Insight4News: Connecting News to Relevant Social Conversations

Bichen Shi, Georgiana Ifrim, and Neil Hurley

Insight Centre for Data Analytics,
University College Dublin
Dublin, Ireland
{bichen.shi,georgiana.ifrim,neil.hurley}@insight-centre.org

Abstract. We present the **Insight4News** system that connects news articles to social conversations, as echoed in microblogs such as Twitter. **Insight4News** tracks feeds from mainstream media, e.g., BBC, Irish Times, and extracts relevant topics that summarize the tweet activity around each article, recommends relevant hashtags, and presents complementary views and statistics on the tweet activity, related news articles, and timeline of the story with regard to Twitter reaction. The user can track their own news article or a topic-focused Twitter stream. While many systems tap on the social knowledge of Twitter to help users stay on top of the information wave, none is available for connecting news to relevant Twitter content on a large scale, in real time, with high precision and recall. **Insight4News** builds on our award winning Twitter topic detection approach and several machine learning components, to deliver news in a social context.

Keywords: news tracking, social media, Twitter, summarization.

1 Introduction

Famously in August 2011, news of an earthquake in Virginia, USA, reached New York by Twitter before the tremors were felt. Media stories such as the death of Michael Jackson have spread rapidly on social media in advance of breaking on the mainstream media. Nowadays, more often than not, news stories break online long before appearing in newspapers. The landscape of news delivery and dissemination has changed dramatically in less than a decade since the widespread take-up of social media. Writer and entrepreneur Chris Anderson captures this through his statement *"The ants have megaphones now."* [5].

Insight4News links news articles from mainstream media (e.g., BBC), to relevant Twitter conversations, as delivered by tweets, hashtags and automatically detected events. It builds on our prior work [1] for automatically mining social media streams to provide users with a set of headlines summarizing the most important topics discussed over a given time period. Our topic detection approach was assessed by practicing journalists and ranked first as the most effective "news miner" with regards to several evaluation criteria, amongst which

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 473–476, 2014.

were precision and recall. Furthermore, **Insight4News** provides social context for news articles via a machine learning algorithm that classifies and ranks hashtags [2] and provides a timeline of each article with respect to relevant tweets, hashtags, topics and photos.

Since Twitter has become very popular as a channel for citizen-driven media, many systems aim to tap this resource. Storyful [7] is a social media news agency that tracks and curates Twitter content for breaking news and potential stories for newsrooms. The focus is mostly on content curation and licensing. The *headlines* feature of Twitter provides links to articles that relate to a specific tweet. The Tweeted Times promises a personal newspaper, by aggregating news from the Twitter stream and ranking them by popularity among the users' friends. Storify [8] provides a service where users can manually search for topics of interests on several platforms (Twitter, Facebook, YouTube) in order to add social context to a story. Blews [4] focuses on political news by tracking blogs and the articles they cite, tagging each article with the number of blogs citing it, political orientation and emotional charge of those blogs. Hash2News [9] takes a hashtag as input and presents relevant news articles for that hashtag. Most of these systems go from the social media to the news, via the urls shared in tweets. This drastically reduces recall, since many related tweets do not post the url explicitly. We propose a system that combines both directions, by connecting mainstream media news articles to the relevant social feeds on Twitter, while allowing the user to track in real-time the newsworthy topics directly from the Twitter stream. Most readers are interested in other people's opinion on the topics, the connections between articles and the development of stories, *"users crave more relevant news with deeper contextualization"* [6]. **Insight4News** aims to fulfil this need. This is also an important step in the development of new digital journalism support tools. It is expected that such tools will become commonplace in the newsroom of the near future [10].

2 The Insight4News System

In this section, we present the key components of **Insight4News** [3], as illustrated in Figure 1. The system is written in Python3 with the Django web framework and is deployed on an Apache web server. Celery, a distributed task queue, is used for back end data collection and processing.

Data Collection and Processing. We poll 14 RSS news feeds every 15 minutes (currently from BBC and Irish Times), covering international and local news. We extract the urls and retrieve the articles (around 400 daily). We automatically extract representative keywords for each article, pool keywords for all articles, and feed them to the Twitter Streaming API, constraining each retrieved tweet to contain at least two article keywords (for more details see [2]). On average, we get about 500k tweets per day. Each article's keywords are streamed for 24 hours, by updating the all-article-keywords-list every 5 minutes. This step aims to retrieve a large set of relevant tweets without being restricted to a set of manually curated user lists, locations or article urls. Via shallow matching of

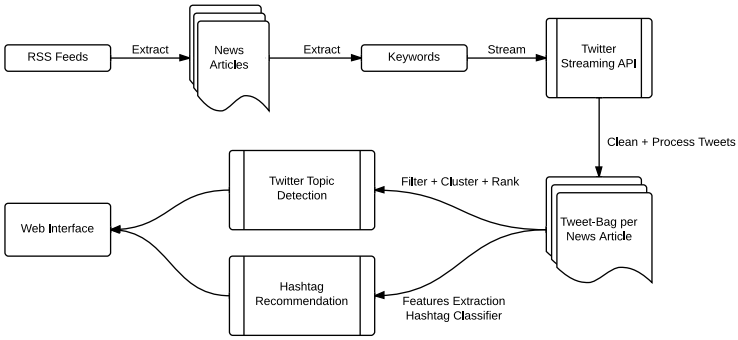

Fig. 1. High-level overview of the **Insight4News** system

tweet and article keywords, we get a local tweet-bag per article which we use for topic detection and hashtag recommendation (recomputed every 5 mins).

