Named Entity Trends Originating from Social Media

Nigel Dewdney
Department of Computer Science
University of Sheffield
acp08njd@sheffield.ac.uk

ABSTRACT
There have been many studies on finding what people are interested in at any time through analysing trends in language use in documents as they are published on the web. Few, however have sought to consider material containing subject matter that originates in social media. The work reported here attempts to distinguish such material by filtering out features that trend primarily in news media. Trends in daily occurrences of nouns and named entities are examined using the ICWSM 2009 corpus of blogs and news articles. A significant number of trends are found to originate in social media and that named entities are more prevalent in them than nouns. Taking features that trend in later news stories as a indication of a topic of wider interest, named entities are shown to be more likely indicators although the strongest trends are seen in nouns.

KEYWORDS: Social media, Trend analysis.

Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, pages 1–16,
COLING 2012, Mumbai, December 2012.

1 Introduction

The detecting and tracking of popular topics of discussion through trend analysis has become a keenly studied an developed area with the rise in use of social media. Various algorithms have been proposed for finding the ”hot topics” of current interest in user communities with various social media providers, such as Twitter.com, providing a trending topics service. Typical topics that are keenly discussed are often around breaking and current news stories. Occasionally the social media may be the first to break the news, as famously with the Haitian earthquake of 2008, or even be at the centre of news stories such as with the events of the ”Arab Spring”. However, sometimes stories and information may originate from social media, rapidly spreading and rising in popularity. Such instances of the spread of information have been described as ”going viral”. The question arises, then, as to what information may lie in social media that might have sufficient potential to be interesting to many others, but is largely lost due to the dominance of current affairs. Are social media topics of interest, other than what is in the news, general in nature or are they about specific things?

Although much of the news available online is sourced from and published by professional media organisations, there are an increasing number of people using web logs, or “blogs” where authors provide original material and opinions on topics of interest to them [18]. Micro-blogs, as popularised by Twitter, provide a more immediate shorter form, but being shorter are less likely to be a rich source of information at the individual message level. Alvanaki et al. [2] have likened tweets (twitter postings) to chat, the longer blogs form being more akin to publication of articles. The study here, therefore, will focus on personal blogs.

Many trending topics have been found to related to current news stories, however Lloyd et al. have found that a small percentage of these originate from personal blogs [21]. In such cases it may be that the popularity of the topic itself becomes the story, as an example of interest “going viral”. A news story of this type, it could be argued, would be a report of the kind of phenomenon of interest here, i.e. a trend originating in social media.

Rising popular activity in social media may not be isolated to national and international situations involving a large population. Speculation around and interest in imminent or recent product releases for example is one area where information may be more readily found in social media than in the main stream. Such information is of interest to marketing companies; social media is an important source for product feedback and marketing strategy monitoring.

A recent example is that of interest surrounding an upcoming release of a computer game called "Black Mesa", a fan remake of "Half Life", a popular commercial PC game from the previous decade. The game has been the subject of much discussion amongst enthusiasts which increased following the announcement of a release date. Discussion even got as far as a news report on the BBC news website on the $3rd$ September 2012. The Graph in Figure 1 shows the number of blog posts published each day during a three week period up to the BBC news story, as measured with Google’s blog search engine.

A natural question to ask, then, is what is the nature of trending topics that *originate* in social media, i.e. those not sparked by topics already in the news? Are there characteristics in trending features that show social media originated trends to be significantly different from news stories, or do we see the citizen journalist [6] in action? The work described here begins to investigate these questions.

As topics that originate in social media are of particular interest here, it will be necessary to

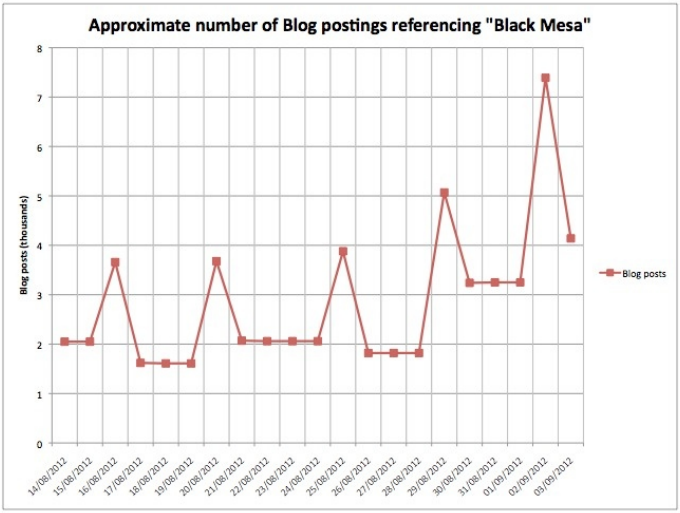

Figure 1: Blog post results for daily search for "Black Mesa game" using Google's blog search engine

identify those that originate in reports made by the mainstream media so that they may be distinguished from those originating in social media. Some of these topics may go on to be of interest in mainstream news, while we would expect many to remain within the "blogosphere". As an initial step towards characterising trending topics with social media origins, we examine the nature of the trending features: Are specific entities or generic nouns more prevalent, and what are the relative strengths of their trends?

