", in *42nd Hawaii International Conference on System Sciences (HICSS)*, IEEE Computer Society, pp. CD-ROM 1-10.
Hughes, D. J., Rowe, M., Batey, M. and Lee, A. (2011), "A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage", *Computers in Human Behavior,* Vol. 28 No. 2, pp. 561–569.
Jacso, P. (2005), "As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases", *Current Science,* Vol. 89 No. 9, pp. 1537-1547.
Java, A., Song, X., Finin, T. and Tseng, B. (2007), "Why we twitter: understanding microblogging usage and communities", in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07*, pp. 56-65.
Karger, D. R. and Quan, D. (2005), "What would it mean to blog on the semantic web?", *Web Semantics: Science, Services and Agents on the World Wide Web,* Vol. 3 No. 2-3, pp. 147-157.
Khrabrov, A. and Cybenko, G. (2010), "Discovering influence in communication networks using dynamic graph analysis", in, pp. 288-294.
Kierkegaard, S. (2010), "Twitter thou doeth?", *Computer Law & Security Review,* Vol. 26 No. 6, pp. 577-594.
Kousha, K. and Thelwall, M. (2009), "Google book search: Citation analysis for social science and the humanities", *Journal of the American Society for Information Science and Technology,* Vol. 60 No. 8, pp. 1537-1549.
Krishnamurthy, B., Gill, P. and Arlitt, M. (2008), "A few chirps about Twitter", in *Proceedings of the 1st Workshop on Online Social Networks: WOSN'08*, ACM, pp. 19-24.
Levine-Clark, M. and Gil, E. (2009), "A comparative analysis of social sciences citation tools", *Online Information Review,* Vol. 33 No. 5, pp. 986-996.
Li, G., Cao, J., Jiang, J., Li, Q. and Yao, L. (2011), "Brand tweets: How to popularize the enterprise Micro-blogs", in *Information Technology and Artificial Intelligence Conference (ITAIC)*, IEEE, pp. 136-139.
Lindgren, S. and Lundstrom, R. (2011), "Pirate culture and hacktivist mobilization: The cultural and social protocols of #WikiLeaks on Twitter", *New Media & Society,* Vol. 13 No. 6, pp. 999-1018.
Mandavilli, A. (2011), "Trial by Twitter", *Nature,* Vol. 469 No. 7330, pp. 286-287.
Marshall, C. C. and Shipman, F. (2011), "Attitudes about Institutional Archiving of Social Media", in *Archiving 2011: Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organizations*, pp. 194-198.
Marwick, A. E. and boyd, d. (2010), "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience", *New Media & Society,* Vol. 13 No. 1, pp. 114-133.
McFedries, P. (2007), "All A Twitter", *IEEE Spectrum,* Vol. 44 No. 10, pp. 84.
Merchant, G. (2011), "Unravelling the social network: theory and research", *Learning, Media and Technology,* Vol. 37 No. 1, pp. 4-19.
Miller, K. J., Fullmer, S. L. and Walls, R. T. (1996), "A Dozen Years of Mainstreaming Literature: A Content Analysis", *Exceptionality,* Vol. 6 No. 2, pp. 99-109.

Muralidharan, S., Rasmussen, L., Patterson, D. and Shin, J.-H. (2011), "Hope for Haiti: An analysis of Facebook and Twitter usage during the earthquake relief efforts", *Public Relations Review,* Vol. 37 No. 2, pp. 175-177.

Naaman, M., Boase, J., Lai, C.-H. and Acm. (2010), "Is it Really About Me? Message Content in Social Awareness Streams", in *2010 ACM Conference on Computer Supported Cooperative Work*, pp. 189-192.

Naveed, N., Gottron, T., Kunegis, J. and Alhadi, A. C. (2011), "Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter", *Proceedings of the ACM WebSci'11*, available at: http://www.websci11.org/fileadmin/websci/Papers/50_paper.pdf (accessed 12 July 2012).

Passant, A., Hastrup, T. and Boj, U. (2008), "Microblogging : A Semantic and Distributed Approach", *Proceedings of the 4th Workshop on Scripting for the Semantic Web*, available at: http://hdl.handle.net/10379/539 (accessed 12 July 2012).