Topic Extraction. Based on our award winning aproach [1], this stage relies on tweet-clustering combined with a few layers of filtering, aggregation and ranking. The detailed steps include hierarchical clustering of tweets, time-dependent n-gram and cluster ranking and headlines re-clustering. For each article tweet-bag, we execute these steps to obtain a set of headlines or topics that summarize the tweet activity relevant to the article. This approach also works on an arbitrary Twitter stream (not article-focused).

Hashtag Recommendation. We pose this as a learning problem. Using the tweets-bag per article, we form article-hashtag pairs, and compute four features for each pair that capture the global (whole stream) and local (article tweet-bag) profile of the hashtag (wrt popularity and relevance). To train a hashtag classifier, we use 2,500 manually labeled article-hashtag pairs. A good source of relevant hashtags that does not require manual effort, are user tweets that post the article url and hashtags for that article. We use a Logistic Regression classifier with 87% Precision and 79% Recall from [2]. The classifier provides a score describing the likelihood that a hashtag is relevant to the article, which we use to rank hashtags for each article, and recommend the top10 hashtags with classification score above 0.5.

The **web interface** currently has 6 views. **TrackedNews** shows the latest news articles, with headline, number of tweets retrieved and published time. **ArticleDetails** shows content and social context, recommended hashtags and extracted topics from related tweets. The default hashtag view is the classifier result, while 3 other views show top10 hashtags based on frequency, recency, and cosine similarity between hashtag and article profiles. **TrackYourArticleHere** allows the user to track their own news article by providing a valid url or the full text. **TrackYourTopicsHere** allows the user to track events by providing keywords (e.g., Russia, Ukraine, EU). **PopularHashtags** lists the top50 most popular hashtags in the last 24 hours. **HashtagDetails** shows a specific hashtag,

its definition as retrieved from http://tagdef.com, and co-occurring hashtags. A plot shows the activity of the hashtag in the last 30 hours. A group of articles and a tweet stream related-to/filtered-by the hashtag are also shown.

3 Insight4News Use Case

On May 13, 2014, an explosion in a coal mine in Soma, Turkey killed at least 280 people. **Insight4News** captures a series of articles on this sad story. *BBC: Turkey coal mine disaster: Desperate search at Soma pit* is one such example, with around 5k related tweets. Clicking on this headline leads to **ArticleDetails**, which shows top10 hashtags, including #PRAYFORSOMA, #Turkey, #turkeymine, #WorkersRights. Top10 topics extracted from the article's tweet-bag are shown (e.g., **1.** *Mourners by miners' graves in the western town of Soma. The toll from Turkey's worst mining accident is now 282.*, **2.** *Bayram Ilki poured water on the grave of his son Saban, as many other coal miners were buried nearby.*). A photo is shown beside each topic, which can be expanded to show the tweets summarized by this topic-headline. Top10 learned hashtags are shown by default. Clicking on #Soma leads to the hashtag definition and related hashtags (#PrayForTurkey, #Protests). A plot shows the activity for #Soma in the last 30 hours. Below the plot, there are articles related to #Soma. One of the articles *Turkish mine disaster prompts violent protests* is a follow up news. It is therefore useful to look at news articles related to the same story by following related hashtags. On the right side of the plot we show the most recent 100 tweets for the current hashtag.

In the future we intend to scale **Insight4News** to more RSS news feeds and tap into additional social media (e.g., Facebook, Reddit) as sources of social context for news.

Acknowledgements. This work was supported by Science Foundation Ireland under grant 07/CE/I1147 (Insight Centre for Data Analytics).

References

1. Ifrim, G., Shi, B., Brigadir, I.: Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering. In: Proceedings of SNOW 2014 Data Challenge, WWW (2014)
2. Shi, B., Ifrim, G., Hurley, N.: Be in The Know: Connecting News Articles and Relevant Twitter Conversations. arXiv:1405.3117 (2014)
3. http://insight4news.ucd.ie/insight4news/
4. Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., König, A.C.: BLEWS: Using Blogs to Provide Context for News Articles. In: ICWSM. AAAI Press (2008)
5. http://www.thetimes.co.uk/tto/business/article2118052.ece
6. http://www.forbes.com/sites/oreillymedia/2012/05/09/the-future-of-the-newspaper/
7. http://storyful.com/products/newsrooms
8. https://storify.com/tour
9. http://hujo.deri.ie/hujo-newshack-ii/
10. http://www.goatmustbefed.com/resources/pdf/goat-must-be-fed.pdf