This paper reports on an analysis of trending common nouns and named entities originating in social media, i.e. after having filtered mainstream news stories out, using the ICWSM 2009 Spinn3r dataset [7]. The rest of the paper is organised as follows: section 2 summarises recent relevant and related work; section 3 provides a description of the data and the method of analysis employed; section 4 describes the analysis of the results; finally section 5 gives conclusions and outlines future work.

2 Related work

Trend analysis has been a popular area of study in recent years with the rise in popularity of social media as a means to disseminate information, provide opinion and facilitate discussion. Discovering what the popular topics within populations are at any particular time is of potential interest to many, including politicians, journalists, and marketing departments. Numerous approaches have been suggested and implemented. Detection of changes in language use as new documents are published is often at the heart of these methods, as new topics emerge and are written about.

A burst in activity may be expected with sudden popular interest in a topic, and reflected in document features. Various different models have been proposed for modelling streams of text to account for bursts. Church found that words that show a high likelihood of re-occurring in a document under a Poisson model, one would often consider to be a "content" word [10]. Sarkar et al. used a double Poisson Bayesian mixture model for term "burstiness", to determine such "content" words [26]. Baroni and Evert have taken a different approach for document term burst modelling [5], proposing the use of document frequency rather than term frequency.

Many approaches to detecting new emerging topics have been based on detecting bursts in term use. For example, Kleinberg examined time gaps between term occurrences in email data and found bursts in email topics seemed to coincide with interest to the author [17], and Kumar et al. observed bursts in links being established in the evolution of the "Blogosphere" [18]; Franco and Kawai have investigated two approaches to detecting emerging news in blogs [13], through blogosphere topic propagation measured by evolution of link numbers, and by blog post clustering; Viet-Ha-Thuc et al. used a log-likelihood estimate of an event within a topic model [16]; and Glance et al. have examined bursts in phrases, mentions of people, and hyperlinks in blogs given a background of blogs published in the preceding two weeks [15].

Other approaches to topic detection and tracking have sought to include structure, see [23], [11] for examples; and topic classification as in [30], or [14], where Gabrilovich et al. used bursts of articles with high divergence from established topic clusters to detect new stories. However new topic detection is difficult though as noted by Allan et al. [1], who comparing the task to that of information filtering, show new story detection in tracking is poor.

Micro-blogs, such as that facilitated by Twitter, have provided a rich source of data for those studying trends and their evolution. Micro-blogs, or "Tweets", are restricted to 140 characters, and has been likened to chat rather than publication by Alvanaki et al. [2]. In their "En Blogue" system, they detect emerging topics by considering frequent tags and co-incident tags (these are augmented by extracted named entities). Twitter provides its own proprietary trending topics service, but others have sought to provide similar functionality. Petrović et al. have investigated first story detection in Twitter micro-blog feeds [24]; Mathioudakis and Koudas describe a system that detects and groups bursting keywords [22]; Cataldi et al. consider a term to be emerging if it frequently occurs in the interval being considered whilst relatively infrequently in a defined prior period, generating emerging topics from co-occurrence vectors for the considered interval [8].

Research has also looked at how trends evolve through social media and how content spreads: Cha et al. have studied how media content is propagates through connected blogs [9]; Simmons et al. have examined how quoted text changes as it is communicated through social media networks [27]; and Lerman and Ghosh have studied the spread of news through the Digg and Twitter social networks [19]. Asur et al. have examined how trends persist and decay through social media [3] finding that the majority of trends follow news stories in Twitter, with re-tweeted items linked to news media providers such as CNN and Reuters.

Trending topics not linked to stories reported in the mainstream media have been found. Lloyd et al. found a small percentage of blog topics trended before the news-stories were published [21]. They compared the most popular named entities in news and blogs on a mentions-per-day basis finding that maximal spikes could be present in one medium before the other. Leskovec et al. in investigating concept of "memes", short phrases, and how they evolved in news websites and blog publication, found a small percentage of quotations to originate in personal blogs

rather than news reports [20]. This small percentage of material indicated by these two studies is makes up the source of interest here.

