FLORIN – A System to Support (Near) Real-Time Applications on User Generated Content on Daily News

Qingyuan Liu, Eduard C. Dragut
Computer and Information Sciences
Temple University
{tuf28438, edragut}@temple.edu

Arjun Mukherjee
Computer Science Department
University of Houston
arjun@cs.uh.edu

Weiyi Meng
Computer Science Department
Binghamton University
meng@cs.binghamton.edu

ABSTRACT

In this paper, we propose a system, FLORIN, which provides support for near real-time applications on user generated content on daily news. FLORIN continuously crawls news outlets for articles and user comments accompanying them. It attaches the articles and comments to daily event stories. It identifies the opinionated content in user comments and performs named entity recognition on news articles. All these pieces of information are organized hierarchically and exportable to other applications. Multiple applications can be built on this data. We have implemented a sentiment analysis system that runs on top of it.

1. INTRODUCTION

Social media are becoming instrumental tools in measuring public opinion and predicting social behavior. Public opinions expressed in social media outlets (e.g., newsgroups, blogs, social networks) are good indicators of various political, social, and economic variables. Previous research showed that the public sentiment on election related topics in social media (e.g., Twitter, Instagram) correlated well with the polling results for the 2008 and 2012 presidential elections [21], and 2009 Obama job approval [3]. Prior works in [18, 19] showed that even stock markets have some correlations with social sentiment. Web search is another dimension of social response, which includes two notable studies by Google: (1) Flu Trends [7] which aims to predict flu outbreaks by tracking Web search behavior, and (2) Early assessments of the end of 2008 economic recession by measuring the drop of users' searches for unemployment benefits [20]. Several related systems have also been developed. For example, Court-O-Meter tracks the opinion on topics about political discussions and Supreme Court on Twitter [22], while that by Godbole et al. [8] tracks opinions expressed in news articles for events by analyzing their content. In both systems the topics of interest are manually set and do not analyze the public reaction on the articles via user comments on articles in real-time.

While all the above works have made progresses, they are rather limited in scale, timespan, and heavily tailored for a specific application. In this demo, we render a platform that continuously "listens" to the public reaction to the news on social and monitors the *social sentiment and the social behavioral responses towards world events* in near real-time. In the long run, the goal is to improve the speed at which changes in social, economic, and political events and conditions can be detected and quantified, leveraging the trend of past social responses to various events. The system leverages methodologies from big-data analysis [6], natural language processing [2, 16], information extraction and retrieval [4, 15, 17] to deliver a new data-driven paradigm for transforming/facilitating public opinion research. The platform would facilitate a number of practical applications, including but not limited to: (1) forecasting opinion trends, (2) opinion surveillance to assess social unrest, intolerance and extremism, (3) real-time analysis of social sentiment on evolving world events, (4) generating time-series data for causal-effects analysis of events, (5) measuring event projection and veracity verification, and (6) truthfulness analysis of fact statements posted on the Web [12]. In Section 3, we show a sentiment analysis application that was developed on the platform.

Proceedings of the VLDB Endowment, Vol. 8, No. 12

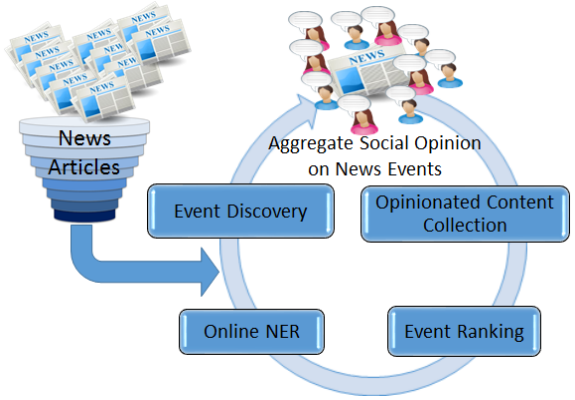

Figure 1: Information flow and main components of FLORIN

2. SYSTEM DESCRIPTION

In this section we describe the key components of FLORIN along with the information flow in the system (Figure 1). The system is designed to handle multiple events concomitantly in a near real-time and continuous manner. Although for the most part of this paper, we detail the system components during one iteration, we will point out places where such behavior may be affected by the results of previous iterations, for instance, avoiding processing duplicate events or comments in subsequent iterations.