Petrovi, S., Osborne, M. and Lavrenko, V. (2010), "The Edinburgh Twitter corpus", *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, available at: http://www.aclweb.org/anthology/W/W10/W10-05.pdf - page=37 (accessed 12 July 2012).

Robinson, L. (2009), "Information science: communication chain and domain analysis", *Journal of Documentation,* Vol. 65 No. 4, pp. 578-591.

Ross, C., Terras, M., Warwick, C. and Welsh, A. (2011), "Enabled backchannel: conference Twitter use by digital humanists", *Journal of Documentation,* Vol. 67 No. 2, pp. 214-237.

Russell, M. A. (2011a) *21 recipes for mining Twitter,* O'Reilly, Sebastopol, Calif.

Russell, M. A. (2011b) *Mining the social web,* O'Reilly, Sebastopol, Calif.

Sadikov, E., Medina, M., Leskovec, J. and Garcia-Molina, H. (2011), "Correcting for missing data in information cascades", in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM* pp. 55-64.

Sakak, T., Okazaki, M. and Matsuo, Y. (2010), "Earthquake shakes twitter users: real-time event detection by social sensors", in *9th international conference on World wide web*, International World Wide Web Conference Committee (IW3C2), pp. 851-860.

Stemler, S. (2001), "An Overview of Content Analysis", *Practical Assessment, Research & Evaluation*, Vol. 7 No. 17, available at: http://PAREonline.net/getvn.asp?v=7&n=17 (accessed 12 July 2012).

Szostak, R. (2008), "Classification, interdisciplinarity, and the study of science", *Journal of Documentation,* Vol. 64 No. 3, pp. 319-332.

Szostak, R. (2011), "Complex concepts into basic concepts", *Journal of the American Society for Information Science and Technology,* Vol. 62 No. 11, pp. 2247-2265.

Talja, S. and Maula, H. (2003), "Reasons for the use and non-use of electronic journals and databases: A domain analytic study in four scholarly disciplines", *Journal of Documentation,* Vol. 59 No. 6, pp. 673-691.

Tennis, J. T. (2003), "Two Axes of Domains for Domain Analysis", *Knowledge Organization,* Vol. 30 No. 3/4, pp. 191-195.

Tenopir, C., King, D. W., Spencer, J. and Wu, L. (2009), "Variations in article seeking and reading patterns of academics: What makes a difference?", *Library & Information Science Research,* Vol. 31 No. 3, pp. 139-148.

Waters, R. D. and Jamal, J. Y. (2011), "Tweet, tweet, tweet: A content analysis of nonprofit organizations' Twitter updates", *Public Relations Review,* Vol. 37 No. 3, pp. 321-324.

Weinberger, S. (2011), "Spies to use Twitter as crystal ball", *Nature,* Vol. 478 No. 7369, pp. 301.

Weller, M. (2011) *The Digital Scholar,* Bloomsburt Academic, London.

Wenger, E., White, N. and Smith, J. D. (2009) *Digital Habitats,* CPsquare, Portland.

Whitelaw, C., Agrawal, M., Rao, H. R. and Onook, O. (2011), "Using social media to study social pheonomena: An example using Twitter data", in *Wireless Telecommunications Symposium (WTS), 2011*, pp. 1-3.

Wigand, F. D. L. (2010), "Twitter takes wing in government: diffusion, roles, and management", in *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, Digital Government Society of North America, pp. 66-71.

Wigley, S. and Lewis, B. K. (2012), "Rules of engagement: Practice what you tweet", *Public Relations Review,* Vol. 38 No. 1, pp. 165-167.

Yamasaki, S. (2011), "A trust rating method for information providers over the social web service: A pragmatic protocol for trust among information explorers and information providers", in *Proceedings - 11th IEEE/IPSJ International Symposium on Applications and the Internet, SAINT* pp. 578-582.

Zhang, J., Qu, Y., Cody, J. and Wu, Y. (2010), "A Case Study of Micro-blogging in the Enterprise: Use, Value, and Related Issues", in *CHI 2010: Organizations and Communities* pp. 123-132.

Zhao, D. and Rosson, M. B. (2009), "How and why people Twitter: the role that micro-blogging plays in informal communication at work", in *Proceedings of the ACM 2009 international conference on Supporting group work - GROUP '09*, pp. 243-252.