3 Data and analytic approach

Blog data for this study comes from the ICWSM 2009 corpus, made available to researchers by the organisers of the 3^{rd} International AAAI Conference on Weblogs and Social Media (2009) [7]. The dataset, provided by Spinn3r.com, comprises some 44 million blog posts and news stories made between August 1st and October 1st, 2008. For the experiments reported here the data is pre-processed. Blog posts that have been classified either as “MAINSTREAM NEWS” or “WEBLOG” are extracted, while “CLASSIFIED” postings and empty postings (here less than 3 characters long) are discarded. Applying trend analysis to each class will allow the likely trend source to be identified.

Many trend analysis approaches analyse simple lexical features, before using other techniques, such as clustering and feature co-occurrence analysis, to improve the semantic richness. (See [22], [8], [2] for examples.) Trending topics involve tangible (named) entities; Azzam et al. suggested that a document be about something – its topic – and that something would revolve about a central entity [4]. There is also evidence that names can be effective in information retrieval tasks [28], and searching for names has been shown to be useful concept in news archive search [25].

Rather than relying solely on lexical statistics to determine both content bearing features and trends, the approach taken applies part-of-speech tagging and named entity recognition prior to statistical analysis. The Stanford CoreNLP toolset, using the supplied 4-class model for English text without any modification [29][12] is used here. The training data used for the model was that supplied for CoNLL 2003 and is made up of Reuters Newswire. Although the training data does not perfectly match blog data, articles of interest may be expected to have some similarity in style, i.e. "reporting". It is assumed that the performance of the natural language processing is sufficiently robust such that output will be substantially correct English given noisy input. (Note: Input data is pre-processed to remove any html mark-up for this investigation.)

We consider a trending topic to be one that shows an increase in the number of occurrences of associated features, be they nouns or named entities, from that expected. For this we employ a a traditional Poisson model parameterised by the observed mean feature occurrence rate. The model assumes that features occur at random and independently, the intervals between occurrences being Poisson distributed. The reciprocal of the expected interval gives the expected frequency. Positive deviations (decrease in arrival time) from expectation are indicative of a trending feature, with strength measured by the size of the deviation.

A random variable X with a Poisson distribution, expectation $E[X] = \lambda$, has the model:

$$P(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, k \geq 0 \tag{1}$$

The mean frequency is simply the inverse of the expected gap between occurrences for the feature k, i.e. $1/\lambda$. As the variance of the Poisson distribution is also λ trend strength can be measured as the number of standard deviations, $\sqrt{\lambda}$, the gap reduction is from the mean. It follows that for feature k with expected frequency $\frac{1}{\lambda_k}$ and observed frequency $\frac{1}{\lambda'_k}$, the strength of a trend in k is given by:

$$T(k) = \frac{\lambda_k - \lambda'_k}{\sqrt{\lambda_k}} \tag{2}$$

A daily trend in feature occurrence is measured in standard deviations given by $T(k)$ from the expected frequency, calculated by averaging observed frequency over preceding days. Feature frequencies are calculated on a daily basis with average frequencies being calculated accumulatively. A number of days of observation are required to establish a reasonable estimate of the average frequency $1/\lambda_X$ for each feature $X = 1, 2, ...$ In this study, one week of observations are used prior to application of trend detection.

Following the method of Lloyd et al. [21],trend analysis is applied independently to the "MAINSTREAM NEWS" and "BLOG" classes of posts in the corpus, thus allowing the likely trend source to be identified. The focus, then, is on named entities and nouns that show trending behaviour, originating in social media blogs.

4 Experiments and Results

Over the two full months of data in the corpus, August and September 2008, there are 1,593,868 posts from mainstream news sources, while there are 36,740,061 blog postings. Of these, 1,428,482 news stories and 27,074,356 blogs contain at least one entity, and all bar 157 blog postings contain English nouns (although there is no guarantee the post is actually in English).

The amount of material produced each day is not consistent however as can be seen from the graphs shown in figures 2 and 3, although News postings show a periodic nature as one might expect. There is a notable increase in noun output in blogs but not in news towards the end of the period, although this increase is not seen named entity output. The number of postings made per day shows no significant change suggesting that the rise in noun output is due to a relatively small number of long blog postings that do not mention a correspondingly higher number of named entities.

We now turn our attention to those features that demonstrated a rising trend in occurrence in blogs during the period of the corpus, either exclusively or prior to a trend in news articles. Minimum criteria for feature selection are that they have a minimum of over five occurrences or show a positive deviation of over five standard deviations from their average daily occurrence at the time of their maximum positive trend. Trends for features that have trended in news articles within the previous seven days are not considered. No trend analysis is carried out for the first seven days to allow a fair estimate of average daily occurrence to be established, so occurrences on the 8^{th} August are the first to be considered, being reported therefore on the 9^{th}. This selection process yields a total of 47639 features that show a positive trend originating in social media from the 8^{th} Augusts 2008 to 30^{th} September 2008. An average of 60.4% of these trending features are also seen in later trends within news articles. The break-down across nouns and named entities is given in table 1.

