Detecting Attention Dominating Moments Across Media Types

Igor Brigadir Derek Greene Pádraig Cunningham

{igor.brigadir, derek.greene, padraig.cunningham}@insight-centre.org
Insight Centre for Data Analytics
University College Dublin, Ireland

Abstract

In this paper we address the problem of identifying attention dominating moments in online media. We are interested in discovering moments when everyone seems to be talking about the same thing. We investigate one particular aspect of breaking news: the tendency of multiple sources to concentrate attention on a single topic, leading to a collapse in diversity of content for a period of time. In this work we show that diversity at a topic level is effective for capturing this effect in blogs, in news articles, and on Twitter. The phenomenon is present in three distinctly different media types, each with their own unique features. We describe the phenomenon using case studies relating to major news stories from September 2015.

1 Introduction

The problem of detecting breaking news events has inspired a host of approaches, extracting useful signals from activity on social networks, newswire, and other types of media. The online communication platforms that have been adopted allow these events to persist in some form. These *digital traces* can never fully capture the original experience, but offer us an opportunity to revisit significant phenomena with different points of view, or help us to characterise and learn something about the processes involved. Many different forms of news media attempt to record and disseminate information deemed important enough to communicate, and as the barriers to broadcasting and sharing information are removed, attention becomes a scarce commodity.

We define the problem of detecting *attention dominating moments* across different media types, as a collapse in diversity in the content generated by a set of online sources in a topic during a given time period. *Media types* here include mainstream news articles, blog posts, and tweets. These media types differ in both the category of topics covered [22], and their use of language [10]. In the context of Twitter, we define *sources* as unique user accounts. For mainstream news and blogs, sources refer to individual publications or outlets. Publications may have different numbers of authors, but as unique author information is not available, we treat each unique blog or news outlet as a single source.

In Section 3, we describe the two stages of our proposed event detection procedure. In the first stage, content generated by the news, blog and tweet sources is grouped into broad topical categories, through the application of matrix factorization to the content generated by these sources. In the second stage, we examine the variation in similarity between content generated by sources within a given topic during a given time period, in order to identify a collapse in diversity within a topic which corresponds to an attention dominating moment. In Section 5, we evaluate this procedure on a collection of one million news articles and blog posts from September 2015, along with a parallel corpus of tweets collected during the same time period.

Rather than formulating the problem as tracking the evolution of topics themselves, we consider the diversity of content within a specific topic over time. The motivation is that, for instance, a collapse in diversity around a major sporting event will be strongly evident in certain news sources, but not evident in others.

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

The distinction is important, as this approach is more suited to retrospective analysis, when the entire collection of documents of interest is available. The topics do not change over time, as opposed to a real-time setting where topics must be updated as new documents arrive [21]. The information need is guided by two major questions. Firstly, when have significant collapses in diversity occurred in a topic of interest? Secondly, are there differences between media types when these events occur?

Our main contributions here are: 1) a diversity-based approach of detecting attention dominating news events; 2) a comparison between traditional news sources, blogs, and Twitter during these events. 3) a parallel corpus of newsworthy tweets for the NewsIR dataset.

2 Related Work

In previous work, attention dominating news stories have been described as *media explosions* [2] or *firestorms* [14]. The idea of combining signals from multiple sources for detecting or tracking evolution of events proved effective in the past. Osborne *et al.* [16] used signals from Wikipedia page views, together with Twitter to improve "first story detection". Concurrent Wikipedia edits were used as a signal for breaking news detection in [19].

Topic modeling applied to parallel corpora of news and tweets has been previously explored by a number of researchers [6, 9, 11]. Extensions to LDA to account for tweet specific features have been proposed [22]. A comparison between Twitter and content from newswires was explored in [18]. A Non-negative Matrix Factorization (NMF) approach is used for topic detection in [20].

How offline phenomena link to bursty behaviour online is discussed in [5] and [12]. In [12] Shannon's Diversity Index was used to detect a "contraction of attention" in a tweet stream by measuring diversity of hashtags. In contrast, we employ a different measure of diversity based on document similarity, applying it to streams from different media types segmented by topic. Methods for automatically detecting anomalies or significant changes in a time series are discussed in [4]. In [15] a change-point detection approach is applied to time series constructed from Tweet keyword frequencies.

As a broad overview, the common components involved in detecting high impact, attention dominating news stories include: selecting relevant subsets of documents; representation and feature extraction; constructing time series from features; event detection and analysis. In this paper we concentrate on a single key feature of breaking news: a collapse in content diversity within a fixed time window.