2.1 Event Discovery

This component discovers interesting events from news articles that attract user engagement via comments. An *event* is broadly interpreted as an issue of public interest (e.g., Ebola breakout, Boston bombing). While several event detection algorithms have been proposed [1] (survey), [9] (tutorial), they do not scale well for our near real-time needs. Hence, we took a different approach. We implemented a web crawler that continuously crawls Google News for

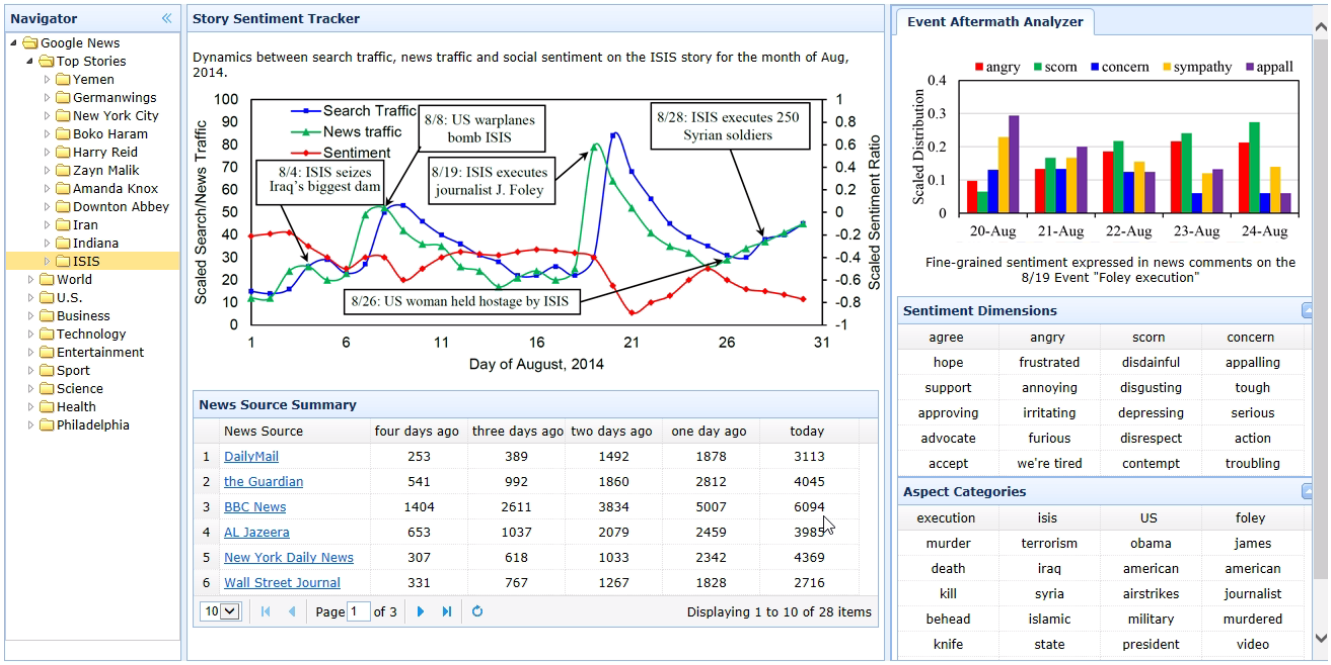

Figure 2. System screenshot featuring the ISIS event. The graph in the center illustrates the relationship between search traffic, news traffic and public sentiment for the month of Aug, 2014 covering the story of terrorist group ISIS with specific event annotations. The graph on the right illustrates Foley 8/19 Event aftermath analysis via relative distribution of fine-grained sentiment expressed in news comments.

