UNIVERSITY OF OSLO
Department of Informatics

Mining Twitter Data for Resource Usage Prediction

Dankun Du

Network and System Administration

Oslo and Akershus University College

of Applied Sciences

May 23, 2012

Mining Twitter Data for Resource Usage Prediction

Dankun Du

Network and System Administration
Oslo and Akershus University College of Applied Sciences

May 23, 2012

Abstract

This thesis investigates the predictability of Twitter traffic for topic-related websites' resource requirements by developing and implementing a data mining methodology. The new traffic correlation mining process is able to extract traffic surges and develop potential predictive mining and correlation techniques between Twitter and the corresponding forum. Thorough testing of this data mining methodology has been performed, and the results show that using Twitter data to predict imminent resource demands is a fruitful area of research. The findings in this thesis confirm the potential of utilizing the significant public interests expressed in Twitter data as a resource usage prediction tool for relevant websites.

Acknowledgements

First and foremost I owe my sincerest gratitude to my supervisor Æleen Frisch. Her extraordinary talent in all aspects of System Administration has always amazed and motivated me in many ways. She has accompanied me through all the difficulties with her unique sense of humor and patience. Without her brilliant instruction and continuous encouragement, I would not be able to achieve the same quality of work I had today. I feel great to have her be my supervisor and I always appreciate for her guidance.

I also offer my deep gratitude to Kyrre M. Begnum who initially inspired me with this interesting idea and provided valuable data sources and help for this project.

Also, I would like to take the opportunity to thank the Department of Informatics in University of Oslo for offering this Master degree program and making me interested in Network and System Administration field of study. In addition, I give my sincere appreciation to Oslo and Akershus University College of Applied Sciences for always providing me facility support, and the especially kind help when my laptop was accidently damaged during the thesis project.

Last but not least, I give lots of thanks to my beloved family and friends who has been understanding and supporting me in all aspects. Their encouragement and companionship brought me strength in completing this project.

Oslo, May 2012
Du Dankun

Contents

Chapter 1

Introduction

System administrators may not be as ambitious as people who work in sales or with the stock market, but they have some of the same information needs nevertheless. What system administrators are most concerned with is providing the best possible performance to users. Currently, cloud computing capabilities give organizations the flexibility to plan for and utilize resources in an efficient way as needs change. Instead of ”hard coding” resource deployment to a minimally sufficient level, companies are capable of scaling their server numbers up and down according to resource demand, attempting to maximize both performance and economy. Thus, understanding resource utilization become the crucial problem. If a site is to be able to modify its resource consumption in real time, dynamically adjusting and allocating resources to provide users with higher performance services in the most economical way becomes possible.

However, detecting the resource usage variations and responding by adjusting resource levels still has its disadvantage. For example, if an alarm indicating that more servers are needed is not generated until after the existing servers are near or at their saturation point, users will still suffer from lost connections since it takes time for the reaction process to go into effect. On the other hand, although there have been studies of predictive algorithms for identifying and predicting periodic resource consumption, most of them focus on long term prediction such as one day or more. They cannot detect any special events or situations causing a traffic increase above the normal level in a very short time and thus will not be able to address such situations.

However, the popularity of online activities like discussion or shopping is dramatically increasing during these years, human interests surging about certain events or newly developed products can crash relevant websites all of a sudden. When the British single board computer Raspberry Pi launched on February 29th 2012, not only the official launch website but also other manufacturers that sell this machine were crashed by overwhelming purchase demand from general public [1]. A similar case happened on February 5th 2012, when the NFL Super Bowl championship game was played. Coca-Cola websites crashed after experiencing overwhelming Super Bowl ad-driven traffic, and its down time was long enough that the company had to put up a maintenance page [2]. All these facts indicate that the traditional methods of long-term prediction without the ability to notice traffic sudden surge can no long be satisfying.

Twitter is as an online social networking and microblogging service that allow users to post messages of up to 140 characters. It provides short, simple and focused information about and fast reactions to the world's news. Users get first-hand accounts of events via Twitter, although some of them are later debunked. It has been estimated that there were an average of 290 million tweets posted per day in February 2012, 10 times more than in November 2009 [3]. As a social medium, Twitter is becoming an essential part of people's life. Users almost share every single piece of thoughts or events in their life, therefore tweets gain the capability of spreading news all over the world, and in turn reflecting public emotions and reactions to world's events. There hasn't been considerable published research on Twitter, but people already realize that social networks as Twitter are natural born data resource and dispersed information hidden in the entire collection of tweets are vast and valuable. Efforts have been made to extract potential predictive information for different research purposes, for example, to provide useful information for sales and election or predict the progress of a swine flu pandemic [4][5][6].

In the system administration area and specifically in the management of web sites, knowing about increased resource requirements even a little while in advance would be very helpful in deploying servers and other resources to optimize performance before, during and after the surge. Since Twitter is capable of revealing public enthusiasm which may lead overwhelming traffic on corresponding websites, this research investigate the phenomenon of whether Twitter traffic increases about a certain topic can be an early indicator for increased resource consumption on related websites? In other words, can Twitter predict the traffic variation of relevant websites by detecting traffic spikes about certain events on its own?

