Finding Potential News from Trends Originating in the Blogosphere

Nigel Dewdney

University of Sheffield
Department of Computer Science
Sheffield. S1 4DP
acp08njd@sheffield.ac.uk

Abstract. Tracking current population interests by trends in online media of entities and topics has become increasingly popular. But while notable world events often spur online public discussion, some have been observed originating in social media postings. A natural question arises: Can analysis of social media trends be used to find mainstream newsworthy material? The work reported here takes initial steps towards answering this by investigating whether some characteristics of trending nouns and entities originating in blogs could predict a subsequent trend in mainstream news. Results show that many trends do originate in blogs, with approximately 12% seen subsequently in news media. Frequency based ranking provides a basis for selecting the most predictive trends. The study also suggests that named entity mentions, and co-occurrences thereof, may provide more focused trends than common nouns.

1 Introduction

The internet contains a wealth of information that is growing and changing all the time. Increasingly, information is not only provided by professional organisations via dedicated websites, but also the public at large via forums, newsgroups, social networking pages and web logs, or "blogs".. Individuals constantly add to the information (and possibly mis-information) on the web. Often these blogs reflect current circumstances for the authors. Could this wealth of material provide *new* interesting news worthy information?

Possession of the most up to date information can be of key commercial or strategic advantage in negotiation and decision making: "Knowledge is power" as the old saying goes. Knowledge of developing situations before they are widely known would therefore be of key advantage. There have in recent years been several well publicised occasions where social media has essentially broken news stories before the mainstream media. News of the Haitian earthquake of 2010 was first alerted to the world at large by messages on Twitter, and more recently the role of social media in supporting the "Arab Spring" popular uprisings are both examples where key indications were available online before being made available by the main stream news agencies. It would be desirable, therefore, to have an automated method of monitoring the social media for the unexpected.

The tracking of popular entities and topics in online media has become an increasingly common way to gain an insight into what populations are concerned about or interested in at any particular time. However, trends may result from current and recent events reported by the mainstream media outlets, or be of no significance to the wider world. We may assume that stories that have been published by the main stream media have been judged to be of interest. Ideally one would wish for new trending topics to be identified that are *not* linked to current or recent news stories. The requirement is for emerging topics to be identified that the user is not already aware of or can easily find in mainstream news outlets.

A step towards being able to find novel interesting information would be to investigate which trends in social media subsequently get picked up by the mainstream: how prevalent are they and could they be distinguished?

This paper reports on an analysis of trends in social media that do not have concurrent trends or coverage in the mainstream press . No attempt is made at this stage to characterise topics that trends relate to, or which will be picked up by news media. Neither does the study seek to quantify the proportion of news that occurs first in social media (others have examined this latter aspect). The aim here is to establish whether a significant proportion of trending features originating from social media could be used in news material selection or prediction, and whether there is advantage in using named entities.

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; section 5 examines feature specificity through bi-grams of trends; finally section 6 gives conclusions and outlines future work.

2 Related work

Traditionally news story production has been dominated by professional sources and research focussed on news detection and tracking. For example see the TREC evaluations that ran a number of novelty detection tracks [1].

A popular approach has been to look for bursts of activity as indications of emerging interest in a topic: Kleinberg [2] in looking for time gaps in term occurrence in email data found bursts in topics coincided with interest to the author; bursts of linking activity have been observed by Kumar et al. [3] in the evolution of the "Blogosphere"; Gabrilovich et al. [4] have investigated the applicability of several distance metrics in finding novel information; Franco and Kawai [5] have investigated two approaches to detecting emerging news in blogs, by measuring linking evolution and by clustering the content of postings. Ha-Thuc and Srinivasan [6] have investigated using a log-likelihood estimate of an event within a topic model as an intensity metric; and Glance et al. [7] have examined bursts in phrases, mentions of people, and hyperlinks in blogs given a background of blogs published in the preceding two weeks. They hypothesise that product mentions in blogs may have predictive power for product success.