A high proportion of nouns that show trending behaviour originating in blogs, about 73%, are within the vocabulary of news articles. The lack of editorial control together with tagger inaccuracies account for much of the remainder. A much lower proportion of named entities that originally trend within blogs are also seen in news at all. We may conclude that while some people, organisations, and places etc. may be of topical interest in the social media, only

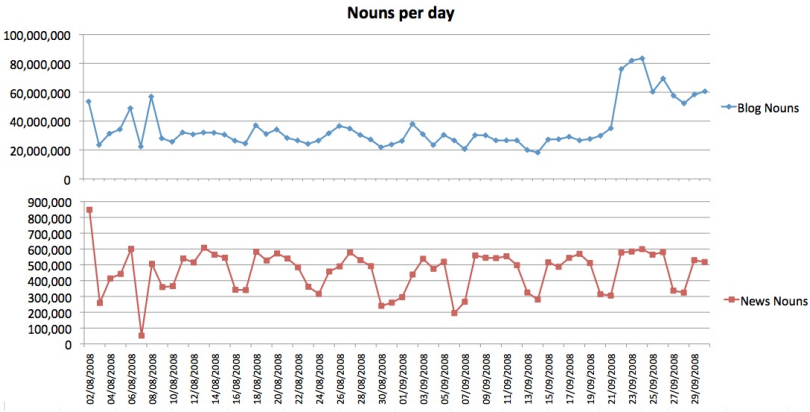

Figure 2: Nouns in blogs and news per day in ICWSM 2009 corpus

Type	No. Trending	No. in News pre/post trend	%
Nouns	12157	8827	72.6%
Misc	11260	5303	47.1%
Location	9809	5331	54.3%
Person	9365	5993	64.0%
Organisation	9823	5889	60.0%
Totals	47639	28776	60.4%

Table 1: Unique social media originating trending feature totals & amount in news use

abut half of them (between 47% and 64%) are also in the sphere of interest of the mainstream media organisations.

As trend strength is measured relative to the average occurrence of a feature rather than in absolute occurrence numbers, the most popular nouns and entities are not necessarily the same as those that show the strongest trend. Table 2 notes the top ten most frequent nouns and the top ten strongest trending nouns with their average frequency and trend strengths at the time of they showed their strongest trend. There are two observations to make: Firstly that a significant number of these "nouns" are not correctly identified by the part-of-speech identifier and named entity tagger, being either broken mark-up or proper nouns; Secondly the strongest trending contain the unidentified proper nouns.

Tables 3,4,5 and 6 show the top ten by average occurrence and by maximum trend strength for Organisation, Person, Location and Miscellaneous names. The most frequent entities are mentioned several thousand times a day (about an order of magnitude less than the most frequent nouns). Their trend strengths range from a few 10's of std deviations from their average daily occurrence to a few thousand, similar in strength to the top occurring nouns. The trend strengths are typically well under those shown by the top ten entities by maximum trend strength, which are in the region of several thousand standard deviations. These too are an order of magnitude less than trend strengths shown by the maximally trending nouns. Overall,

Top Occurring			Top Trending		
Noun	Avg per Day	Max Trend	Noun	Avg per Day	Max Trend
QUE	67072.4	958.9	WON	184.2	92152.5
%	60444.3	1002.7	––––––––––...–	14.8	83949.8
THINGS	52755.2	410.4	3A	36.1	76574.3
COM	51408.7	5322.4	BEHAR	59.4	75912.3
SOMETHING	48963.0	547.2	PEARCE	87.6	69001.3
DA	46427.7	4269.0	PROP	163.2	68665.7
GIRL	44684.9	1255.9	<BR?/>	810.1	51689.2
MUSIC	44454.6	740.1	PIVEN	705.3	50291.6
DVD	44327.5	4001.1	ANTOFAGASTA	16.2	48510.5
EL	41919.9	666.7	JEUDI	142.1	46578.9

Table 2: Top ten 'nouns' by average daily occurrence and by trend strength in blogs

Top Occurring			Top Trending		
Noun	Avg per Day	Max Trend	Noun	Avg per Day	Max Trend
GOOGLE	10443.0	303.9	ILWU	0.8	5929.3
APPLE	3138.6	577.7	ADM	24.4	5459.2
UA	3083.5	2359.9	OHIO STATE	211.1	5452.2
YAHOO	2279.1	2409.3	STATE FARM	15.4	5147.4
VMWARE	2142.2	1494.6	SOA	243.7	4935.9
HOUSE	2009.3	146.9	IBM	1944.4	4341.8
IDF	1945.2	1663.6	HEALTH MINISTRY	52.9	4122.3
IBM	1944.4	4341.8	BUCS	82.4	4000.3
MCKINSEY	1726.6	1059.5	ACORN	109.3	3967.0
HET	1641.5	1610.4	USAF	115.4	3965.7