For an example of the potential predictability of Twitter, consider the following scenario. A popular singer is injured on the day of an important concert and is not able to give the performance. Twitter users are the first to report this news and spread the word across the Internet. In the following hours and days, there will be intense focus on and discussion about this event on related websites such as the popular music-related forums, the concert venue's site, the singer's own site, the site for the hospital where the singer is being treated, and so on. Thus a hypothesis is proposed: the initial Twitter traffic has the potential for giving advance notice of an unexpected surge in traffic for some or all of these websites.

By analyzing the data variation of Twitter traffic and the postings from its topic-related websites, it might be possible to find a correlation relationship between tweets and posts. If the predicatablity of Twitter can be verified, resource usage variation for relevant websites might be achievable in nearly real-time.

1.1 Problem statement

The goal of this thesis is to analyze the predictive ability of Twitter traffic for related website resource requirements by examining the data variation correlation between Twitter events and corresponding web forum posting in order to develop a predictive algorithm.

This research uses the data collected in advance of this thesis, from April 2011 through January 2012:

- **Twitter traffic** is measured by extracting data directly from Twitter platform by a Perl script. This data indicates how many tweets have been posted about a certain topic by people in a fixed time interval.
- **The forum resource usage** is also obtained by extracting real data directly from the forum website by a Perl script. This data tells how many posts have been generated in the forum by people over a fixed time period.

The thesis uses data mining methodology and statistical correlation procedures in order to analyze the data. This thesis is organized as follows. Chapter 2 presents a survey of related technical knowledge and work on data mining of social network sites. Chapter 3 describes the development of the data mining methodology for Twitter traffic and forum resource usage. In Chapter 4, results and analysis of testing and running this methodology on various data sets of different topics are presented. Chapter 5 provides discussions about the predictability of Twitter and evaluates the project design. The last chapter is a brief conclusion of the achievements of this project.

Chapter 2

Background

2.1 Predictability of Network Traffic

Network traffic is significant to server performance, and unexpected increased traffic will compromise the quality of service. Under these circumstances, a website will be at risk of losing users. Therefore, forecasting the traffic becomes an important problem in server resource management and performance optimization.

According to previous extensive studies, traffic prediction can be achieved with different time scales. Long term traffic prediction, such as days or weeks, can be used for service design or backup plans. Short term traffic prediction, especially real-time prediction, can be an effective solution to achieve dynamic resource allocation such as network bandwidth allocation and dynamic scaling web servers[7][8].

However, the character of network traffic makes predictability difficult. Network traffic varies considerably in trend and scale over time. Also, traffic variation is more complicated on small time scales[9], resulting in difficulties for short-term prediction based on previous traffic trends or patterns[10][11]. A study investigating the predictability of online game server resources by Jon-Erik Tyvann proposed a predictive algorithm[12]. His approach forecasted the resource usage for the following day based on training with previous repetitive resource usage data. In the end, he concluded that while his algorithm worked reasonably well in predicting the general shape of traffic patterns, it could be disrupted and experience poorer results when it encountered sudden significant deviations from the pattern even for a short time period.

2.2 Data Mining

Data mining is the name for a variety of efforts for identifying and analyzing information that is present (hidden) within large existing data sets [13][14]. It takes advantage of statistical and/or artificial intelligence techniques in order to discover patterns or relationships within existing data in order to predict future trends, behaviors or events.

Data mining tools help people to extract valuable predictive information which is not obvious to casual inspection or simple analysis. This powerful technique is significant in many fields such as stock trading and sales/marketing. For example, based on the result of mining massive quantities of relevant data, a company is able to make decisions about the next season's sale strategies or answer business questions in advance such as "Which customers are most likely to enjoy and purchase our new products? How many customers are likely to be interested in our promotion?" Traditionally, finding answers to questions like these has been time consuming and of limited accuracy, and it was mostly done with traditional market research methods like surveys and projection from past sales. Data mining offers the possibility of finding answers from new and more detailed sources of data, such as long term customer buying patterns in this example, which might prove more accurate and helpful in the long run.

Data mining requires modeling techniques, significant computing resources, and a lengthy research process. With a sufficient data resource, it focus on automatically discovering previously unknown knowledge and predicting future behaviors. Human beings benefit greatly from existing software, including the large variety of performance data analysis tools. However, this convenience never comes for free, and there is no "one size fits all." The analysis situation is very complicated for data mining purposes, providing significant challenges even for data experts.

2.2.1 Data Mining Process

Data mining requires modeling techniques, significant computing resources, and a lengthy research process. With sufficient data resources, it focuses on automatically discovering previously unknown knowledge and predicting future behaviors. Typically there are two types of data mining approaches[15]. The first one is similar to traditional statistical methods. These main analyze the moving trends within and distribution of the data set, trying to identify and construct a model for future events prediction. The second type is concerned with locating small deviations from normal behavior, trying to detect unusual patterns. For example, software which identifies a user's identity by their movement pattern on a touch screen should be able to notice the different movement performed by a stranger. Similarly, security software for a bank should detect unusual purchase patterns by a credit card customer which may ne caused by fraud.

General procedure of mining data includes the following steps[14]:

- Data Preparation: data mining results largely depend on the data source fed into the mining procedure, so in this step, data relevant to the specific mining purpose must be collected from multiple data sources.
- Clean and Integrate the data: this is an important pre-processing step for data mining. It provides cleaned, multi-dimensional data for future knowledge discovery.
- Rough Analysis using Traditional Tools: this step goes through all historical data, viewing the features of the data and make initial, possibly naive predictions using, for example, the mean, standard deviation, percentages, etc.
- Modeling: This the the key technique in data mining. However choosing and developing a proper model for future prediction is not obvious. Different modeling techniques should be applied, along with varying all of the model's different parameters. This step is usually complicated and repeated many times.
- Model Evaluation: This step is to measure and evaluate the model with respect to cost, confidence level and other aspects, to see if it is a success.
- Predict Event: Deploy the prediction model to improve whatever processes was the goal at the beginning.