More recently work on emerging topic and trend detection has focussed on data from the micro-blogging web service Twitter, in which messages are restricted to 140 characters, and has been likened to chat rather than publication by Alvanaki et al. [8]. In their system, named "En Blogue", they detect emerging topics by considering pairs of tags (augmented by extracted entities) at least one of which is frequent. Twitter provides its own proprietary trending topics service, but others have sought to provide similar functionality, e.g. [9], [10], and Benhardus [11] has compared different term weighting methods in Twitter trend detection.

The use of social media to predict future trends would seem to be a natural area for investigation given the establishment of evolving trends therein. Predicting the future from social media has seen interest in sectors such as movie commercial success - for example [12] - and political success - Tumasjan et al. [13] have investigated whether trends in party mentions can predict election outcomes. However prediction of topic popularity beyond news stories has not seen much investigation.

Research has also looked trend evolution in social media and how content spreads: Cha et al. [14] have studied the propagation of media content through the Blogosphere social network; Simmons et al. [15] have examined how quoted text changes as it is communicated; Clough et al. [16] have investigated whether a news story dependence on news agency text can be measured; while Asur et al. [17] have examined how trends persist and decay through social media, noting that many originate from providers such as CNN.

Evidence that social media content could pre-empt publication in the mainstream began to emerge in 2006. Lloyd et al. [18] in comparing the most popular named entities in news and blogs on a mentions-per-day basis found a small percentage of topics discussed in blogs existed before corresponding news-stories were published. A similar two-stream method to examine the characteristics of named entity trends and nouns that originated in blogs is followed here. In a different approach Leskovec et al. [19] looked at the evolution of "memes", or short phrases, Although the majority of quotation was found in blogs, around 3.5% "meme" transfer was from blog entries to news, indicating social media origins. It is the type of material found to occur by these studies that is of interest here.

3 Data and analytic approach

Topics are often about tangible (named) entities: Azzam et al. [20] suggested that a document be about a central entity, and there is evidence that names can be effective in information retrieval tasks [21], [22]. This work investigates nouns and four types of entity as potentially trending features.

For data, the ICWSM-2009 [23], is used. The dataset, provided by Spinn3r.com, is a set of 44 million blog posts and news stories made between August 1st and October 1st, 2008. The data was pre-processed for the experiments here: Blog posts that have been classified either as "MAINSTREAM NEWS" or "WEBLOG" are extracted, while all others are discarded. Posts are not reliably

language tagged,: no language filtering is applied, thereby maximising recall, which may also result in cross-language name reference inclusion. English part-of-speech tagging and named entity recognition are applied using the Stanford CoreNLP toolset without any modification [24] [25]. Data is grouped by day.

A standard Poisson model is employed for each feature frequency: This 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. If a random variable X has a Poisson distribution with expectation $E[X] = \lambda$ then

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

The mean frequency is simply the inverse of the expected gap between occurrences for the feature k, $1/\lambda$. The variance of the Poisson distribution is also λ. A trend is detected as a significant positive deviation from the expected distribution, its strength measured as the number of standard deviations in the associated gap reduction. 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:

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

Frequencies for features are observed over a day. The daily trend in feature occurrence is measured in standard deviations given by $T(k)$ from the expected frequency. Expected frequencies are estimated by averaging those observed over preceding days. This does require a certain amount of "burn-in" time to establish a reasonable estimate of the average frequency $1/\lambda_X$ for each feature $X = 1, 2, ...$ Feature frequencies are calculated, therefore, on a daily basis while average frequencies are calculated on an accumulative one, i.e. no "window" is applied. (In a larger study a rolling interval would be more appropriate to account for long term drifts in language use.) Counts are calculated for each feature in each media category and Laplacian smoothing applied to account for unseen (new) features.