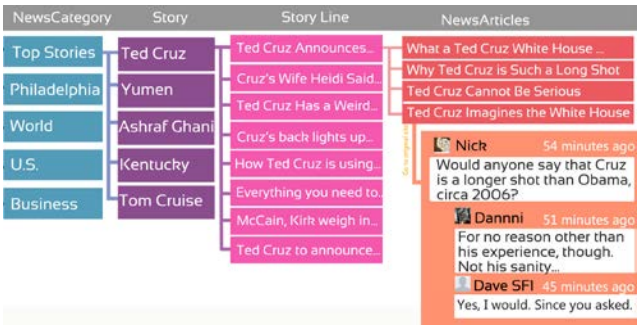

Figure 3: Data structure for extracted data from daily news.

the top stories published on a number of major news media websites (e.g., BBC, Al Jazeera, The Guardian) and discovers potential events of interest based on the news traffic volume. The set of events collected by the system consists of (1) the top news stories in Google News and (2) the top-k (k = 10, 20) most frequent n-grams (n = 1, 2, 3, and 4) from the headline/article titles (they appear in the left hand side pane in a Web browser, Figure 2). In our experience, this approach discovers events quickly and does not require crawling the entire content of the news articles. The list of candidate events is continuously refined over time. We draw attention to an interesting problem here: different terms may refer to the same event and sometimes the same term may be used to refer to different events. To address this problem, we cluster the event terms based on co-occurrence and semantic distance. Therefore, FLORIN represents an event e as a (soft) clustering of semantically related terms extracted from news titles (e.g., {ISIS, Foley, American journalist} represents the 8/19 event "Foley execution"). We denote $E = \{e\}$ as the set of events generated by this component.

2.2 Opinionated Content Collection

We aim to collect two types of opinionated content, news article comments and social media posts. In the current implementation of FLORIN, we only support the former. This alone gives us hundreds of thousands of user generated content per event. We observed this for sensitive events (e.g., Foley execution; Ferguson shooting). Most major news websites allow users to contribute comments on news articles that help us gauge the public sentiment on events. We attempt to collect all relevant user comments, R_e, from each news article a belonging to event e as follows.

Upon crawling an article a for an event $e \in E$, we parse its source to mine certain HTML constructs indicating the presence of user comments. (Note that not all articles on news website are opened to user comments.) We use a number of heuristics to detect whether an article is accompanied by user comments and look for patterns such as "%more%comments", "%comments%count%", or "(post|write)%[new]%comment", in the HTLM source code. '%' stands for "Any string of zero or more characters." '[ ]' stands for optional, whiles '$(x|y)$' stands for "Match x or y." For each comment, we gather all comment metadata (e.g., user details, posting time, time relative to article's original publishing time, content). For each article, we collect its publication time [13] and also preserve the hierarchical comment thread structure of replies (Figure 3). Currently, FLORIN collects comments from six major news outlets: BBC.com, aljazeera.com, nydailynews.com, theguardian.com, dailymail.com and wsj.com.

2.3 Event Ranking

Public engagement on events varies over time/geo-location. Events such as the protest in Ferguson, MO, the conflict in Ukraine, or ISIS capture the front pages of news outlets and social responses for many months, while events such as the Super Bowl or Oscars seize

people's attention for only few weeks/days. Hence, FLORIN maintains a list of "current" events from which a user can select to monitor. As the first version of our implementation, the list is limited to only 20 entries at all times. (Recall, that for each event we collect all posted news articles via Google News and the associated user comments.) The ranking is a weighted linear expression between three factors: (1) the number of news outlets covering the event, (2) the total time spent on being featured in Google News, and (3) the rate of user comments. (1) and (3) are computed relative to a time interval, which can be configured by the user. The options are: *days* or a time interval specified by the user. (1) is computed from the snippets in Google News without visiting the actual sources. (2) is computed from the times of the articles stored in our database. And, (3) is computed based on the rate of user generated comments in the interval given by the user. In the current implementation, these factors are all equally weighted.

