JMLR: Workshop and Conference Proceedings 36:182–197, 2014 BIGMINE 2014

From Tweets to Stories: Using Stream-Dashboard to weave the twitter data stream into dynamic cluster models

Basheer Hawwash BHAWWASH@GMAIL.COM
Olfa Nasraoui OLFA.NASRAOUI@LOUISVILLE.EDU
University of Louisville, Louisville, KY

Editors: Wei Fan, Albert Bifet, Qiang Yang and Philip Yu

Abstract

Social media has recently emerged as an invaluable source of information for decision making. Social media information reflects the interests of virtual communities in a spontaneous and timely manner. The need to understand the massive streams of data generated by social media platforms, such as Twitter and Facebook, has motivated researchers to use machine learning techniques to try to discover knowledge in real time. In this paper, we adapt our recently developed stream cluster mining, tracking and validation framework, Stream-Dashboard, to support detecting and tracking evolving discussion clusters in Twitter. The effectiveness of Stream-Dashboard in telling stories is illustrated by analyzing a couple of stories related to the Louisville Cardinals' basketball championship. We further validate the detected story lines, that are automatically mined from user-generated tweets using as an alternative source, Google Trends, which are based on search queries.
Keywords: Social Media, Data Stream Clustering, Visualization

1. Introduction

Social media can be described as a set of platforms that allows users to share information in virtual communities in many forms, including short texts (e.g. Twitter and Facebook), photos, videos, blogs...etc. Some of the social media's main features, that distinguish it from traditional data sources, include immediacy, frequency and reach, to name a few. These characteristics made social media an attractive source of data for many researchers and organizations, who saw a game-changing opportunity to improve their productivity and efficacy.

Several applications use social media to help solve real-world problems, including education Grosseck and Holotescu (2008), finance Bollen and Mao (2011), disaster relief Gao et al. (2011) and marketing Evans (2010). However, the very same characteristics that have made social media attractive, have also made it challenging, motivating researchers to turn to machine learning for help Lin and Kolcz (2012); Pennacchiotti and Popescu (2011). Machine learning, in this arena, can discover useful knowledge by identifying hidden patterns or relationships from social media-related data.

One of the most important social media sources is Twitter, a micro-blogging website used by millions of users. Twitter is a perfect and accurate example of what is considered a Big Data platform, where millions of tweets are generated every day. Twitter generates a massive and continuous data stream of tweets with an estimated 500 million tweets generated per day

[1]. The real time, informal and spontaneous nature of the tweets have made Twitter attractive for decision makers, for instance to extract users' interests and opinions Jansen et al. (2009); Bifet and Frank (2010); Mendoza et al. (2010). What distinguishes Twitter from other social media platforms is its immediacy, which made it attractive to many low-latency applications such as natural disaster detection Sakaki et al. (2010). Due to the massive amounts of data generated by Twitter, it became imperative to use advanced machine learning methods to summarize and identify interesting patterns or topics on the fly. To this date, research has focused more on detecting trending topics, and not enough on tracking the evolution of these topics over time.

In this paper, we apply Stream-Dashboard Hawwash and Nasraoui (2012) on Twitter data to detect and track evolving topics over time. More specifically, we cast the discovered knowledge in a similar framework to telling stories, where stories are first detected, and then tracked over time. Furthermore, the main milestones or major changes, affecting the stories' evolution, are identified and quantified automatically. One contribution of this work is mainly to discover and present trending clusters in Twitter messages in a format that is similar to stories, where we are inspired by picture story books. To this extent, we define a twitter story in a very simplified way, as a set of events in a chronological order defined by a set of pictures and metrics that describe those events. For each of the stories, we define its start and end times, as well as major events that are crucial to understanding the story. Although the notion of stories that we show here are very simple, they constitute a required step toward richer future story telling efforts, because what we present, was extracted in a completely automated manner, and in a completely unsupervised manner (no Human guidance or previous training data). We further validate the detected story lines, that are automatically mined from user-generated tweets using as an alternative source, Google Trends, which are based on search queries.

The rest of this paper is organized as follow: Section 2 will present some background on Stream-Dashboard and other techniques used in this paper. Section 3 will illustrate the use of Stream-Dashboard on Twitter, and finally we conclude in Section 4.