We may expect small changes in occurrences to be more significant for features that do not occur very often than for high frequency features. The use of Poisson models parameterised by feature counts for each stream affords relative trend measurements, so that deviations in feature frequencies that have different means in each media stream can be compared. Similarly, deviations in the frequencies of different features can be compared.

4 Trend Analysis

For each experiment a bedding-in time of seven days was arbitrarily chosen. Thereafter feature counts were calculated for each day and compared to the rolling average, reporting at the start of the next day. On any one day in the experimental period the top trending features by deviation from their average

daily frequency to date are selected subject to a minimum of ten standard deviations above the average, and having more than five occurrences on the day. These thresholds were chosen to minimise trending due to movement caused by natural variation and to have some resilience to poor frequency estimation for very rare features. Any feature trending in the news is tagged and is not considered for the following seven days. If a news-trended feature does not trend in the news for a period of seven days it is then re-established as potentially blog-trending.

Feature types examined are those tagged as Nouns, Person names, Place names, Organisation names and Miscellaneous entities (others of undermined type). The number of trending features that were unique to blogs in the period and the number that subsequently trended in the main stream news media are shown in Table 1. Features may have trended more than once during the period but are only counted once.

Table 1. No. of trends originating in blogs & subsequently trending in news

Type	Trenders	News post trend	%
Nouns	9450	1741	18.4%
Misc	4350	221	5.1%
Location	4650	571	12.3%
Person	5450	740	13.6%
Organisation	5250	589	11.2%
Totals	29150	3862	13.2%

Blog trending features may not necessarily trend subsequently in the news, or even appear at all. Ranking by trend strength would seem to be a natural choice for prioritising and selecting trends as likely to trend subsequently in the news. However, as shown in Figure 1 graphs (a..d), having separated trending features into classes of those that subsequently trend in news, those that appear without trending in news, and those that appear uniquely in blog vocabulary, one finds that ranking by trend strength , measured in std. deviations, favours vocabulary that is unique to blogs. Ranking by feature frequency on the day of the trend, however, favours vocabulary that appears in the news, including that which subsequently trends. Note that ranking by frequency does not yield a proportion of subsequently trending features that is proportional to the number of ranked features. This is more marked for nouns than named entities.

Graphs (e) and (f) in Figure 1 show the proportions of feature type broken out from subsequently trending features and blog-unique features respectively. Miscellaneous entities are most likely to be unique to blogs, while Locations are the most likely to trend subsequently in news for the highest ranking features. However, overall, subsequent news-trending nouns are more likely to be selected than any particular entity type while being least likely to be a feature unique to blogs.