Table 3: Top ten Organisations by average daily occurrence and by trend strength in blogs

Top Occurring			Top Trending		
Noun	Avg per Day	Max Trend	Noun	Avg per Day	Max Trend
OBAMA	34008.6	64.1	BEHAR	16.4	7119.4
MCCAIN	14677.7	53.3	KWAME KILPATRICK	517.2	6698.0
JOHN MCCAIN	12280.5	68.7	ALICE COOPER	63.2	6345.6
JACK	5415.7	3098.2	FREEMAN	111.9	6155.8
JESUS	3924.7	1994.1	CORSI	166.0	5619.6
JENSEN	2661.4	1675.2	MRS. CLINTON	71.3	5575.4
RYAN	2163.1	2375.9	OLMERT	134.1	5238.0
DAVID	2156.5	1503.5	SANTANA	72.1	5127.6
PETER	1703.3	2613.4	BUFFETT	93.4	5101.8
GOD	1688.5	2767.8	CARL ICAHN	37.0	5028.9

Table 4: Top ten Persons by average daily occurrence and by trend strength in blogs

Top Occurring			Top Trending		
Noun	Avg per Day	Max Trend	Noun	Avg per Day	Max Trend
NEW YORK	6267.8	67.2	GOLD COAST	10.6	4005.6
INDIA	6188.3	70.4	BISHKEK	103.6	3912.0
UK	4146.9	293.8	LIMA	358.6	3683.3
FRANCE	3269.7	56.4	WELLS	18.6	3674.3
FLORIDA	3207.1	69.3	WIRED.COM	91.3	3632.4
DELHI	3171.2	1805.5	HAMPTON	30.4	3630.7
IRAN	2755.7	269.6	CALCUTTA	54.8	3613.2
ISRAEL	2750.3	371.8	HULU	167.2	3491.8
LOS ANGELES	2320.7	74.1	YUNNAN	64.9	3463.6
PARIS	2110.3	183.4	TRIPOLI	180.4	3421.1

Table 5: Top ten Locations by average daily occurrence and by trend strength in blogs

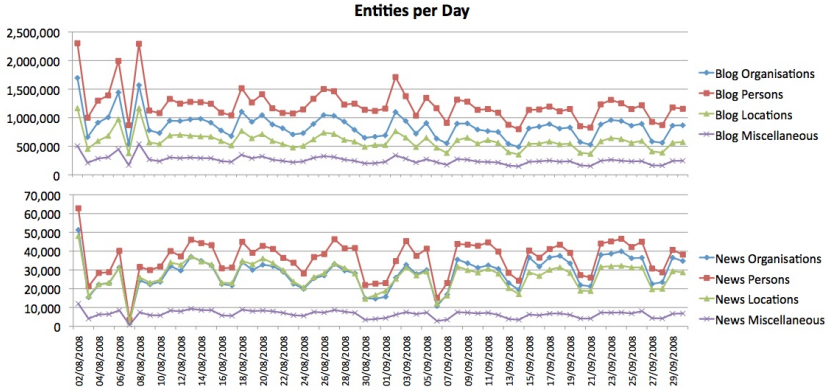

Figure 3: Entities in blogs and news per day in ICWSM 2009 corpus

Top Occurring			Top Trending		
Noun	Avg per Day	Max Trend	Noun	Avg per Day	Max Trend
INTERNET	6428.6	85.4	ALPS	42.8	2523.1
WINDOWS	1882.7	1416.4	GENESIS	76.9	2430.3
DEMOCRAT	1669.7	78.0	LITTLE LEAGUE WORLD SERIES	49.4	2263.7
GMT	1479.2	1183.7	SUMMER OLYMPIC	11.0	2050.8
ALS	1267.9	865.3	TEAM	21.3	2029.0
FACEBOOK	1217.0	1385.2	VIETNAM WAR	76.5	1928.7
TWITTER	791.6	1029.4	BOLIVARIAN ALTERNATIVE	3.9	1927.5
CHRISTMAS	786.4	1174.1	BRITONS	71.9	1772.3
JAVA	745.7	923.8	SERIE A	18.3	1765.6
MUSLIMS	703.2	665.0	CHINA OPEN	62.6	1696.2

Table 6: Top ten Miscellaneous by average daily occurrence and by trend strength in blogs

people tend to appear in trends more strongly than organisations and places, as well as showing higher average daily occurrences.

To get a sense of any linkage between average daily occurrence and maximum trends strengths in social media originated trends, the two can be plotted against one another. Distributions can be further divided into those features that are unique to language seen in social media, that which is also seen in news articles and those that also trend in news articles after a trend is seen originating from blogs. Plots for nouns and each entity type are shown in Figure 4. Also shown are the mean and standard deviation of the distributions in log occurrences and log maximum trend strength.