2. Background

2.1. Stream-Dashboard

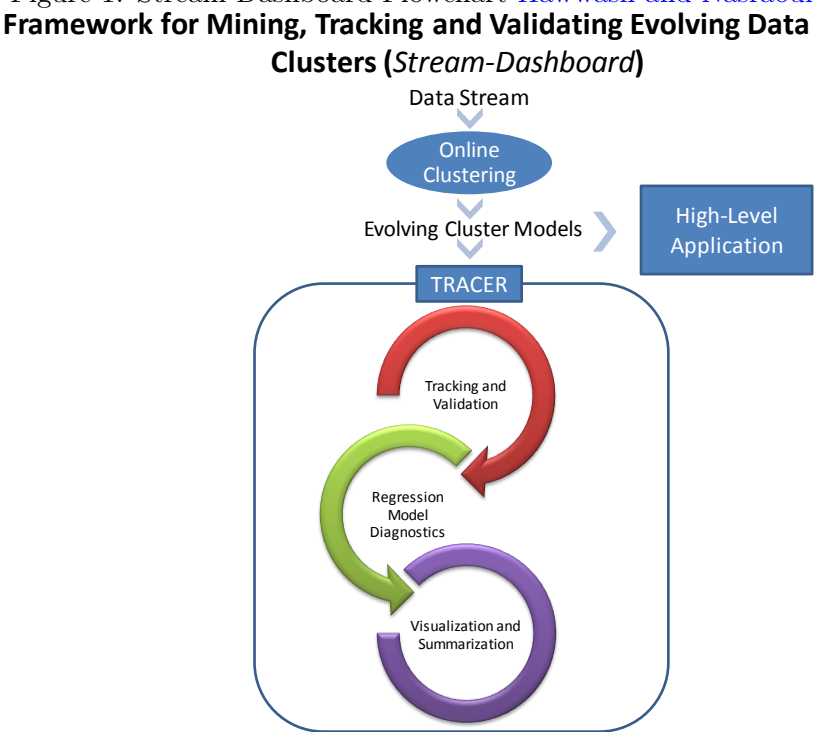
Stream-Dashboard Hawwash and Nasraoui (2012) is a complete framework to simultaneously *mine*, *track* and *validate* clusters in big data streams. It consists of two main components: an *online clustering* component and a *tracking and validation* component (TRACER).

The online clustering component can be any stream clustering algorithm that incrementally maintains a clustering model of the data stream (in this paper, posts on Twitter), and generates a set of properties or metrics, describing each cluster (in this paper, clusters correspond to topics).

The tracking and validation component (TRACER) monitors the characteristics of the clustering model (i.e. the properties of the clusters) and builds and maintains *regression models* for them. The output of TRACER can be used to track and visualize the evolution of clusters over the data stream lifetime, as well as an input to another application, such as

1. https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

Figure 1: Stream Dashboard Flowchart Hawwash and Nasraoui (2012)



an online recommender system, that needs to adapt to concept drift. Stream-Dashboard is shown in Figure 1.

2.2. RINO-Streams

As the first component in Stream-Dashboard, we use RINO-Streams Hawwash and Nasraoui (2010), an incremental density-based stream clustering algorithm. RINO-Streams extracts evolving clusters from a massive data stream in one pass, while also detecting and resisting outliers. It incrementally updates the clustering model using an estimation of centroids and scales rooted in robust statistics Ricardo A. Maronna (2006). Moreover, it detects outliers and merges clusters using a robust distribution-independent statistical test, called the Chebyshev test Marshall and Olkin (1960), which ensures robustness to outliers and cluster compactness.

3. Twitter Story Teller

In this section, we illustrate how Stream-Dashboard can be used as an apparatus to tell stories inferred from a stream of tweets, by going through the fundamental KDD stages of data preparation, mining, and examination of the results.

3.1. Data Preparation

Twitter provides an Application Programming Interface (API)[2] which allows collecting tweets by third party users. The free API, which is limited to a 1% sample of all the tweets, was used to collect tweets starting from October 2011. We collect tweets for 15 minutes every hour and store them in a database. We did not use any filtering keywords when initially querying the API, thus generating a wild collection of random tweets. However, for some of the experiments shown below, we will use term-based filtering from the collected set to focus on certain events.