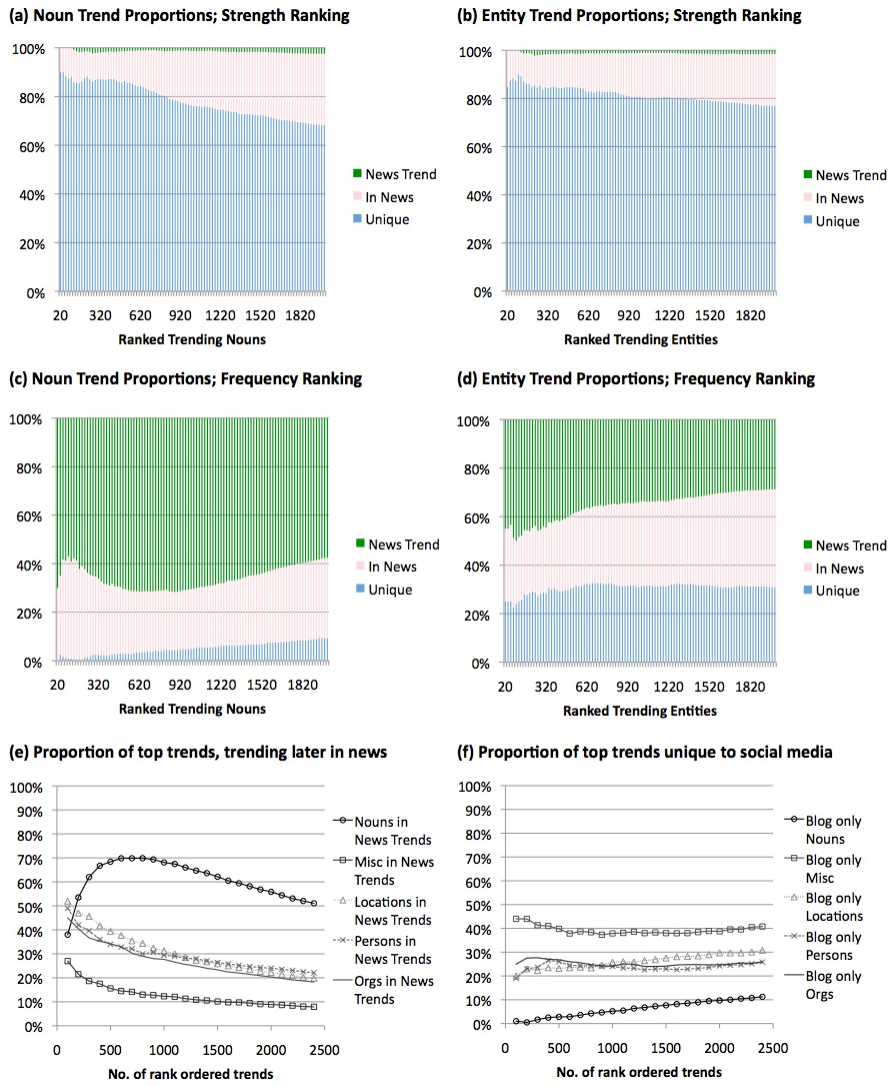

Fig. 1. Proportions of feature types in future news trends when ranked: (a) Nouns by trend strength, (b) Entities by trend strength, (c) Nouns by raw count, (d) Entities by raw count. (e) Prediction of news trend precision with raw count ranking. (f) Blog-unique features with raw count ranking.

For each feature type the top fifty blogs trends, having not trended in the previous seven days of news stories, were further examined. Note that this represents only a minority of all trends in social media, but the most significant for the purposes here. Although only the most significant trends originating in social media are examined, one should not expect all trends to be reflected in

subsequent news stories; they may simply not reflect news-worthy material, or have been overlooked by media organisations. This is indeed the case with on average 38% of blog trending features (46% of top fifty by type) seen subsequently trending in news.

The number of mentions of selected terms in weblogs per day is greater than that in news stories. However, as the counts have not been normalised for number of posts, this should not be surprising as the number of blog posts is much higher than mainstream media articles (by about 20:1 for this corpus). Blog posts mentioning top trending nouns are particularly higher in number than the number of corresponding news articles, while generally less so for named entities An explanation for this is could be the higher number in average use of nouns compared with particular named entity mentions indicating a large background use of the nouns in question. This would suggest that trending named entities are more topically specific than trending nouns.

Figure 2 shows the evolution of the top two trends for each of the feature classes. In these graphs, only positive trending behaviour is shown. The trend strength is measured in number of standard deviations from the expected value. Trend activity in news is plotted on the negative y-axis (i.e. trend strength $\times$ -1). A solid line on the plot indicates when the trend behaviour in the blog is valid, i.e. not filtered from a previous or current news trend. Note that with even within the top selections illustrated one can observe a range in feature trend patterns. It may be possible to identify some patterns for features that would allow elimination or promotion (e.g. periodic trending). This is left for future work.

Just because a feature may trend in social media and then in the news does not mean that they are topically linked. It is well known that some features, be they generic nouns or named entities, are more specific in what they refer to than others. Unsurprisingly many topics can be found in posts giving rise to trends in generic nouns seen here. For example posts featuring "design" on 9$^{\text{th}}$ August include such diverse topics as the Olympics, a design conference, website design, guild badges, and peta-scale computing to name just a few. Subsequent news stories for "design" are similarly diverse, covering jobs, product reviews, etc.