3 Proposed Method

Our objective is to detect when multiple articles in a topical stream become less diverse, signalling the emergence of an attention dominating news story. We consider attention to a phenomenon as the main driving force behind the decision to produce or broadcast a communication. Using the diversity of content within a time window, we attempt to characterise instances where a particular piece of information becomes dominant. Concretely, for each type of media, NMF is used to assign topics to documents; for documents in a topic, we calculate diversity between documents in a time window. This type of analysis allows us to examine the extent to which the onset of an important breaking news event is accompanied by a collapse in textual content diversity, both within a group of news sources and across different media types.

3.1 Finding Topics

We apply a Non-negative Matrix Factorization (NMF) topic modeling approach to extract potentially interesting topics from a stream of tweets or set of articles. For each *media source*, we build a tf-idf weighted term-document matrix and use this as input to NMF.

We also considered LDA to infer topics in these datasets. The choice of NMF over LDA was primarily due to computation time. LDA was significantly more computationally expensive than NMF with NNDSVD [1] initialisation. NMF also tends to produce more coherent topics [17].

3.2 Measuring Diversity

The same tf-idf representation used for topic modeling is used in diversity calculations. Each article, blog post or tweet is a tf-idf vector. A separate document-term matrix is built for each *media type*. Stopwords and words occurring in fewer than 10 documents are removed.

To measure diversity, we calculate the mean cosine similarity between all unique pairs of articles within a topic for a fixed time window. Given a set of documents D in a time window, the diversity is:

$$diversity(D) = -\frac{\sum_{i,j \in D, i \neq j} cosSim(D_i, D_j)}{\sum_{i=1}^{|D|-1} i}$$

Where $cosSim(D_i, D_j)$ is the cosine similarity of tf-idf vectors of documents i and j in a time window. In practice, calculating similarities between all pairs of documents can be efficiently performed in parallel, and can be calculated in a matter of seconds.

Longer time windows consider more document pairs, which naturally result in smoother trends. In contrast, shorter time windows are more sensitive to brief attention dominating events, but also false positive spikes—where a small number of articles happen to be similar in content, but do not constitute an attention dominating story.

An alternative to content diversity is also considered. Ignoring document content, and just considering the sources of articles, diversity is calculated with Shannon's Diversity Index:

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

Where p_i is the proportion of documents produced by the ith source in a time window of interest, R is total number of sources in a given media type.

Both diversity measures produce a single diversity value per time window, generating a univariate time series. Changes in diversity that are 2 standard deviations away from the mean are naively considered to be important enough to warrant attention. Exploring more robust and well established methods for change point detection such as [15, 4] is left for future work.

For the case studies described in Section 5, the window length was set to 8 hours. While the fast-paced "24/7 news cycle" is described as a constant flood of information, we find that all three mediums largely follow a more traditional publishing cycle, with prominent spikes in number of published articles on weekday mornings, and low numbers of articles published outside of normal office hours. A more detailed analysis of publishing times and characteristics will be explored in future work.

4 Datasets

To explore attention dominating news stories, we apply the method described above to three media sources: mainstream news, blogs, and tweets. For the first two sources, the NewsIR dataset[1] is used. For the final source, we use our own parallel corpus collected from Twitter[2]. In contrast to previous work [6, 11] where tweets are retrieved based on keywords extracted from news articles, the parallel corpus was derived from a large set of newsworthy sources, curated by journalists [3]. Journalists on Twitter curate lists[3] of useful sources by location or general topic of interest—for example "US Politics" may contain accounts of US politicians and other journalists who tend to cover US politics related stories.

Gathering all members of such lists covering different countries and topics follows the *expert-digest* strategy from [7]. A tweet dataset collected independently of news and blog articles preserves Twitter-specific features and topics. Source and document counts are summarised in Table 1.

Media Type	*Sources*	*Documents*	*Docs. per 24h*
News	18,948	730,634	8,177
Blogs	73,403	253,488	23,568
Tweets	30,448	3,274,089	125,568

Table 1: Summary of overall source and document counts by media type after filtering, and average number of documents in a 24 hour window.

Of the original 1 million articles provided, 15,878 were filtered as non-English[4] or outside the date range of interest (*i.e.* created between 2015-09-01 and 2015-09-31). Tweet language filtering was performed using meta-data provided in the tweet.

5 Attention Dominating Events

In order to compare the same topics across different media types, we compare the top 10 terms representing the topics from different models. Specifically, when topics from two different models have strongly-overlapping (using Jaccard similarity) top term lists, this indicates that similar events were discussed in both media types.