2.4 Online Named Entity Recognition (NER)

This component generates a set of initial *seed named entities* to be used by other applications, including sentiment analysis, which is included in this demo (Section 3). Despite extensive work in NER [14] (tutorial), its accuracy is far from perfect for Web text. Most existing methods are not amenable to real-time needs as they depend on Part-of-Speech (POS) tagging, or Wikipedia, which are expensive and unsuitable for online processing [11].

FLORIN includes an efficient one-scan mining based on the following hypotheses. (1) Many named entities relevant to an event e appear in news articles with an initial capital letter as news articles follow proper English. (2) The comments on an article a tend to re-mention the named entities from a. (3) The set of event e's articles share a large portion of the entities relevant to e. These hypotheses allow us to discover a large initial set of the named entities relevant to the event e and even to perform a first-hand (soft) clustering of synonymous entity mentions (e.g., Foley, James Foley, and Jim Foley). Applications built on top of our infrastructure can further refine the list of named entities produced by FLORIN. We give such an example in Section 4.1.

2.5 System Implementation

FLORIN is implemented using a combination of technologies: MySQL, Java, JSON, JSOUP, and Open Flash Chart[1]. Figure 2 shows our system screenshot featuring the ISIS event.

3. SENTIMENT ANALYSIS APPLICATION

We leverage the real-time news comment data collected for two sentiment analysis applications: (1) Story polar sentiment tracking with event annotations, and (2) Fine-grained sentiment analysis of event aftermath. For (1), we leverage the event set discovered in Section 2.1 on a story and the user comments on news articles belonging to an article to track the social sentiment responses to various events in that story. We record the daily sentiment for each event and obtain a data-point at the end of each day. The data points are aggregated over a certain period of time (in this case, a month) with event annotations.

As an illustration, we analyzed the social response to the terrorist group ISIS during August 2014. Figure 2 (center) shows the time-series of the search traffic and news traffic of "ISIS" from Google Trends and Google News respectively. Search traffic refers to the number of searches. News traffic refers to the number of sources (e.g., news articles, blogs) covering the story. Both news and search traffics are rescaled to $[0,100]$ for a trend comparison. We also report the daily sentiment score s_t obtained from user comments across news sites (Section 2.2). For day t, s_t is defined as the ratio of the count of positive words to the count of negative words in user comments on the story. The value of s_t is scaled to $[-1,1]$ where values close to $+1/-1$ indicate posts dominated by positive/negative opinions, respectively. In sentiment analysis, each opinionated word conveys a general positive (e.g., great, awesome, nice) or negative (e.g., poor, worse, pathetic) polarity. Our sentiment analysis application system uses our sentiment lexicons in [5, 10].

Comment 1: I sincerely hope and advocate Obama to take action and order to target these ISIS miscreants. Completely agree with US airstrikes in Iraq.

Comment 2: I support Obama's policy of not paying ransom and approving airstrikes and military action. I'd agree that this doesn't solve terrorism. But we're tired and frustrated of islamic extremists and terrorism.

Comment 3: James Foley execution is disgusting. US must take action. Terrorism is a serious concern and threat.

Figure 4. Analysis example of user opinions on the disturbing news event of Foley execution. We show the estimated aspects and sentiments (Tables in Figure 2) using different colors.

4. DEMONSTRATION OVERVIEW

The goal of our demo plan is to demonstrate to visitors that FLORIN is a step forward in the ongoing effort of the scientific community to build tools that measure public opinion and predict social behavior using large-scale social media content. We describe here several demo scenarios we will prepare for our audience that showcase FLORIN's data collection and aggregation capabilities, as well as the two sentiment analysis applications developed on top of them.