Trending features that are unique to blogs tend to be fewer and weaker than those that also appear in news vocabulary, although the separation is greatest for nouns. However, the presence of these for nouns at all may be for the reasons of noise and tagger errors described above. Many trending named entities occurring uniquely in blogs have average daily occurrence of less than ten per day.

Features that show trends in news after the original trend in social media tend to be the most

	Unique		Blog trending		Subsequent News trend	
	Log Occurences		Log Occurences		Log Occurences	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Nouns	0.6976	0.8714	1.7296	0.7543	2.8397	0.6804
Misc	-0.2744	0.9127	0.3864	0.7157	1.4419	0.7532
Locations	-0.1258	0.9134	0.4861	0.7241	1.5045	0.8019
Persons	0.0159	0.8519	0.7401	0.8558	1.7236	0.6577
Organisation	0.0118	0.8254	0.6590	0.7933	1.6546	0.6888
	Log Max Dev		Log Max Dev		Log Max Dev	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Nouns	3.3728	0.2587	3.9692	0.1692	3.8430	0.4191
Misc	2.4378	0.2214	2.5897	0.2414	2.9137	0.2129
Locations	2.5991	0.2104	2.7958	0.2299	3.0807	0.2757
Persons	2.7170	0.2390	2.9716	0.2125	3.3035	0.1868
Organisation	2.6515	0.2465	2.9165	0.2141	3.2103	0.1797

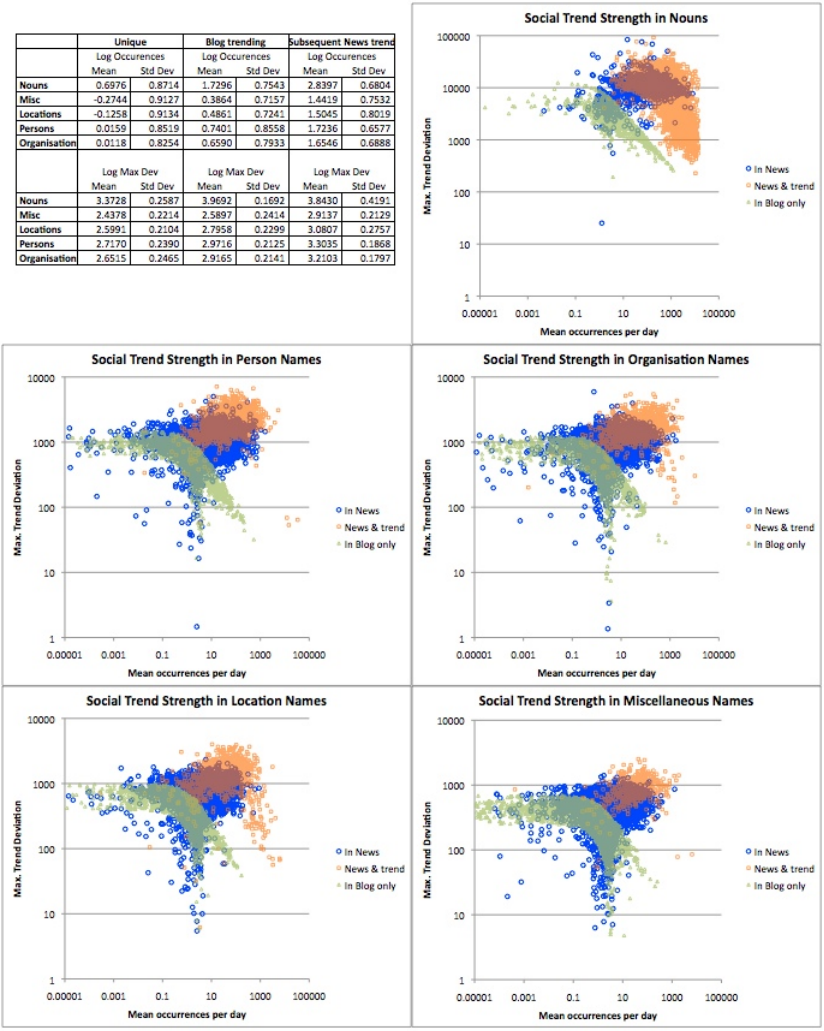

Figure 4: Distributions of occurrence per day and trend strengths for trends originating in blogs