3.1.1. Pre-processing

After storing the raw tweets in the database, we perform the following pre-processing steps:

1. Detecting the language of the tweet and keeping only the English-written tweets
2. Cleaning the tweets by:
   (a) extracting the web links, user names and hash tags and storing them in the database,
   (b) removing non-English characters and digits
3. Removing stop words
4. Lemmatization of words [3]
5. Removing tweets containing less than 3 words after cleaning
6. Extracting other properties of the tweet besides the text (e.g. the user name) and storing it in the database.

3.1.2. Data Statistics

For the purpose of this work, we extracted one year-worth of tweets without specifying any filtering terms. More specifically, we collected 500 random tweets every hour starting from March 2012 and ending in April 2013. Some of the statistics of the collected data are listed in Table 1.

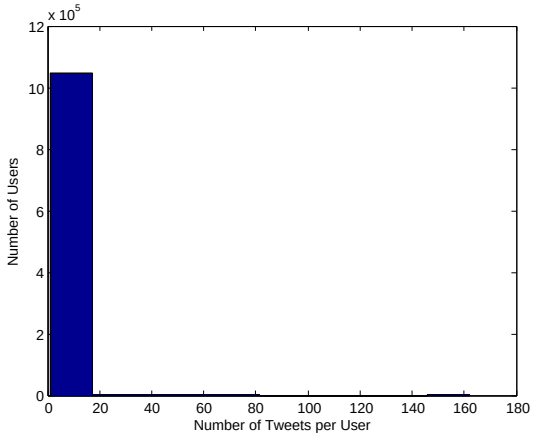
The number of tweets per user is shown in Figure 2(a), and the number of hashtags per tweet is shown in Figure 2(b). Both figures show that the distribution follows a power law, i.e. most users have a few tweets and most tweets have a few hashtags.

2. https://dev.twitter.com//
3. we used the open source tool, MorphAdorner, for lemmatization: http://morphadorner.northwestern.edu/

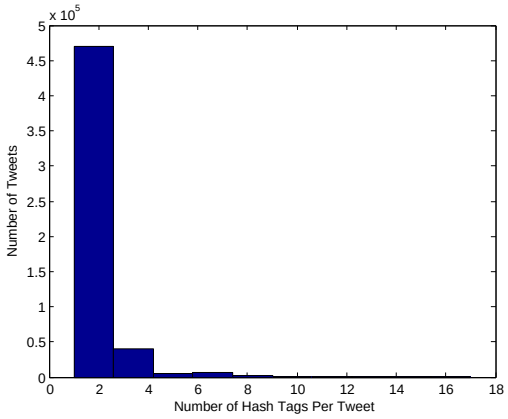
Table 1: Unfiltered Twitter Dataset Properties

Statistic	Value
Number of unique users	3,421,720
Number of tweets	4,164,402
Number of unique tweets	3,602,550
Number of tweets using hashtags	524,719
Number of unique tags	266,739

Figure 2: Twitter Properties



(a) Number of tweets per user



(b) Number of hashtags per tweet

3.1.3. Popular hashtags per month

Since hashtags are automatically generated by the users, their frequency represents the topics of interests (i.e trending topics). Figure 3 shows the popular (top 20 per month) hashtags for several months in a *tag cloud* visualization format, where the size of the font of each hashtag is proportional to its frequency. Some of the hashtags are always popular such as #TEAMFOLLOWBACK, hence, they are not very informative. On the other hand, some hashtags provide some insights about the trending topics during certain time periods. For example the #London2012 hashtag in August 2012 (Figure 3(a)) refers to the Olympics taking place in London at that time, whereas the #VMA hashtag in September 2012 (Figure 3(b)) refers to the MTV Video Music Award taking place during that month.

3.2. Detecting Trending Clusters

After pre-processing the stream of tweets into bag of words term vectors, the latter were used as input to RINO-Streams to discover clusters, then we took a snapshot of the centroids (i.e. the representative of each cluster) at the end of each day. Figure 4 shows some of these clusters detected on different days. The keywords representing each cluster in the figures correspond to the top 10 terms (i.e. with highest frequency) in its centroid. It can be seen that some of the topics in the clusters are meaningful, for instance about shopping, as shown in Figure 4(a), while others are generic, as shown in Figure 4(d). It is worth noting that the generic clusters are a result of using the bag of words model, since some words are very common, and hence, are used to create generic clusters that attract words that do not belong to the more detailed clusters. Despite the fact that these generic clusters are sometimes not meaningful, they are still valid (based on the periodic quality tests done by RINO-Streams), since these generic clusters are continuously being updated by the common words. This effect can be reduced by using different weights such as TF-IDF, however, it is challenging to find the right threshold value to identify which words are meaningful and which are very common and should be removed.