2.2.2 Data Mining and Statistics

Both statistics and data mining place an emphasis on discovering knowledge, learning facts from data, so they overlap at many points. Some people consider categorizing data mining as a part of statistics, but this proposal causes lot controversy [16]. Data mining is defined mostly on the Internet is as the repetitive process of identifying novel patterns or models in existing data which mean the knowledge mined is previously unknown, and this is what traditional statistical analysis cannot always achieve. However, statistics do play an important role in data mining. Data mining developed from it, and it uses statistical analysis technique to construct and correct models during the machine learning procedure[15].

The following are common techniques of conventional statistical analysis methods:

- Description and Visualization: methods like calculating the average, median, percentiles and generating histograms and graphs in order to measure data variation. They are useful in interpreting large data sets. They are used used as the first step to help people gain an overall idea about the data.
- Correlation Analysis: measures the relationship between two variables or two data sets, in order to see how the changes in one variable/data set reflect the changes happening in the other.
- Regression Analysis: Based on correlation analysis, this measures how the strength of the relationship between two variables or two data sets. Results of regression analysis could be linear, multiple linear, or curvilinear models.

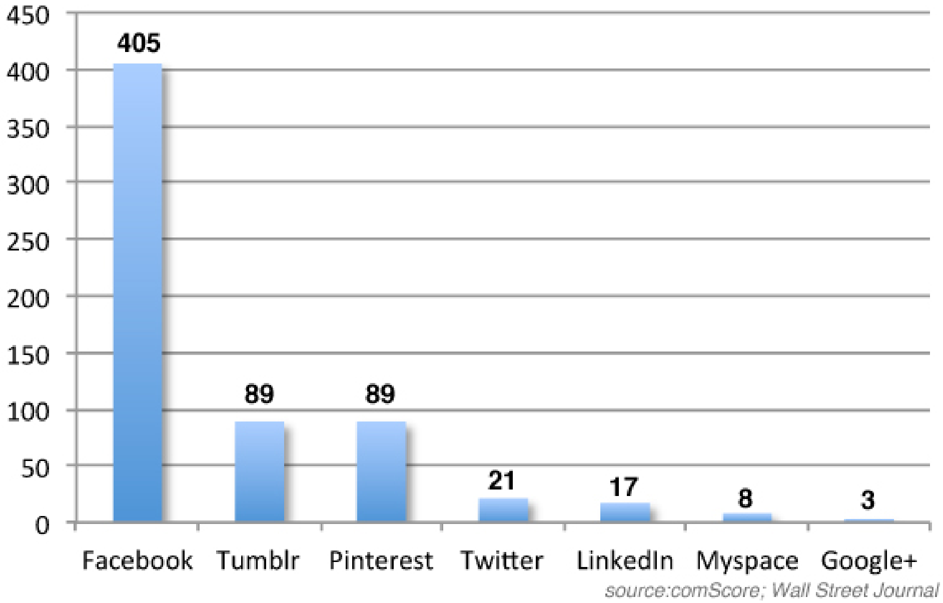
Other methods like cluster analysis, factor analysis, discriminant analysis are all widely used, but will not be not discussed here.

2.3 Social Networking Sites

Social networking sites such as Facebook, MySpace, YouTube and Twitter have attracted considerable interest since they were introduced in the 1990s[17][18]. Such sites are defined as an online service where users construct a profile and link themselves to other users with whom they share some connections[19]. They provide a platform for people to both build their online social relationships but and also to view the social network of their friends and family[17].

There is no doubt that online social networking has grown in popularity worldwide and become a routine part of people's lives. Figure 2.1 shows the average visiting period for several well-known social networking sites (SNSs)[20]. People share their opinions, interests and activity on these sites, maintain contact with old friends, make new friends and even start romantic relationships. SNSs contribute to bridging different continents, diverse culture and religions, making people who are more aware of and politically engaged in world's events. Recall the revolutions happened in Arab 2011 spring, people used Facebook, Twitter to communicate and organize the protests, and then used YouTube to share their movements with world. Social networks and social media made a remarkable contribution to their success[21][22][23].

Since SNSs have very large user groups, and the topics they discuss every day vary from entertainment to technology and from politics to business, SNSs provide rich sources of naturalistic behavioral data. The constantly generated information seems quite attractive to scholarly research[24][25]. Interesting questions concerning SNSs impact on human beings and its power to predict



source:comScore; Wall Street Journal

Figure 2.1: Average Minutes Per Visitor to Social Media Sites

the real-world event have been put forward these years[26][27]. To date, locating patterns and trends hidden in social networking data is in its very earliest stages, with a great deal of research projects in progress.

2.4 Twitter

Twitter, launched in 2006, is one of the most popular online social networking services, ranking the third in 2012 among the top ten leading social media websites [28]. Twitter is known as a microblogging service that allows users to post and share short messages of up to 140 characters, known as tweets. Users can follow others and see instant status posts by people that they subscribe to. Someone who subscribes to a person's tweets is known as a follower of that person. Twitter constructs a directed graph of user connections, which means tweets are only allowed to be shared by each users' followers, but not the other way around.