Named entities seem little better. For example, trending location entities often result from a higher than average number of posts that refer to different events from those places. For example, "Mass." arises from many posts about activities from multiple locations within Massachusetts. Subsequent news stories also have a wide spread of topics. However, some topics referring to a location are related to later news stories: "Pune" trends from posts of multiple topics including rain storms, Indian celebrities, and personal travel, but also commentary on a gang rape and murder that had occurred there. The subsequent news story is about lack of progress in the case.

Some named entities are more specific, though, and therefor topically more predictive. "Fannie" refers to the U.S. mortgage company Fannie Mae. The economic crash of 2008 had just got underway when ICWSM 2009 corpus was

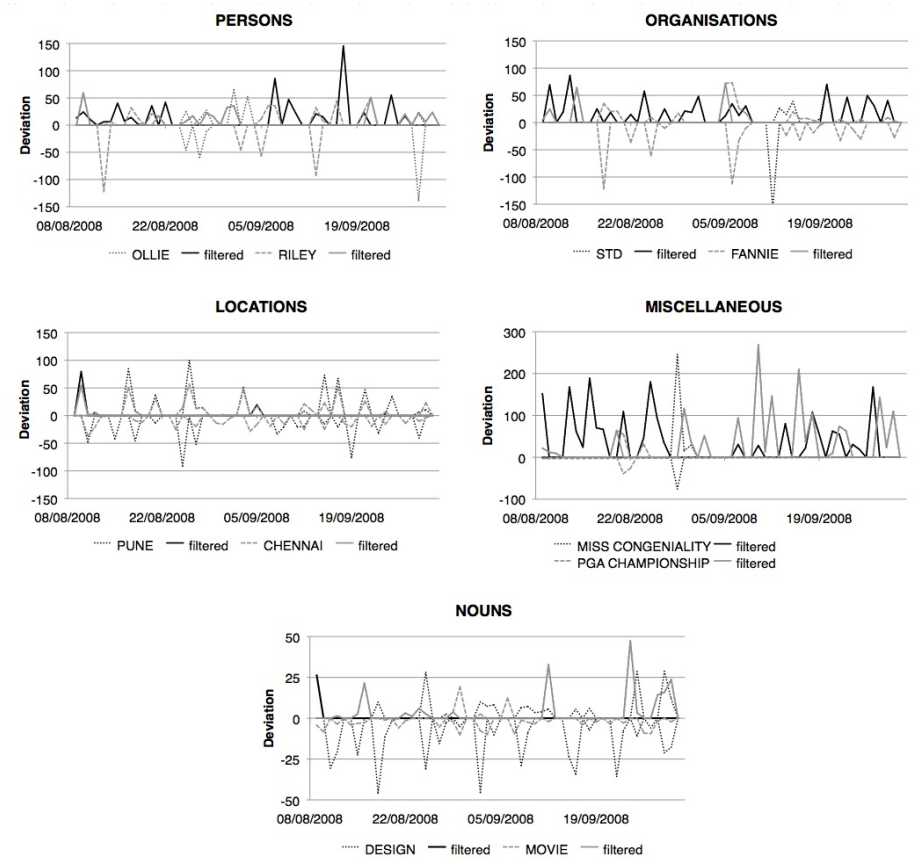

Fig. 2. Trend history for top two nouns and entity types trending in blogs prior to news. Trends (positive deviation from expected) in blogs plotted as positive deviation, and in news plotted as negative deviation.

collected, and there was much speculation about whether Fannie Mae (and its counterpart Freddie Mac) would need a U.S. government bailout. The news story trend occurred as various U.S. political figures reacted to market concerns, calling for appropriate aid, the following headline from the Chicago tribune being typical:

"U.S. plans mortgage bailout"
- http://chicagotribune.com/business/

Another example where an entity name was sufficiently specific to a topic is "PGA Championship" which referred to the 90^{th} PGA golf championship. Blogs centred on discussion surrounding players ahead of the playoffs as in:

"Jack Nicklaus Isn't Sure if Sergio Garcia Will Ever Win a Major"
- http://golf.fanhouse.com/2008/08/20/

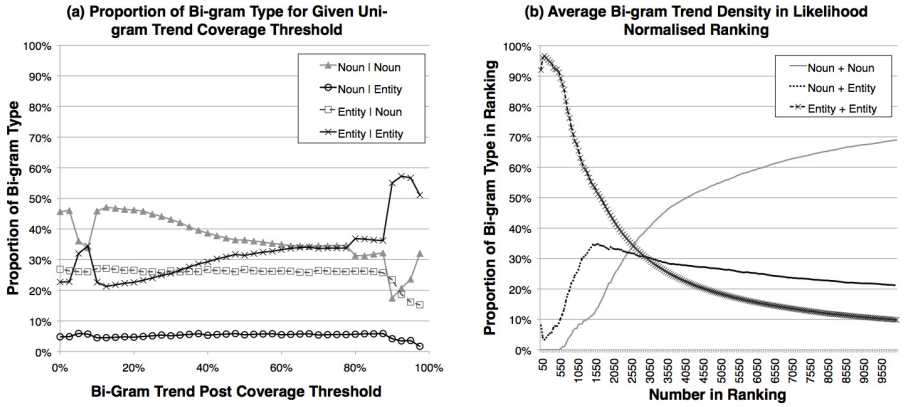

Fig. 3. Bi-gram type densities: (a) as proportion of uni-gram selected blog posts; (b) ranked by ratio of post frequency to expected frequency given independent uni-grams

while the next day news articles reported the playoffs' opening session:

"Mahan fires 62 to open PGA playoffs"
- http://sportsnet.ca/golf/2008/08/21/

"Mahan leads by 4 shots at The Barclays"
- http://cbs.sportsline.com/golf/story/10942213/rss

Feature specificity would seem to be important to find meaningful and potentially predictive information. Employing methods to refine trending features into the topics that gave rise to them is likely to be beneficial therefore.

5 Feature specificity

One may argue that to impact information about something one must express some relation between it and some other concept. The set of relationships between the key concepts, such as entities, expressed in a document could be said, then, to be the document's topic. Here we employ a similar technique to Alvanaki et al. [8] taking pairs of trending features (a trend bi-gram) to be our key concepts and examining their co-occurrence in source blog postings, the assumption being that frequently co-occurring nouns and named entities would be more likely to be linked topically.

Counts of trend co-occurrences are calculated on a by blog post basis for the day the individual trends occur. Nouns and named entity types are not examined separately here so that the analysis includes finding information across different feature types (e.g. a Person and a Location). If we assume that a topically related collection of documents would contain more than one feature in common and that for trending stories this would include more than one trending feature, we

can examine topical consistency in posts selected by a single trending feature and by a trend bi-gram. Correspondingly this will give an indication of how topically specific any single trend feature is.

A bi-gram of each trending feature and its most highly co-occurring trend for each trending feature can be created. The proportion of the uni-gram selected posts the bi-gram appears in gives an indication of how well related the posts are and therefore specificity of the unigram under an assumption of feature independence. The average proportion for bigrams of nouns was found to be 15% (variance 1.6%), but 20% (variance 6.7%) for entity bigrams suggesting entity selected posts are more likely to be topically related. Assuming that the higher the proportion of bi-grams in a sub-set of posts, the more likely that they are to be topically linked, a rank-ordering of proportion of uni-gram selected posts by bi-gram coverage allows bi-gram types to be compared against likely topical coherence. Graph (a) in Figure 3 shows that when bi-grams are present in more than 90% of the posts containing a given uni-gram trend, there is a better than 55% likelihood that both the trending features will be entities, while there is less than 31% likelihood that both trend features in the bi-gram will be nouns.