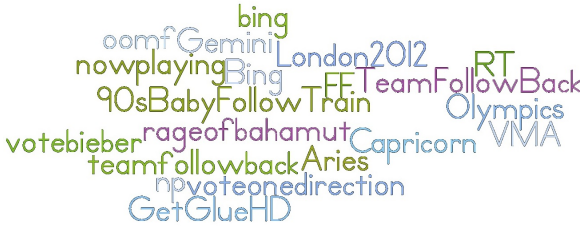
3.3. Twitter Case Study: Louisville Cardinals

To further analyze the use of Stream-Dashboard on Twitter data, we extracted a subset of tweets that are related to Louisville, KY. The subset was extracted by finding all the tweets that contain a set of keywords such as "louisville","uofl" and/or "cardinals". It is worth noting here that we selected the tweets for Louisville because we can relate and validate these stories since we are affiliated with the University of Louisville. However, the same analyses, that we are about to discuss in this section, can be applied to other unknown topics.

3.3.1. Pre-processing

The dataset being highly sparse, needed further pre-processing. Figures 5(a) and (b) show the number of terms per tweet and the frequency of each term respectively. Based on these figures, we removed all the tweets that had less than 2 terms, and then based on the frequency of occurrence of the terms in tweets, we removed the highest 1% and lowest 1% terms. After pre-processing, we obtained a total of 10,153 tweets using 4,354 terms.