Twitter attracts huge attention and it is known for reacting to the world's news and spreading information very quickly[29]. Due to its small message size and excellent mobile apps developed by Twitter itself and third parties, users are capable of tweeting anywhere and anytime. This gives Twitter the capability of focusing on news in real-time, making it a natural news spreading platform. Nowadays users most likely get first-hand accounts of breaking

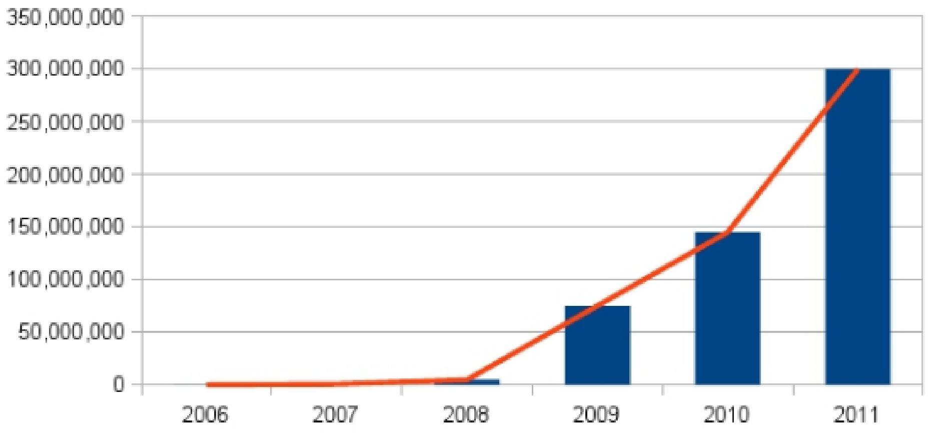


Figure 2.2: Twitter Growth in 2006-2011

news on Twitter[30][31]. Even though Facebook has more users and visitors every day, it can't compete with Twitter in reporting news[32]. In a word, Twitter is never late for breaking news and sometimes can be considered to be overreacting.

Figure 2.2 illustrates the Twitter usage increase from 2006 to 2011[33]. The number increases dramatically from 2008 to the present. In February 2012, it has been reported that there are 290 million tweets posted each day, 10 times larger than reported in November 2009 [3]. As a social network site, Twitter is obvious becoming an essential part of people's life. People share their thoughts about news, movies, feelings, and common interests on Twitter, and the information generated every day is vast. The amount of tweets and the vast range of topics makes Twitter seem to be a natural born data resource for research of many kinds.

2.4.1 Twitter as Data Resource

As mentioned above, the characteristics of Twitter make it a rich resource of instantly updated messages, providing possibilities for many different research purposes. Third party researchers are allowed to connect Twitter and obtain raw data such as general tweet content and number of tweets via the Twitter API (personally identifiable information is not available)[21]. This API enables users to extract information by querying for specific keywords like ″boxing,″ ″nfl,″ and ″nba.″ The service limits the maximum number of search results per query to roughly 1500[34].

Perl was the language chosen to integrate with the Twitter platform data collection purposes. The Net::Twitter:Search module provides a Perl interface to the Twitter API. A searching tool which returns the timestamp and number of tweets every five minutes was created by Dr. Kyrre Begnum (then an Associate Professor at Oslo and Akershus University College of Applied Science). Raw data are collected by running a Perl script in cron job. Data was collected by him over the period April-December 2011, and data was collected for January through March 2012 by this researcher. The keywords chosen by the original research were for the most part related to professional sports areas of interest.

2.5 Internet Forums

Internet forums are online discussion sites where people can hold conversations about a topic of interest in the form of posted messages[35]. A forum often focuses on a certain interest area such as sports, movies, technology, politics, etc. It therefore attracts people who are birds-of-feather with common interests. Forums differ from online social network sites in that a forum does not reveal a person's social relationship or allow instant message between users like Twitter and Facebook. Instead, it is a place people ask for help or express their views to others.

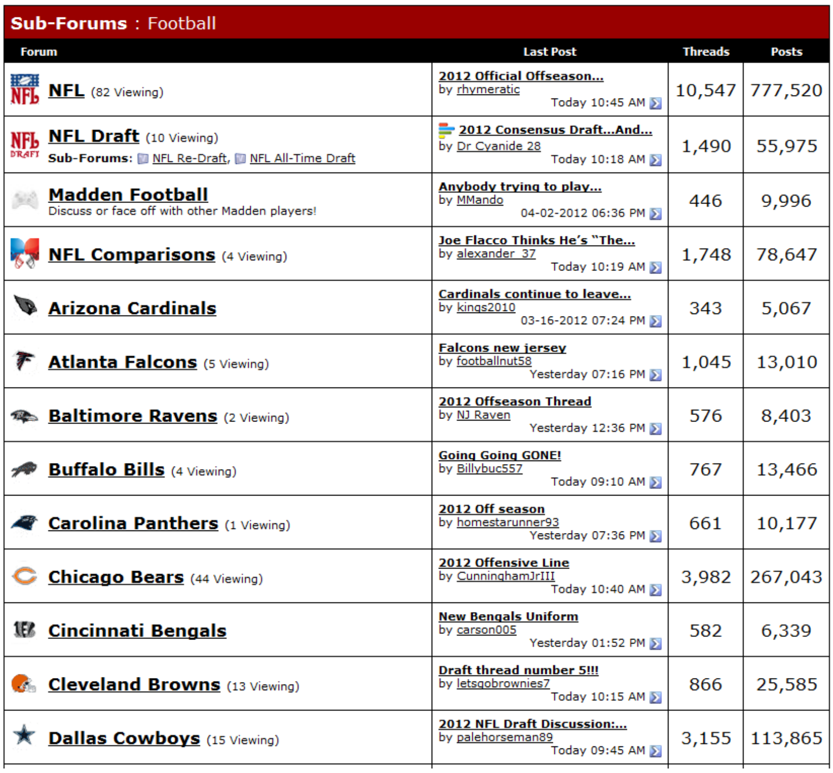
A conversation within a forum often contains lots of posts, and an interesting thread can cause hundreds or even thousands of responses. A thread is the name for a post and all of its follow-up replies. People get excited after sports competitions, breaking news events, new product launches, and similar events. All of these potentially lead to lively discussions and controversy on the relevant forums. Total forum traffic consists of a certain percentage of people who make posts and a much greater number who merely subscribe to the forum and just read them.