We estimate specificity by the point-wise mutual information of the bigram features at the document frequency level. This is similar to the topic coherence measure proposed by Newman et al. [26] which sums PMI in term frequency, but calculated intrinsically as in the coherence measure proposed by Mimno et al. [27]. Denoting by $S_t(a)$ the number of posts containing feature a on day t the PMI for a bi-gram feature $\{a, b\}$ is given by:

$$R(a,b) = \frac{S_t(a,b)}{S_t(a)S_t(b)} P_t \tag{3}$$

where P_t is the number of posts with at least one trending feature on day t.

Graph (b) in Figure 3 shows the proportion of each bi-gram type in an averaged rank ordering of the daily bi-gram trends. Bi-grams consisting of two entities are found predominantly at the top of the ranking, followed by those with including one entity. Assuming from the analysis above that bi-grams consisting of entities are more topically specific, this suggests that trending topics can be found by selecting posts containing bi-grams of trending entities scored by how unlikely they are to co-occur at random.

6 Conclusions

This study has considered whether there is evidence to support the idea that social media content could be used to predict or inform news stories. By filtering out trends originating in news, this study has focussed on the small percentage of topics that previous studies, [18], [19], have found *not* to have originated in mainstream news articles. It has also examined whether named entities could be more useful than common nouns as a feature for finding information.

The analysis has shown that named entities, and nouns in general, can be found trending in blog postings that have not previously trended in the mainstream media. On average approximately 12% of these features (18.4% of tagged nouns, 11.1% of tagged entities) subsequently trended in news stories (Although the total number of trending entities exceeded that of nouns). Given 3% of news stories start in social media, this suggest that either the majority of significant social media originated trends are not sufficiently interesting to professional news organisations, or are missed. However, pre-emptive trending features do exist, and simple ranking of trending features by the number of occurrences on the day of trending gives reasonable performance in promoting those that subsequently trend in news. We can conclude, then, that there is potential data in blogs with which to investigate methods for prediction of (or sourcing material for) news stories.

Improved entity tagging, filtering by language could be beneficial. Some trending entities are referred to by different forms of their names which suggests and co-reference resolution could all help in characterising and detecting trends.

Trending entities may not be the subject of topic(s) of interest, and subsequent stories may be indirectly related, on developing events, or take a new angle (a "meta-story" if you will). Examples of all these have been seen in the data examined here. However, not all subsequent trending news stories are related in any significant way to the preceding blog topics sharing the mention, and trends often arise from from multiple topics having the feature in common. These seems to be more likely the case with nouns than with named entities.

Preliminary investigation into topic specificity of trending features has shown that blog posts that are more likely to be related can be identified through co-occurring trending features, and that sets of features are more likely to contain entities than nouns. Selection of bi-grams can be achieved through a rank ordering of how unexpected its co-occurrence is.

Future work will examine whether trend selection or ranking can be improved. It will also look further at trend co-occurrence as it seems like this may be a basis for selecting posts more likely to be topically linked. Clustering of posts with trending features may also be a suitable method. These refinement methods could also assist the observer in assessing the likely interest as a news story. Such topic specificity refinement may also allow sufficient characterisation of trending topics originating in social media to make predictions as to which are likely to be picked up by news media.