4.1 Off-Line Data Demo Scenario

In the off-line demo scenario, we will identify a number of recent events of large public interest, crawl and organize them for the demo session. These will be used to promote lively interaction with our audience. Here we give an example based on the data we crawled during the month August 2014 for the event ISIS terrorism. Figure 2 depicts the data for this event.

Search Traffic versus News Traffic. On the plots rendered in our tool, we will show the directional trend of the search traffic volume for the terms of an event (e.g., ISIS) showing that they match well with the trend of the news traffic. For example, for ISIS we have a correlation coefficient of $\rho = 0.71$. We can also show the correlation between the news traffic and the social media traffic ($\rho = 0.78$ for ISIS, not shown in Figure 2).

We will also present cases where the search traffic lags the news traffic. For ISIS, this can be seen in the specific peak events annotated in the plot. This is plausible as the users' search activity is often influenced by the news.

Sentiment Analysis. We will display and comment on the outcome of our sentiment analysis tools. For ISIS, the overall sentiment for the coverage of the story is negative with the relative rise and fall. This is evidenced by the sharp plummeting of the sentiment score on the events occurring on 8/8, 8/19, 8/26, and 8/28. Thus, using the data collected by FLORIN during that period, we find that the public sentiment accurately reflects the actual events taking place.

[1] http://teethgrinder.co.uk/open-flash-chart-2/

The sentiment time-series has a slight lag ($\approx$ 1 day) as it takes a few hours before comments are accumulated.

We will identify fine nuances revealed by our sentiment analysis tools. For instance, though it may appear surprising, the 8/8 event of US performing military action on ISIS to combat terrorism, was perceived negatively by the public. This result of our prototype actually corroborates with the national polls conducted by Gallup and Pollingreport (see details in Section 4.3).

Fine-Grained Sentiment Analysis. Sentiment is much more expressive than just a polar value. Hence, for the selected events we will present an event aftermath fine-grained sentiment analysis leveraging the large-scale comment data per event. We will show how the result of soft-clustering performed in FLORIN for each event $e \in E$ are employed as seeds to perform fine-grained aspect and aspect specific sentiment extraction (bottom-right table in Figure 2). Specifically, we employ our models in [16] to discover aspects and fine-grained sentiment categories for an event. Figure 4 shows a few representative comments of the case of 8/19 event of Foley execution. Fitting the sentiment models for the stream of comments on an event facilitates (near) real-time event aftermath sentiment analysis over time. In Figure 2 (right), we plot the relative distribution of the multi-dimensional sentiment expressed in comments following the event for 5 days. We can see that the initial reaction to the Foley execution event was *concern*, *sympathy*, and *appall* which subsided gradually with the rise of *anger*, *scorn* over the 5-day aftermath analysis. This task has a latency of 4-6 hours before daily histograms are generated upon receiving the previous day's full comment set. The latency is due to the time needed to fit our statistical models in [16] to the large-scale comment dataset.

4.2 Live Data Demo Scenario

In this demo scenario, we will set up a live data collection and walk the audience through the setup steps of FLORIN, showing live how the data is collected, stored in the backend database, the key main memory data structures and how the temporal variables (e.g., of the articles, comments and events) are updated. We invite the audience to pick an event to be motorized. We will describe the steps to explore the sentiment on events. Specifically, the fitting of a previously learned sentiment model on large news data that allows for fine-grained modeling under real-time needs will be explained.

4.3 Validation via Third-Party Entities

One interesting demonstration scenario is to showcase instances where the results obtained with our data and tools corroborate with that of third-party entities over the same events and time. We describe here the validation conducted on the ISIS terrorism case study. A key reason for choosing to present it throughout the paper is because we were able to corroborate some of our results with independent survey data from Gallup and Pollingreport. For instance, the question "do you favor or oppose US military action and sending ground troops to fight ISIS" appears in both Gallup[2] and Pollingreport[3] (with different paraphrasing). To evaluate our results, we compute the relative scores of sentiment dimensions "agreement" vs. "disagreement" towards the aspect "military action". It is worthwhile to note that our prototype (Figure 2) shows a dip in the overall sentiment on the 8/8 event "US warplanes bomb ISIS" that corroborates with the polling result of ~60% opposition in Gallup and Polling report (see footnotes 2, 3).