Like all online services, forums require enough resources to provide users with good performance. A robust forum site should be capable of handling all the traffic generated by posts and visitors. Generally, all the threads and posts are stored in a database, and they are retrieved as needed when accessed by users. Therefore, increased traffic can result in a large demand on forum server's resources since those posts could be in any form including as text, images and video.

For online discussion sites, large resource consumption would generally degrade their quality of service. Once users have an unpleasant experience, the forum takes the risk of losing that user. The general trend for resource usage is the more posts, the higher consumption, although the exact relationship between them is not something that forum sites reveal. Nevertheless, analyzing user posting rates provides a first-order indication of the demand on a website's resources.

Prosportsdaily.com

Prosportsdaily.com is an independent website which focuses on the news and events related professional sports in the USA, such as the National Football League (NFL), the National Basketball Association (NBA), Major League Baseball (MLB), and so on. It provides related new from every major newspaper in America and well as other services and features, including a series of user forums. Prof. Begnum captured posting rate data for some of the forums provided by this site over the same period as the Twitter data mentioned previously.



Sub-Forums : Football

Forum	Last Post	Threads	Posts
NFL (82 Viewing)	**2012 Official Offseason...** by rhymeratic Today 10:45 AM	10,547	777,520
NFL Draft (10 Viewing) **Sub-Forums:** NFL Re-Draft, NFL All-Time Draft	**2012 Consensus Draft...And...** by Dr Cyanide 28 Today 10:18 AM	1,490	55,975
Madden Football Discuss or face off with other Madden players!	**Anybody trying to play...** by MMando 04-02-2012 06:36 PM	446	9,996
NFL Comparisons (4 Viewing)	**Joe Flacco Thinks He's "The...** by alexander_37 Today 10:19 AM	1,748	78,647
Arizona Cardinals	**Cardinals continue to leave...** by kings2010 03-16-2012 07:24 PM	343	5,067
Atlanta Falcons (5 Viewing)	**Falcons new jersey** by footballnut58 Yesterday 07:16 PM	1,045	13,010
Baltimore Ravens (2 Viewing)	**2012 Offseason Thread** by NJ Raven Yesterday 12:36 PM	576	8,403
Buffalo Bills (4 Viewing)	**Going Going GONE!** by Billybuc557 Today 09:10 AM	767	13,466
Carolina Panthers (1 Viewing)	**2012 Off season** by homestarunner93 Yesterday 07:36 PM	661	10,177
Chicago Bears (44 Viewing)	**2012 Offensive Line** by CunninghamJrIII Today 10:40 AM	3,982	267,043
Cincinnati Bengals	**New Bengals Uniform** by carson005 Yesterday 01:52 PM	582	6,339
Cleveland Browns (13 Viewing)	**Draft thread number 5!!!** by letsgobrownies7 Today 10:15 AM	866	25,585
Dallas Cowboys (15 Viewing)	**2012 NFL Draft Discussion:...** by palehorseman89 Today 09:45 AM	3,155	113,865

Figure 2.3: ProSportsDaily.com Forum: Football

Figure 2.3 illustrates how the football forum looks in Prosportsdaily.com. The football forum consists of one main, big sub-forum named NFL as well as many other small sub-forums for different NFL teams: Arizona Cardinals, Atlanta Falcons, Buffalo Bills, and so on. Each sub-forum has hundreds or thousands of threads, depending on its popularity and supporters. This Super Bowl discussion platform provides the exact number of threads and posts for each sub-forums. A data fetching tool can be generated to scan this web page constantly and record the current threads and posts numbers in order to determine the total amount of posting during a given time period. Again, raw data are collected by running a Perl script in cron job.

Sherdog.net Forums

Sherdog.net forum is the largest and most active American website devoted to the sport of mixed martial arts. Although most martial arts originally comes from Asia, MMA has attracted millions of followers in the USA with the rise of UFC (Ultimate Fighting Championship). Sherdog tracks every piece of MMA news, providing information about individual records of fighters, reviews and previews of MMA events, interviews with fighters, and the like. Sherdog forums have over 60,000 MMA fans and fighters, over 20 MMA topic forums, and over 4 million posts. The structure of Sherdog.net forums is the way similar to Prosportsdaily.com, and data collecting job is done by running a Perl script in cron job too.