References

1. Soboroff, I., Harman, D.: Novelty detection: the trec experience. In: HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 105–112. Association for Computational Linguistics, Morristown, NJ, USA (2005)
2. Kleinberg, J.: Bursty and hierarchical structure in streams. Data Min. Knowl. Discov. 7(4), 373–397 (2003)

3. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the bursty evolution of blogspace. In: Proceedings of the 12th international conference on World Wide Web. pp. 568–576. ACM, New York, NY, USA (2003)
4. Gabrilovich, E., Dumais, S., Horvitz, E.: Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In: WWW '04: Proceedings of the 13th international conference on World Wide Web. pp. 482–490. ACM, New York, NY, USA (2004)
5. Franco, L., Kawai, H.: News detection in the blogosphere: Two approaches based on structure and content analysis (2010)
6. Ha-Thuc, V., Srinivasan, P.: Topic models and a revisit of text-related applications. In: PIKM'08: Proceedings of the 2nd PhD workshop on Information and knowledge management. pp. 25–32. ACM, New York, NY, USA (2008)
7. Glance, N.S., Hurst, M., Tomokiyo, T.: Blogpulse: Automated trend discovery for weblogs. In: WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. ACM (2004)
8. Alvanaki, F., Sebastian, M., Ramamritham, K., Weikum, G.: Enblogue: emergent topic detection in web 2.0 streams. In: Proceedings of the 2011 international conference on Management of data. pp. 1271–1274. SIGMOD '11, ACM, New York, NY, USA (2011),
9. Petrović, S., Osborne, M., Lavrenko, V.: 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. pp. 181–189. Association for Computational Linguistics, Los Angeles, CA, USA (June 2010),
10. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining. pp. 4:1–4:10. MDMKDD '10, ACM, New York, NY, USA (2010),
11. Benhardus, J.: Streaming trend detection in twitter. Tech. rep. (2010)
12. Joshi, M., Das, D., Gimpel, K., Smith, N.A.: Movie reviews and revenues: an experiment in text regression. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 293–296. Association for Computational Linguistics, Stroudsburg, PA, USA (2010),
13. Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010),
14. Cha, M., Antonio, J., Pérez, N., Haddadi, H.: 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)
15. Simmons, M., Adamic, L., Adar, E.: Memes online: Extracted, subtracted, injected, and recollected. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011),

16. Clough, P., Gaizauskas, R., Piao, S.S.L., Wilks, Y.: Meter: Measuring text reuse. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 152–159. Association for Computational Linguistics, Morristown, NJ, USA (2002)
17. Asur, S., Huberman, B.A., Szabó, G., Wang, C.: Trends in social media : Persistence and decay. CoRR abs/1102.1402 (2011)
18. Lloyd, L., Kaulgud, P., Skiena, S.: Newspapers vs. blogs: Who gets the scoop. In: AAAI spring symposium on Computational Approaches to Analyzing Weblogs. pp. 117–124 (2006)
19. Leskovec, J., Backstrom, L., Kleinberg, J.: 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. pp. 497–506. ACM, New York, NY, USA (2009)
20. Azzam, S., Humphreys, K., Gaizauskas, R.: Using coreference chains for text summarization. In: CorefApp '99: Proceedings of the Workshop on Coreference and its Applications. pp. 77–84. Association for Computational Linguistics, Morristown, NJ, USA (1999)
21. Thompson, P., Dozier, C.: Name searching and information retrieval. In: In Proceedings of Second Conference on Empirical Methods in Natural Language Processing. pp. 134–140 (1997)
22. Saggion, H., Barker, E., Gaizauskas, R., Foster, J.: Integrating nlp tools to support information access to news archives. In: Proceedings of the 5th Int'l conference. on Recent Advances in Natural Language Processing (2005)
23. Burton, K., Java, A., Soboroff, I.: The ICWSM 2009 Spinn3r Dataset. In: Third Annual Conference on Weblogs and Social Media (ICWSM 2009). AAAI, San Jose, CA (May 2009), http://icwsm.org/2009/data/
24. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: 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 1. pp. 173–180. Association for Computational Linguistics, Stroudsburg, PA, USA (2003),
25. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370. Association for Computational Linguistics, Stroudsburg, PA, USA (2005),
26. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. Association for Computational Linguistics (2010)
27. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 262–272. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011),