Topics in a model that do not have any overlapping terms with topics in other models, suggest that content unique to a platform is prominent. For example: the "*live, periscope, follow, stream, updates*" topic in the tweet corpus has no equivalent among the news or blog topics. This reflects the fact that the Periscope app became popular with journalists for broadcasting short live video streams and Twitter is the main platform where these streams are announced. The "*music, album, song, video, band*" topic is prominent in the blogs and Twitter, but is not present in news. This may reflect the fact that most Twitter accounts and blogs are far more personal in nature.

An indicative, but not necessary feature of attention domination news is the presence of a similar topic on multiple platforms. To illustrate the phenomenon of topical diversity collapse, we now describe three case studies.

[1] Available from: `http://research.signalmedia.co/newsir16/signal-dataset.html`

[2] Data: `https://dx.doi.org/10.6084/m9.figshare.2074105`

[3] Examples of such lists are available `https://twitter.com/storyful/lists/` and `https://twitter.com/syflmid/lists`

[4] `https://github.com/optimaize/language-detector` was used for language detection. Interestingly, language detection proved effective for filtering "spammy" articles containing obfuscated text, large numbers of urls, or containing tabular data.

For each case study, we present the following: Top 10 topic terms for a topic in a media type, and a plot of diversity over time, where:

- Solid lines show diversity of documents over time.
- Dashed lines show Shannon Diversity of sources.
- Highlighted time periods are when major developments occurred—based on Wikipedia Current Events Portal[5] for September 2015.
- Dot and Triangle markers indicate periods when diversity drops 2 standard deviations below the mean.

5.1 European Refugee Crisis

The European crisis began in 2015, as increasing numbers of refugees from areas in Syria, Afghanistan, and Western Balkans [8] sought asylum in the EU. Figure 1 shows a plot of diversity for the documents assigned to this topic in each 8 hour time window, for the three media types. To help with visualisation, raw diversity values are standardised with z-scores on the y axis, while the x axis grid separates days.

Media	*Top 10 Topic Terms*
Blogs	refugees, syria, syrian, war, president, government, military, europe, russia, iran
News	refugees, migrants, border, hungary, eu, europe, european, refugee, asylum, germany
Tweets	refugees, syrian, hungary, help, migrants, europe, border, germany, austria, asylum

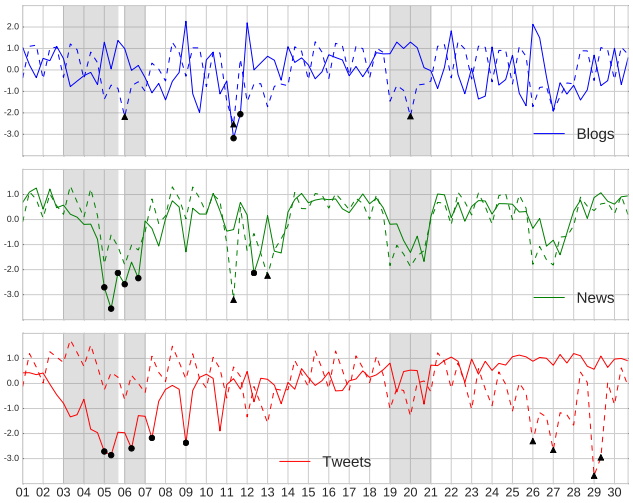

Figure 1: Standardised diversity scores for the European refugee crisis topic during September 2015, across three media types.

The downward trend in diversity between September 3rd and 5th in the refugee crisis topic can be explained by the death of Aylan Kurdi. News of his drowning quickly spread online and made global headlines. This was a particularly far-reaching story, dominating news coverage until an announcement on relaxing controls on the Austro-Hungarian border by Chancellors Faymann of Austria and Merkel of Germany. Both Twitter and mainstream news streams experienced a diversity collapse, while Blogs maintained more diverse set of articles. Between 19th and 21st, smaller drops in diversity coinside with Pope Francis' visit, where the issue of refugees was a prominent topic of discussion.

5.2 Donald Trump Presidential Campaign

Donald Trump's presidential campaign has attracted considerable attention across all types of media[6]. Positions on issues of immigration and religion are particularly polarising, frequently causing controversies in mainstream media.