To demonstrate to our audience the robustness of our data collection along with the analytics tools developed on it, besides the ISIS case study, we will prepare a number of (off-line) case studies that will be demonstrated and discussed with our audience. We aim to present an interactive and inquisitive demo session for each visitor. Given the interest in collecting and analyzing the social media content nowadays, we believe that our system will result in a lively demo session.

5. REFERENCES

[1] Atefeh, F. and Khreich, W. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*. 31, 1, 2015.

[2] Chen, Z., Mukherjee, A. and Liu, B. Aspect Extraction with Automated Prior Knowledge Learning. In *ACL*. 2014.

[3] Cottle, S. Media and the Arab uprisings of 2011: Research notes. In *Journalism*. 12, 5, 2011, 647–659.

[4] Dragut, E., Meng, W. and Yu, C. *Deep Web Query Interface Understanding and Integration*. Morgan & Claypool Publishers. 2012

[5] Dragut, E.C., Yu, C., Sistla, P. and Meng, W. Construction of a sentimental word dictionary. In *CIKM*. 2010, 1761–1764.

[6] Gamon, M., Mukherjee, A. and Pantel, P. Predicting interesting things in text. *In COLING*. 2014.

[7] Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. and Brilliant, L. Detecting influenza epidemics using search engine query data. In *Nature*. 457, 2009, 1012–1014.

[8] Godbole, N., Srinivasaiah, M. and Skiena, S. Large-Scale Sentiment Analysis for News and Blogs. In ICWSM. 2007.

[9] Gupta, M., Li, R. and Chang, K.C.-C. Towards a Social Media Analytics Platform: Event Detection and User Profiling for Twitter. In *WWW*. 2014, 193–194.

[10] Hu, M. and Liu, B. Mining and summarizing customer reviews. *In SIGKDD*. 2004.

[11] Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A. and Lee, B.-S. TwiNER: Named Entity Recognition in Targeted Twitter Stream. In *SIGIR*. 2012. 721–730.

[12] Li, X., Meng, W. and Yu, C.Y. T-verifier: Verifying Truthfulness of Fact Statements. In *ICDE*. 2011, 63–74.

[13] Lu, Y., Meng, W., Zhang, W., Liu, K.-L. and Yu, C.T. Automatic Extraction of Publication Time from News Search Results. In *ICDE Workshops*. 2006.

[14] Meij, E., Balog, K. and Odijk, D. Entity Linking and Retrieval Tutorial. In *SIGIR*. 2013, 1127.

[15] Meng, W. and Yu, C.T. *Advanced Metasearch Engine Technology*. Morgan & Claypool Publishers. 2010

[16] Mukherjee, A. and Liu, B. Aspect Extraction through Semi-Supervised Modeling. In *ACL*. 2012.

[17] Mukherjee, A. and Liu, B. Mining Contentions from Discussions and Debates. In *SIGKDD*. 2012.

[18] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X. Exploiting Topic based Twitter Sentiment for Stock Prediction. *In ACL*. 2013.

[19] Si, J., Mukherjee, A., Liu, B., et al. Exploiting Social Relations and Sentiment for Stock Prediction. In EMNLP. 2014.

[20] Varian, H.R. and Choi, H. Predicting the Present with Google Trends. In *SSRN Electronic Journal*. 2009.

[21] Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. (Jul. 2012), 115–120.

[22] Wang, Y., Clark, T., et al. Towards Tracking Political Sentiment through Microblog. In ACL Workshops. 2014.

[2] http://www.gallup.com/poll/177263/slightly-fewer-back-isis-military-action-past-actions.aspx

[3] http://www.pollingreport.com/iraq.htm