2.6 Tools and Source Platform

There are several specialized tools that will be used during the course of this research.

- RRDTool: whose name stands for Round Robin Database, is a powerful open source tool to store and process time series data. It specializes in creating time series graphs which visualizes the trend of data over a given time period. In this study, RRDtool is used to generate graphs of each day or week to give an visual understanding of traffic variation.
- R: is a free software for statistical computing and graphing. It is an extremely flexible package for professional statistics analysis. In this research, R is used to calculate statistical results, and generate histograms and distribution graphs.
- Perl: is a high-level, general-purpose, interpreted, dynamic programming language, widely used for text processing and tool development. Perl was used to fetch the data from Twitter and the forums. Perl is also used for data preparation and some parts of the data analysis.

2.7 Related Work: Data Mining Twitter

Many researchers are interested in exploring information hidden within Twitter data. There have been studies focused on mining Twitter for potential trends and correlations between social media and real world behaviors. For example, Cambridge Aviation Research proposed a simple and crude algorithm to examine the consumer attitudes towards some major U.S. airlines by mining Twitter [18]. They search and collect Twitter text of airline mentions and then counting the number of positive and negative words towards each airline to score and summary the sentiment of tweets for each airline, then they compared their results with American Consumer Satisfaction Index web site and confirmed that tweets were able to reveal public satisfaction level towards U.S. major airlines.

Moveover, Daniele Quercia and Michal Ksinski's research predicted users' personality types based on their Twitter activity and profile. They identified each user's type by their followers and subscription amounts and scored their personality based on how active they appeared to be on Twitter. Personality data was collected from 355 Twitter users and then used to study the relationship between user type and their personality traits. The researches could effectively predict users' personality types from their public Twitter data.[36]

On the other hand, exploiting Twitter as a predictive tool becomes very popular with high hopes for a significant outcome. Recent studies mainly focus on forecasting some real world outcome such as the box office results for a Hollywood movie, the sales market of a new product, or the spread of the swine flu[5]. Stock market analysts constantly try to find patterns in public information in the hope of generating large returns on investment. For instance, Johan Bollen and Huina Mao from Indiana University Bloomington demonstrated that Twitter could be used to detect public sentiment which played an important role in the financial markets. By analyzing 9,853,498 tweets posted by 2.7 million users in 2008, they pointed out a potential predictive correlation between Twitter mood and stock markets.[4]

Some other on-going research projects are interested in the social impacts on scholarly articles of science-related tweets. Researchers at the Yale University Bioinformatics department are investigating the relationship between tweets and traditional sources for measuring the scientific impact of journal articles and other reports of research[37]. A similar study of journal articles in Twitter was done last year, Gunther Eysenbach proves tweets is able to predict citations of scholarly articles without years' accumulation and he pointed out that Twitter could be seen as a metric to measure public interest in a specific topic[38].

Chapter 3

Data Mining Methodology Development

This project's goal is determining whether there are correlations between Twitter traffic and resource usage in related online forums. As noted in the previous chapter, serious research into Twitter and other social networking sites as data sources is just beginning. Accordingly, the key difficulty facing this research lies in the lack of systematic knowledge and methodologies in mining tweet- and post-related data. There is also no existing model for traffic correlation between different websites. Within the given time and resource constraints of this project, the complexity of dealing with data mining in an unknown area is obvious.

As is usual with data mining processes in general, this project was split into three steps, each will be explained in detail in later sections:

- Preparing the data: Data indicating network traffic and resource consumption on Twitter and related forums needs to be collected and stored in some way. It then must be processed for completeness and potential invalid/erroneous items. This "cleaning" job must be completed before the data can be used within the mining process.
- Data behavior analysis: Data should be described and interpreted in an understandable way. Specific data patterns for both the tweets and posts data sets must be defined and identified for future study.
- Correlation analysis: The relationships between Twitter data patterns and forum data patterns must be analyzed statistically in order to determine whether any mathematical relationship exists between them. The events will be considered in terms of both time and scale.

The customary data mining steps of visualization and simple statistical analysis are incorporated primarily in the second step above, but are also used somewhat in the first step as well. The development of the predictive model and its evaluation comprise the third step.

The data mining procedure was developed using data for the NFL as a training set. The procedure was then applied to other data sets.

3.1 Preparing the data

3.1.1 Fetching the data

This project attempts to exploit Twitter as a predictive tool indicating traffic trends and resource consumption on a specific related website. Therefore, the data for network traffic and resource consumption on Twitter and the chosen forum should be collected at constant intervals. The number of tweets about a topic during a given time period indicates the activeness of Twitter users. Therefore, retrieving the number of tweets about specific topics within short intervals can show real-time traffic variation on Twitter.

It is hoped that forum traffic will have tight connections with Twitter activities since they both reveal human interest in current world activities. The number of posts reflects user demands for resource usage in two areas: storage required for the posts themselves and network bandwidth consumed by the posters and readers. Generally, more posts indicate higher resource consumption, so collecting and analyzing user post rates has a relationships to the forum's demands on its resources, although the exact relationship between them is not known.

All the tweet and post data should be counted within the same theme and collected over the same time intervals.