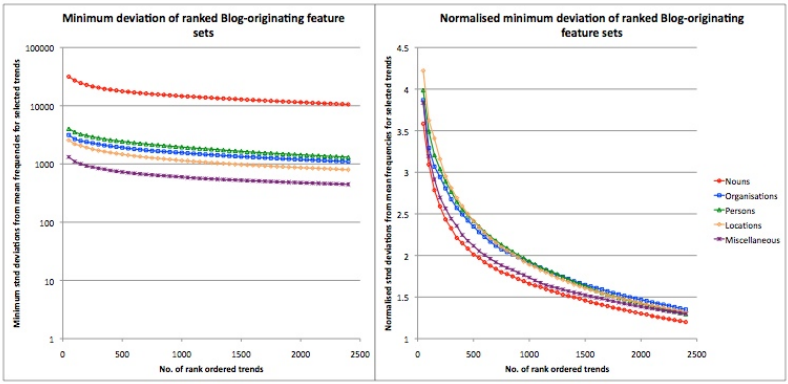

Figure 5: (a) Minimum trend strength of top n trending features (b) Minimum of top n ranked normalised trend strengths

frequently occurring. Within these features, entities also tend to show higher trend strengths and, while some nouns show higher trend strengths, nouns overall have a similar spread in trend strength as those not trending subsequently in the news: A separation in distributions of trend strength for features later trending in news and those not, is not present. Overall these distributions have significant overlap with those of corresponding feature types appearing in news articles without subsequent trends therein. The vast majority of those features showing subsequent trends in news articles have an average occurrence of at least one mention per day at the time of the trend.

These distributions suggest that entities being written about by bloggers that may be of wider interest at any particular time, tend to show trend strengths of a few hundred standard deviations from their average daily occurrence, although this can be less for very common entities (those with daily occurrence in excess of 1,000). Strengths for nouns in topics of potential wider interest tend to be an order of magnitude higher (average daily occurrence also being about an order of magnitude higher). However, this magnitude difference in trend strength is also true for nouns not subsequently trending in news articles. This suggests that comparisons between feature types would be better made having normalised by the average trend strength within a feature type.

A comparison of maximum trend strengths in feature types given the top n trending features is shown in Figure 5: Graph (a) shows the raw trend strengths while Graph (b) shows the normalised trend strengths. Note that normalisation of trend strength by feature type average de-emphasises the dominance of nouns, while the relative difference between entity types shows little change, although Locations have slightly more prominence.

In a monitoring application, it is likely one would wish to select only the most significant trending features. This suggests applying a threshold to observed trend strength. Graph (a) in Figure 6 shows the number of features that would be selected from this corpus given a normalised feature trend strength threshold. Each feature type is plotted separately as well as

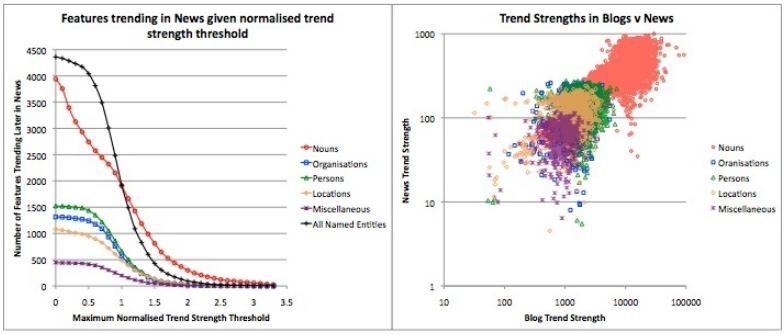

Figure 6: (a) Number of features selected for given normalised trend strength threshold (b) Relative trend strengths for those seen first in blogs and subsequently in news

for combined named entities. Note here that although the total number of trending entities outnumbers that for nouns, the spread of entity trend strengths has a narrower spread that for them. Figure 6 Graph (b) shows the un-normalised trend strength for features that also show a later trend in mainstream news articles, against that subsequent news trend strength. This shows that there is some correlation between the strength shown in the news trend and that shown in the original social media trend. If trending in news stories is indicative of wider interest then this suggests that trending features originating in social media are quite likely to be of topical appeal. Furthermore, as we have seen, many of these features are likely to be named entities.

5 Conclusions

This initial study into the type of language in trends originating in social media has shown that although much that is discussed by bloggers is whatever is currently topical in mainstream media, there is a significant amount of material of wider interest that originates in blogs. Furthermore it suggests that a significant proportion of this material may be linked to that which is later topical in news articles. The amount of material produced by bloggers is approximately 20 times greater in number of articles than professional news organisations, and the amount of individual nouns and named entities suggest their postings are also longer. Size of vocabulary is also much greater amongst social bloggers than within the mainstream media (although some of this one would consider to be erroneous or “noisy” text). There is great potential, then, for finding material in social media that is of wider interest.