Media	*Top 10 Topic Terms*
Blogs	trump, donald, republican, presidential, debate, gop, president, candidates, candidate, bush
News	trump, republican, presidential, donald, debate, clinton, bush, fiorina, candidates, campaign
Tweets	trump, im, love, donald, going, debate, happy, gop, president, think

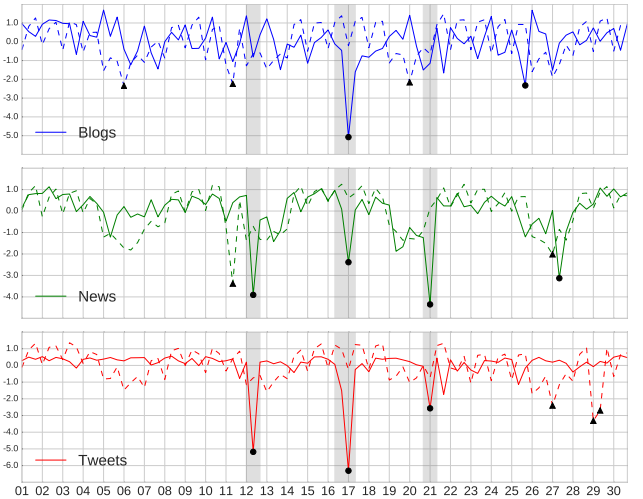

Figure 2: Standardised diversity scores for Donald Trump Presidential Campaign topic

Significant events marked around 12th, 17th, 21st in Figure 2 relate to: Trump's comments on Senator Rand Paul on Twitter which was discussed on mainstream news around 12th, but not as prominently on blogs. On the 16th-17th coverage of a republican presidential debate hosted by CNN; and 21st—mainstream news coverage of reactions to events on 17th: during

[5] https://en.wikipedia.org/wiki/Portal:Current_events/September_2015

[6] https://en.wikipedia.org/wiki/Donald_Trump_presidential_campaign,_2016

a town hall meeting in Rochester, Donald Trump declined to correct a man who said that President Obama is a Muslim.

The statement prompted a significant drop in the diversity of stories across all platforms. On the 25th, during a speech given to conservative voters in Washington, Trump called fellow Republican presidential candidate Marco Rubio "a clown". Based on the data, it appears that the reaction to the latter on Twitter was not as pronounced as among journalists and bloggers.

5.3 Pope Francis visits North America

The visit of Pope Francis spanned 19 to 27 September 2015, where the itinerary included venues in both Cuba and the United States. This event is a good illustrative example as it was widely documented[7], and highlights a case where a collapse in diversity did not occur at the same time on different media platforms.

Media	*Top 10 Topic Terms*
Blogs	pope, francis, church, catholic, visit, cuba, popes, climate, philadelphia, vatican
News	pope, francis, catholic, church, philadelphia, popes, cuba, united, vatican, visit
Tweets	pope, francis, visit, house, congress, popeindc, cuba, white, popeinphilly, philadelphia

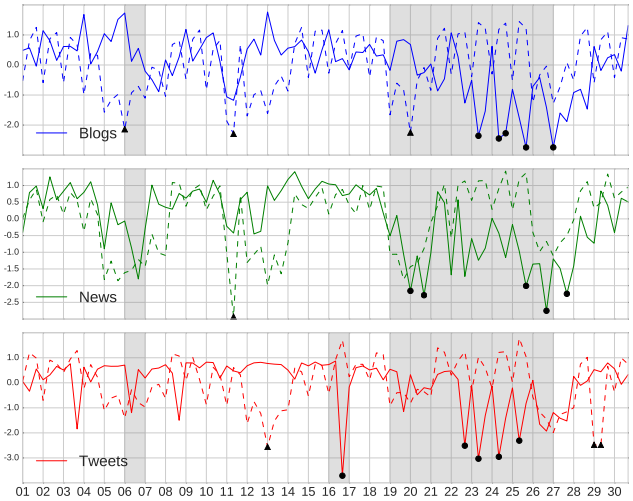

Figure 3: Standardised diversity scores for the Papal visit topic during September 2015.

In the case of news publishers, the largest drop in diversity coincided with the beginning of the Pope's visit to Havana. Twitter users and bloggers reacted more on September 23rd and 24th, when the Pope met with Barack Obama and became the first Pope to address a joint session of US Congress.

[7] https://en.wikipedia.org/wiki/Pope_Francis'_2015_visit_to_North_America

In the Twitter stream, the notable event around 16th-17th is due to large numbers of similar tweets as preparations for the visit were being discussed, and #TellThePope trended briefly.