The majority of the data for this project existed prior to its start. A Perl script collected and calculated sum values every 5 minutes for tweets and posts in different forums(NFL, NBA, NHL and boxing posts from ProSports-Daily.com forum, MMA posts from Sherdog.net forum) during May 2011 to January 2012. This existing data was provided by Dr. Kyrre Begnum (then an Associate Professor at Oslo and Akershus University College of Applied Science). Additional data was collected later for the NFL data set during this project from January to March 2012 (in order to capture the championship period for the NFL).

For each sport, forum posts and threads amount are recorded in two separate files, containing only two columns: timestamp and total posts/threads. The Twitter data for each sport is recorded in a separate file with three columns: timestamp, number of tweets and output web pages. The latter is confirmed by the tool's developer to be for debugging purpose only.

Tables 3.1 and 3.2 show the Twitter and forum data files and their sizes:

Sports Type	Tweets Size	Data Points
NFL	1.3M	93774
NBA	1.3M	78282
NHL	1.3M	80450
Boxing	1.3M	80255
MMA	1.3M	80154

Table 3.1: Twitter Data Files

Sports Type	Posts Size	Threads Size	Data Points(post)	Data Points(thread)
NFL	519K	28K	40850	2253
NBA	706K	48K	38023	2854
NHL	152K	5.9K	8631	377
Boxing	38K	2.5K	2378	166
MMA	379K	158K	19401	8974

Table 3.2: Data Files for posts in ProSportsDaily.com and Sherdog.net Forums

Besides, tweets with the topic of "beer" was also collected from during May 2011 to January 2012. No corresponding posts data from forums was found, but beer data can provide an opportunity for further comparative study. The file size of Beer tweets is 1.3M too with 79901 data points.

3.1.2 Data cleaning

According to data mining principles in chapter 2, data should be cleaned before feeding it into mining process. The better data one provides, the more accurate result one obtains.

Data can be polluted in different ways. Thus it is important to determine how much data is trustworthy. In this project, the existing data format only includes two elements to verify the data's validity and reliability: the timestamp and the tweets/posts amount. The data cleaning and validation job consists of the following:

- **Time series**: All the data was supposed to be collected every 5 minutes by the data fetching tools, so the timestamps in each datafile should follow in sequence with a constant interval. One must check the time series to see if everything went well during data fetching process.
- **Data gaps**: Data fetching is not always reliable. Thus, there could be holes inside each data set. Decisions should be made about how to deal

with data loss. Small gaps can handled by filling them in with average values. Large data holes must be excluded from analysis since it is meaningless to seal data loss without traffic variation over a long period.

- **Other constraints**: Facts about the tweets/posts value ranges should also be taken into consideration. The Twitter Search API returns at most 1500 search results, so any value larger than 1500 should be discarded. In addition, other impossible values like negative values or numbers orders of magnitude larger than normal post amounts should be discarded as well.

In this project, data from Twitter and the forum are handled separately, with different data cleaning methodologies.

3.1.3 Twitter Data

Judging from the data collected, the Twitter Search API appeals quite unstable. It was planned to fetch tweets number every 5 minutes. However, the actual time interval within the data varies from negative values to more than 7000 seconds. Possible explanations for those incredible large data gaps can be that the Twitter server got very busy sometimes, so the query needed to wait until it responded, or that query rates are limited somehow, since Twitter doesn't publish how they measure the exact rate limit against requesting client IP [39].

The following listing shows the raw data collected from querying the Twitter Search API at a constant time interval:

```
1302040906,1,3
1302040998,48,3
1302042319,426,6
1302042609,80,2
1302042909,62,2
1302043080,51,2
1302043207,40,2
1302043509,157,3
1302043808,67,2
```

The data file consists with two data columns: the timestamp and number of tweets during the specific interval; the third column – the number of pages – is only for debugging purposes.

Using simple sorting and arithmetic computations, it is easy to spot different types of errors existing in the raw data. The following listings give examples of invalid data. The third column here is the time difference between two consecutive timestamps. Most of the timestamps lie in a seemingly random range around 300, instead of being exactly 300. Beginning from 1304265013, the timestamp are entirely messed up in the following few records. The time difference can be as large as 2374 seconds – nearly 8 times larger than 300 seconds – and can be as small as 4 seconds, although the query was not supposed to be called after such a small period. In addition, timestamps 1304265020 and 1304265004 are out of order.

```
1304261443 ; 108 ; 300
1304261748 ; 107 ; 305
1304262038 ; 74 ; 290
1304262372 ; 109 ; 334
1304262639 ; 85 ; 267
1304265013 ; 2 ; 2374
1304265020 ; 3 ; 7
1304265004 ; 1036 ; -16
1304265041 ; 15 ; 37
1304265045 ; 4 ; 4
1304265337 ; 158 ; 292
1304265636 ; 138 ; 299
1304265941 ; 172 ; 305
```

There are several possible explanations for these problems. Since the data was collected by a system cron job, it might be uncertain how long it took to gather the data and how timestamps were rounded. Also, when collecting data, the running server could be down or without an Internet connection. It has been confirmed by Dr. Kyrre Begnum that these situations did happen a couple of times (e.g., losing power in a storm). One might also blame the Twitter Search API for these incredibly large data gaps as noted above.

Data pollution needs to be taken into account seriously since valid source data is the most important prerequisite for the later data mining process. Data loss and errors have been calculated and analyzed by statistical methods. The histogram 3.1 below indicates the distribution of the Twitter data interval. The values are mostly centered around 300 with dozens of differences, and a few are spread out along X axis. Statistical analysis shows that 96.5% of the time interval are located between 250 to 350 seconds. Since a randomly varying range makes it quite difficult to seal data gaps, it was decided to set 250s-350s as a safe time interval range for retaining data. Other values beyond this scope will be discarded. A Perl script was created to fulfill the cleaning task.