Although maximum trend strengths shown for nouns are considerably greater than those shown in named entities, named entities are marginally more frequent in social media originated trends. Higher trend strengths are displayed by those features that are seen, and particularly later trend, in news articles, although these strengths are relative to the distribution seen for the feature type. Nouns trends that will later trend in news articles are not separable by trend strength alone. Selecting trends purely by highest trend strength is unlikely to be be optimal, therefore, as many trending entities of potential interest may be missed. A better strategy for

selecting trends likely to be indicative of topics of wider interest would be to select the strongest trends within classes of nouns and named entities, and possibly applying appropriate thresholds. Normalisation of trend strength by average class type trend strength may be another possibility, as this seems to make trend scores for feature types more comparable. Normalised trend scores show a narrower distribution around the mean score for entities that subsequently trend in news stories than nouns, suggesting that a threshold could be effectively applied in deciding what should be considered a genuinely trending feature.

If we believe that news stories have a wide interest, then these results suggest it is more likely that the trending feature in social media is a named entity than a noun. (Even though the very strongest trend strengths seem to be displayed by nouns.) The identification and analysis of named entities as separate features to detect trends in is, therefore, potentially of great benefit when seeking to find emerging topics of interest. The Named Entity Recogniser was not trained or tuned for social media, but rather well prepared newswire text. One would expect errors to occur both in recognition of named entities and in mis-typing of detected entities, and some errors were observed. However, a sufficiently high accuracy for differences in trends to be detected was observed. The extent to which named entity detection and recognition performance may impact remains to be determined.

Given that there are a significant number of trends originating from social media, it is natural to ask whether one can predict which will go one to be subjects in the news, and what the delay between social media interest and mainstream media interest is. Further work may also focus on determining the topic(s) named entities are involved in. These are areas for future study.

References

[1] J. Allan, V. Lavrenko, and H. Jin. First story detection in tdt is hard. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 374–381, New York, NY, USA, 2000. ACM.

[2] F. Alvanaki, M. Sebastian, K. Ramamritham, and G. Weikum. Enblogue: emergent topic detection in web 2.0 streams. In *Proceedings of the 2011 international conference on Management of data*, SIGMOD '11, pages 1271–1274, New York, NY, USA, 2011. ACM.

[3] S. Asur, B. A. Huberman, G. Szabó, and C. Wang. Trends in social media : Persistence and decay. *CoRR*, abs/1102.1402, 2011.

[4] S. Azzam, K. Humphreys, and R. Gaizauskas. Using coreference chains for text summarization. In *CorefApp '99: Proceedings of the Workshop on Coreference and its Applications*, pages 77–84, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[5] M. Baroni and S. Evert. Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 904–911, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[6] S. Bowman and C. Willis. We media: How audiences are sharing the future of news and information. Technical report, The Media Center at the American Press Institute, 2003.

[7] K. Burton, A. Java, and I. Soboroff. The ICWSM 2009 Spinn3r Dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA, May 2009. AAAI. http://icwsm.org/2009/data/.

[8] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[9] M. Cha, J. Antonio, N. Pérez, and H. Haddadi. Flash floods and ripples: The spread of media content through the blogosphere. In *ICWSM 2009: Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*. AAAI, 2009.

[10] K. W. Church. Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p2. In *Proceedings of the 18th conference on Computational linguistics*, pages 180–186, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[11] A. Feng and J. Allan. Finding and linking incidents in news. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 821–830, New York, NY, USA, 2007. ACM.

[12] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[13] L. Franco and H. Kawai. News detection in the blogosphere: Two approaches based on structure and content analysis. 2010.

[14] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 482–490, New York, NY, USA, 2004. ACM.

[15] N. S. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. ACM, 2004.

[16] V. Ha-Thuc and P. Srinivasan. Topic models and a revisit of text-related applications. In *PIKM '08: Proceeding of the 2nd PhD workshop on Information and knowledge management*, pages 25–32, New York, NY, USA, 2008. ACM.

[17] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Min. Knowl. Discov.*, 7(4):373–397, 2003.

[18] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM.

[19] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks, 2010.

[20] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, New York, NY, USA, 2009. ACM.

[21] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop. In *AAAI spring symposium on Computational Approaches to Analyzing Weblogs*, pages 117–124, 2006.

[22] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[23] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 446–453, New York, NY, USA, 2004. ACM.

[24] 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, Los Angeles, California, June 2010. Association for Computational Linguistics.

[25] H. Saggion, E. Barker, R. Gaizauskas, and J. Foster. Integrating nlp tools to support information access to news archives. In *Proceedings of the 5th International conference on Recent Advances in Natural Language Processing (RANLP)*, 2005.

[26] A. Sarkar, P. H. Garthwaite, and A. De Roeck. A bayesian mixture model for term re-occurrence and burstiness. In *CONLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 48–55, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[27] M. Simmons, L. Adamic, and E. Adar. Memes online: Extracted, subtracted, injected, and recollected. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[28] P. Thompson and C. Dozier. Name searching and information retrieval. In *In Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, pages 134–140, 1997.

[29] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[30] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 688–693, New York, NY, USA, 2002. ACM.