Figure 3: Twitter: Popular Hashtags Per Month



(*a*) August 2012

(*b*) September 2012

(*c*) January 2013

(*d*) March 2013

Figure 4: Twitter: Detected Trending Topics for Several Days

(*a*) Cluster 1 : Shopping

(*b*) Cluster 2 : Smart Phones

(*c*) Cluster 3: Sports News

(*d*) Cluster 4 : General

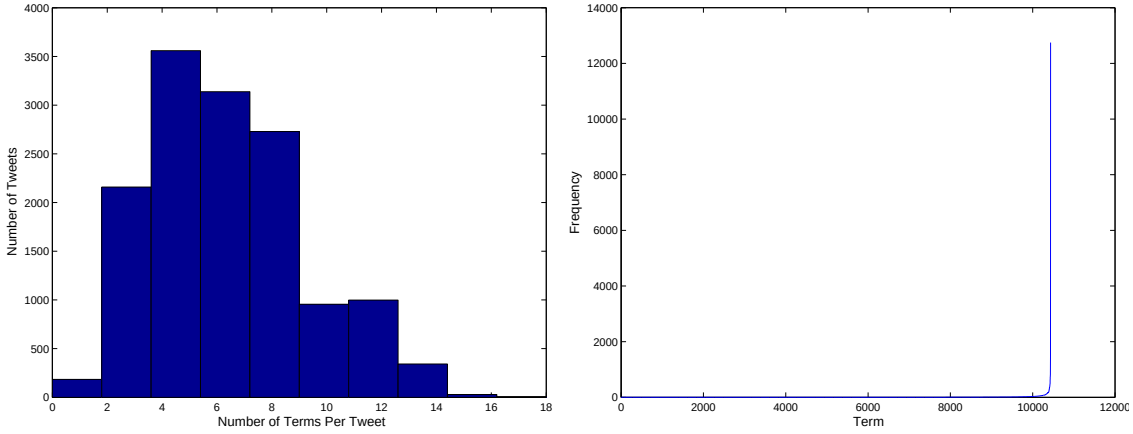
(*e*) Cluster 5 : Cities

(*f*) Cluster 6 : TV Shows

(*g*) Cluster 7 : Christmas

(*h*) Cluster 8 : General

Figure 5: Louisville Tweets



(*a*) Number of terms per tweet

(*b*) Frequency of terms sorted in an ascending order

3.3.2. Tracking Louisville Twitter Stories through Cluster Evolution Discovery

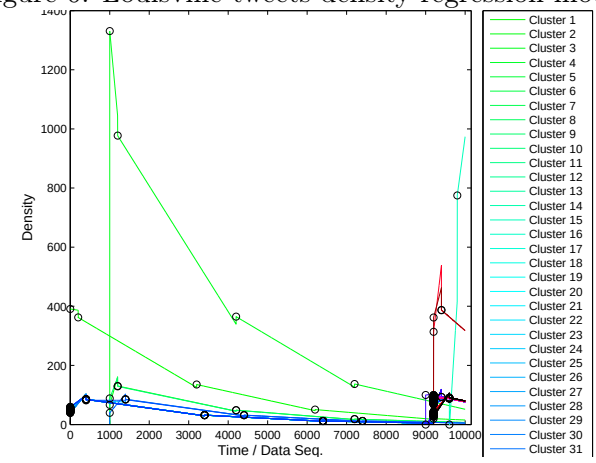
We used Stream-Dashboard to mine the Louisville tweets, and it was able to detect several trending clusters as well as their behavioral changes over time as shown in the density regression models detected in Figure 6. Density is one of the properties of the detected clusters, and it reflects the ratio between a cluster's cardinality (i.e. number of tweets in the cluster) and its variance (i.e. influence area of the cluster). TRACER, the second component of Stream-Dashboard, builds and maintains a regression model from the density values over time and these regression models can be visualized as shown in Figure 6. Furthermore, TRACER tests for major changes in the regression models, called milestones, and these are shown as circles, wherever they occur.

One contribution of this work is mainly to present trending clusters in Twitter messages in a format that is similar to a story, where we are inspired by picture story books. To this extent, we define a twitter story in a very simplified way, as a set of events, corresponding to cluster milestones, in a chronological order further described by a set of pictures and/or metrics that describe those events. For each of the stories, we define its start and end times, as well as major events that are crucial to understanding the story. We will analyze three example stories that were discovered, and some of the interesting topics related to the NCAA tournament where Louisville won the national championship. Although the notion of stories that we show here may appear as too simplified, they constitute a required step toward richer future story telling efforts, because what we present, was extracted in a completely automated manner and in a completely unsupervised manner (no Human guidance or previous training data).

The Sugar Bowl 2013 Cluster In January 2013, the Louisville Cardinals won the Super Bowl final game against the Florida Gators[4]. Stream-Dashboard was able to detect a

4. http://en.wikipedia.org/wiki/2013_Sugar_Bowl

Figure 6: Louisville tweets density regression models



trending cluster related to the super bowl and to track its evolution over time. Table 2[5] lists the dates and the top raw tweet of four milestones/events of the Sugar Bowl topic. Figures 7(a), (b) and (c) show the cluster's centroid as a cloud of the top terms, the cardinality regression model, and the similarity matrix of the clustering model at the time the story was detected to quickly gauge each cluster's size and validity, respectively.

The results show that the users started tweeting about the game at the first milestone (i.e. the start of the story), then the topic gained more popularity just before the game, which was at the second milestone. The topic popularity spiked after the Louisville Cardinals won, and this occurred at the third milestone, until finally it lost popularity after a couple of days (i.e. the end of the story). The quality of the cluster can be validated by observing a dense and dark block (denoting compactness) along the diagonal of the similarity matrix, which is highlighted in Figure 7(c). To further validate the detected story and its trends, we used Google Trends[6] for similar keywords in Figure 8, and it shows similar behavior during the same period. It is worth noting that Google trends are solely based on search query trends, thus offering an alternative source of validating the story line that was extracted from twitter. Moreover, the Y-scale represents how many searches have been done for a particular term, relative to the total number of searches done on Google over time. It does not represent the absolute search volume numbers, because the data is normalized and presented on a scale from 0-100. Each point on the graph is divided by the highest point and multiplied by 100.