Earlier in the month, we see evidence of overlapping attention dominating events. Between 6th and 7th September, the Pope announced the Vatican's churches will welcome families of refugees. This announcement followed a significant development in the ongoing European refugee crisis: around 6,500 refugees arrived in Vienna following Austria's and Germany's decision to waive asylum system rules. This suggests that an attention dominating news event in one topic can trigger events in other topics, especially where prominent public figures are involved.

6 Discussion

While the diversity measure we propose is relatively simple, it can be easily augmented to account for other factors. In the simplest form, every similarity value between a unique pair of articles within a time window carries an equal weight in the diversity calculation, implying that a strong similarity between two highly influential publishers is just as important as between two inconsequential publishers with a small audience. However, this weight could be tuned, either manually or automatically using external information (*e.g.* Alexa rankings). Accounting for social context [13] could also be achieved by augmenting the topic modeling stage of the process. Instead of using a classic tf-idf vector space model, alternative representations that capture more semantic similarity between documents can be used. We aim to explore extensions to this measure in future work.

The sequence of events in the European refugee crisis and papal visit case studies suggest that it may be possible to identify and track major developments with global impact by linking attention dominating moments across multiple topics, as well as across sources on different platforms. Social media communities both influence and are influenced by traditional news media [11]. Stories break both on Twitter and through traditional news publishers. Tracking or linking instances of diversity collapse to explain the direction of influence between the different media types is also a potential avenue for future work.

Acknowledgments: This publication has emanated from research conducted with the support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

[1] C. Boutsidis and E. Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4), 2008.

[2] A. E. Boydstun. *Making the news: Politics, the media, and agenda setting.* University of Chicago Press, 2013.

[3] I. Brigadir, D. Greene, and P. Cunningham. Adaptive representations for tracking breaking news on twitter. *CoRR*, abs/1403.2923, 2014.

[4] P. Esling and C. Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[5] Y. Gandica, J. Carvalho, F. S. D. Aidos, R. Lambiotte, and T. Carletti. On the origin of burstiness in human behavior: The wikipedia edits case, 2016.

[6] W. Gao, P. Li, and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In *Proc. 21st ACM international conference on Information and knowledge management*, pages 1173–1182. ACM, 2012.

[7] S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly, and K. Gummadi. On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1739–1744. ACM, 2013.

[8] E.-M. P. Giulio Sabbati and S. Saliba. Asylum in the eu: Facts and figures. *European Parliamentary Research Service*, (PE 551.332), mar 2015.

[9] Y. Hu, A. John, F. Wang, and S. Kambhampati. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *AAAI Conference on Artificial Intelligence*, 2012.

[10] Y. Hu, K. Talamadupula, and S. Kambhampati. *Dude, srsly?: The surprisingly formal nature of Twitter's language*, pages 244–253. AAAI press, 2013.

[11] T. Hua, F. Chen, C.-T. Lu, and N. Ramakrishnan. Topical analysis of interactions between news and social media. *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[12] A. Jungherr and J. Pascal. Forecasting the pulse: how deviations from regular patterns in online data can identify offline phenomena. *Internet Research*, 23(5):589–607, 2013.

[13] J. Kalyanam, A. Mantrach, D. Saez-Trumper, H. Vahabi, and G. Lanckriet. Leveraging social context for modeling topic evolution. In *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 517–526, 2015.

[14] H. Lamba, M. M. Malik, and J. Pfeffer. A tempest in a teacup? analyzing firestorms on twitter. In *Proc. International Conference on Advances in Social Networks Analysis and Mining*, pages 17–24, 2015.

[15] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-Point Detection in Time-Series Data by Relative Density-Ratio Estimation. *ArXiv e-prints*, Mar. 2012.

[16] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First story detection using twitter and wikipedia. In *SIGIR Workshop on Time-aware Information Access*, 2012.

[17] D. OCallaghan, D. Greene, J. Carthy, and P. Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645 – 5657, 2015.

[18] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton. Can twitter replace newswire for breaking news? In *Proc. 7th International Conference on Weblogs and Social Media, ICWSM*, 2013.

[19] T. Steiner, S. van Hooland, and E. Summers. Mj no more: Using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection. In *Proc. 2nnd International Conference on World Wide Web*, pages 791–794, 2013.

[20] C. K. Vaca, A. Mantrach, A. Jaimes, and M. Saerens. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd international conference on World wide web*, pages 527–538. ACM, 2014.

[21] K. Zhai and J. Boyd-Graber. Online latent dirichlet allocation with infinite vocabulary. In *Proc. 30th International Conference on Machine Learning*, pages 561–569, 2013.

[22] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.