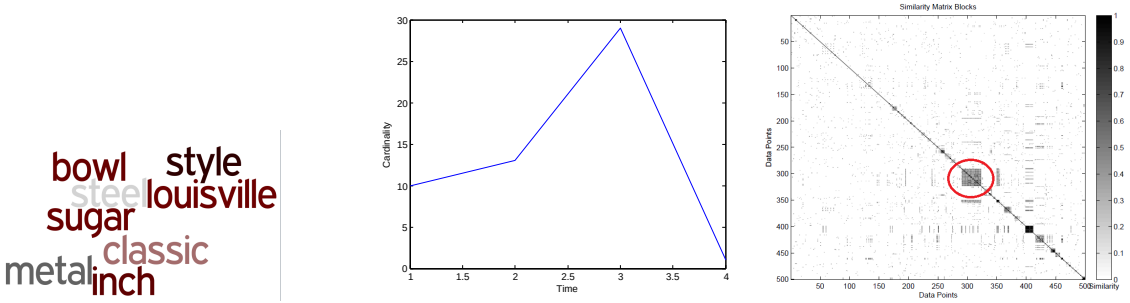
The Charlie Strong Contract Extension Cluster Another topic that was related to the Sugar Bowl, also taking place around the same time, was news about extending the contract of Louisville's football coach, Charlie Strong [7]. Table 3 lists the dates and the top tweets at five detected milestones for an automatically detected cluster, clearly related to

5. note absence of stop words and lemmatization from pre-processed tweets
6. http://www.google.com/trends/
7. http://sportsillustrated.cnn.com/college-football/news/20130123/louisville-charlie-strong-contract-extension.ap/

Table 2: Twitter Stories: Sugar Bowl 2013 Cluster Evolution Properties, with automatically detected milestones.

Milestone	Dates	Top Tweet (before pre-processing)
1	11/25/12-12/7/12	louisville plays florida the sugar bowl lets get
2	12/8/12-1/2/13	who wins tonight louisville florida sugar bowl time tonight
3	1/2/13-1/8/13	louisville cardinals sugar bowl champions
4	1/16/13	official congratulations the sugar bowl champion louisville cardinals

Figure 7: Twitter Stories: Sugar Bowl 2013 Cluster Evolution



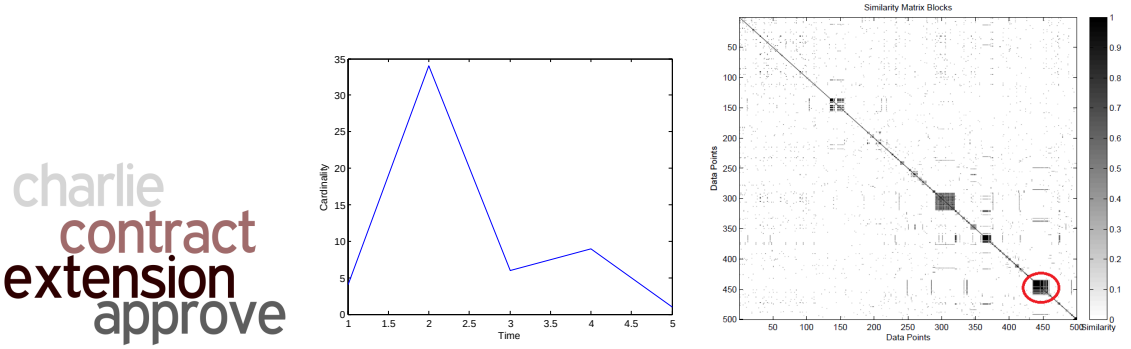
(*a*) Cluster Centroid (*b*) Cardinality Regression Models (time corresponds to milestones) (*c*) Similarity Matrix with the story block highlighted

Figure 8: Google Trends for keywords: Sugar Bowl, Louisville Cardinals and Florida Gators

Table 3: Twitter Stories: Charlie Strong Cluster Evolution Properties, with automatically detected milestones

Milestone	Dates	Top Tweet (before pre-processing)
1	11/25/12-12/7/12	vote for charlie strong university louisville for
2	11/28/12-12/6/12	sources louisville working strong extension louisville negotiating contract extension with charlie
3	12/10/12 – 1/2/13	vote charlie strong university louisville for
4	1/2/13-1/3/13	heartfelt congratulations charlie strong and louisville
5	1/23/13	louisville new contract for football coach charlie strong includes buyout the courierjournal

Figure 9: Twitter Stories: Charlie Strong Topic Evolution



(*a*) Cluster Centroid (*b*) Cardinality Regression Models (time corresponds to milestones) (*c*) Similarity Matrix with the story block highlighted

Charlie Strong. Figures 9(a), (b) and (c) show the topic centroid as a cloud of top centroid terms, the cardinality regression model, and the similarity matrix validation, respectively.

This cluster was detected as people started tweeting to vote for Charlie Strong's contract extension. It spiked when there were more sources to confirm the extension at the second milestone, and it increased popularity again when the extension was finally approved at the fourth milestone. The quality of the cluster is validated in the similarity matrix, where it corresponds to one of the a dark blocks that can be observed on the diagonal. Similar to the previous story, we further validated the detected story line, that was automatically mined from user-generated tweets using Google Trends, which are based on search queries, as shown in Figure 10, with the two sources exhibiting similar behavior.

The Kevin Ware Injury Story The gruesome injury of Louisville guard, Kevin Ware [8], shocked the fans, resulting in a lot of angst and discussion on Twitter, with first shock,

8. Kevin's bone poked out from his skin in a very graphic scene on TV, http://www.nytimes.com/2013/04/01/sports/ncaabasketball/kevin-wares-gruesome-injury-shakes-and-rallies-louisville.html?_r=0

Figure 10: Google Trends for keywords: Louisville Cardinals and Charlie Strong

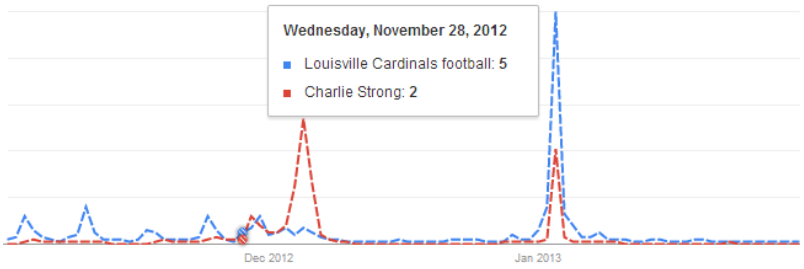


Table 4: Twitter Stories: Kevin Ware Injury Cluster Evolution Properties

Milestone	Dates	Top Tweet (before pre-processing)
1	3/29/13-3/31/13	praying for kevin ware louisville cardinals recovery
2	3/31/13	pray for kevin ware and the louisville team the worst ncaa injury ive ever seen for your respect
3	3/31/13	praying for ware after seeing the teams reaction firmly rooting for louisville now rivalry damned
4	4/1/13-4/4/13	kevin ware out the hospital and heading back louisville back brothers

then prayers, and ending with his exit from the hospital. Table 2 lists the dates and the top raw tweet of four milestones that were detected of this cluster. Figures 11(a), (b) and (c) show the cluster centroid as a cloud of top terms, the cardinality regression model, and the similarity matrix for validation, respectively. Google Trends for similar keywords, as shown in Figure 12, further validates the detected behavior of this cluster's story line.

The cluster appeared right when the injury took place, and it maintained popularity during the second and third milestones, where the tweets were very sympathetic with Kevin Ware. The topic spiked again at the fourth milestone, when Kevin Ware left the hospital and went back to join his team again.

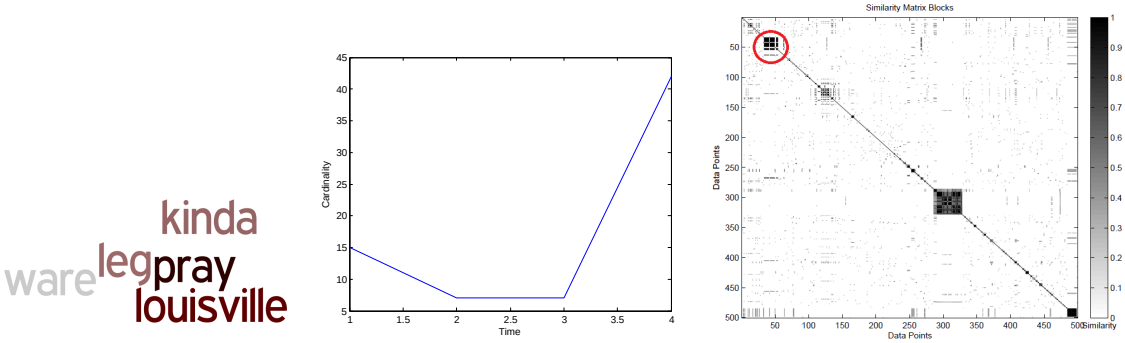
The NCAA Basketball Championship Game Clusters The Louisville basketball team won the NCAA championship in April 2013 [9]. Stream-Dashboard was able to detect several clusters related to this event from the same data stream described above, and several cluster centroids are shown, as top term clouds, in Figure 13. Some of the interesting discussions were caught in distinct clusters and their top tweets are shown in Table 5.

4. Conclusion

We presented an approach to use our unsupervised stream clustering and cluster evolution tracking and validation framework, Stream-Dashboard, as an apparatus to detect and track trending stories from a stream of tweets. We illustrated the effectiveness of Stream-Dashboard by analyzing four discovered clusters related to the Louisville Cardinals football and basketball teams during 2012-2013. The results show an ability to automatically detect trending stories as well as their major milestones, such as the start, end, and major intermediate events of the story (e.g. player Kevin Ware's injury). The stream clustering and tracking results can provide valuable information, on the fly, that reflect the evolving in-

9. http://en.wikipedia.org/wiki/2013_NCAA_Men's_Division_I_Basketball_Tournament

Figure 11: Twitter Stories: Kevin Ware Injury Cluster Evolution



(a) Cluster Centroid (b) Cardinality Regression Models (time corresponds to milestones) (c) Similarity Matrix with the story block highlighted

Figure 12: Google Trends for Keyword: Kevin Ware

Figure 13: Twitter Stories: Four NCAA basketball Championship game cluster centroids

(a) Vote for winner (b) Kevin Ware Recovery

(c) Albrecht (Michigan) vs Hancock (Louisville) (d) ESPN

Table 5: Twitter Stories: NCAA tweet discussions (Note: some slang and inappropriate terms are masked using the symbols: **)

Cluster's description	Top Tweets (note absence of stop words and lemmatization)
2 stellar players: Albrecht (Michigan) and Hancock (Louisville)	louisville reminds centralia their red and white that white boy killin louisville lil white boy from louisville look like one the little rascals louisville has white boy who can shoot too
Michigan surprised Louisville with Albrecht	louisville man who this white boy from michigan louisville all like have white guy nobody ever heard too albrecht unreal louisvilles pitino needs find answer albrecht your embarrassing louisville you werent even the scouting report
Fan Love/Anger: blaming sympathy for Louisville on Kevin Ware's Injury	yall didnt know louisville existed till that n**** broke his leg n**** said louisville sacrificed kevin ware leg for the championship the state kentucky going nuts last year louisville this year aint nobody f**** with basketball half these h**s louisville d*** now but was all uks last year

terests of virtual communities, and possibly support decision making tools in time-sensitive applications.

Acknowledgment

This work was supported by US National Science Foundation Grant IIS-0916489.

References

Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer, 2010.

Johan Bollen and Huina Mao. Twitter mood as a stock market predictor. *Computer*, 44 (10):0091–94, 2011.

Liana Evans. *Social media marketing: strategies for engaging in Facebook, Twitter & other social media*. Pearson Education, 2010.

Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, 2011.

Gabriela Grosseck and Carmen Holotescu. Can we use twitter for educational activities. In *4th international scientific conference, eLearning and software for education, Bucharest, Romania*, 2008.

Basheer Hawwash and Olfa Nasraoui. Robust clustering of data streams using incremental optimization. In *The IVth Alberto Mendelzon International Workshop on Foundations of Data Management. AMW*, Buenos Aires, Argentina, 2010. CEUR-WS.org.

Basheer Hawwash and Olfa Nasraoui. Stream-dashboard: a framework for mining, tracking and validating clusters in a data stream. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, pages 109–117, Beijing, China, 2012. ACM, ACM.

Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

Jimmy Lin and Alek Kolcz. Large-scale machine learning at twitter. 2012.

Albert W. Marshall and Ingram Olkin. Multivariate chebyshev inequalities. *The Annals of Mathematical Statistics*, 31:1001–1014, 1960.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. 2011.

Víctor J. Yohai Ricardo A. Maronna, Douglas R. Martin. *Robust Statistics*. Wiley-Interscience, 2006.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.