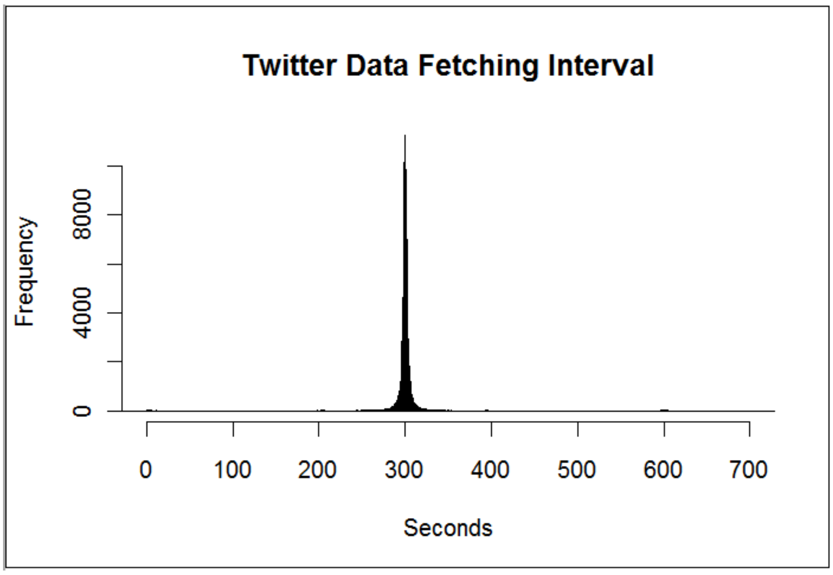


Figure 3.1: Histogram of Twitter Data Fetching Intervals

3.1.4 Forum Data

Forum data was extracted from ProSportsDaily.com every 5 minutes by a Perl script. The original data file contains two columns: timestamp and total posts number. The following listing gives an example of this raw data:

```
1304343909 560410
1304344209 560411
1304344508 560412
1304344808 560413
1304347207 560414
1304348407 560415
1304349008 560417
```

According to the statistics analysis, the forum data appears more reliable than the Twitter data discussed above. Time intervals are mostly fixed to 300 seconds, and errors occurred only in form of data loss. The third column in following listing shows calculation results for time differences between two successive queries. There are 3 missing timestamps during the period 1304389809 - 1304391007:

```
1304389207 ; 1 ; 298
1304389507 ; 2 ; 300
1304389809 ; 2 ; 302
1304391007 ; 2 ; 1198
1304391308 ; 3 ; 301
1304391607 ; 1 ; 299
```

```
1304391907 ; 1 ; 300
1304392208 ; 2 ; 301
1304392508 ; 1 ; 300
1304392808 ; 1 ; 300
```

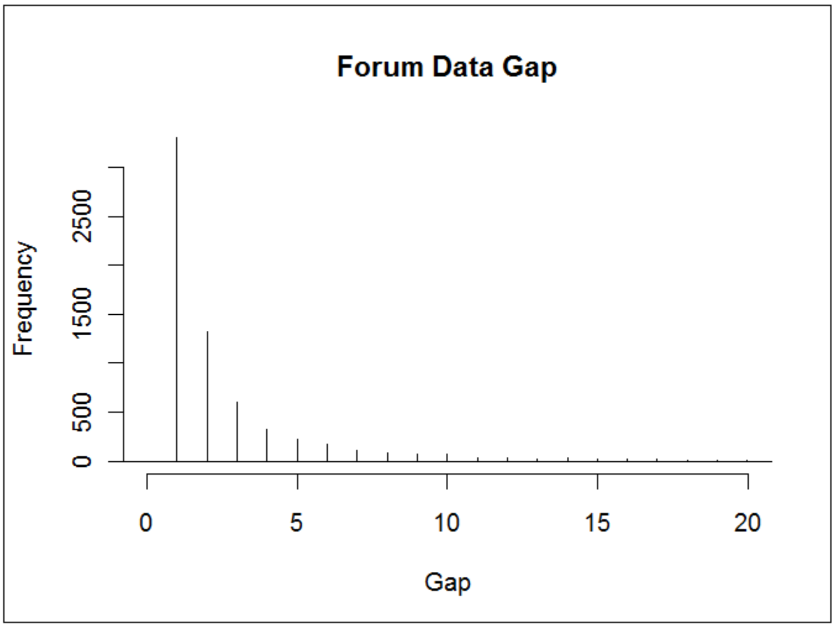


Figure 3.2: Data Gap Distribution of NFL Forum Data

Figure 3.2 illustrates the frequencies for the data gap lengths occurring in the forum data, indicating that there are more small gaps than large ones in this data set. The small data loss can be sealed easily by averaging the total posts into each missing interval. Larger data loss should be discarded immediately since its meaningless to fix long period's data missing without any variations.

The distribution shows the majority of data gap lengths are within 5 collection intervals (5 intervals = 25 minutes), so the decision was made to seal the data loss only when the gap length is smaller than 6. This is a simple and effective way to fill small holes in the dataset and reduce ″blank″ periods in a day's record. A Perl script was created to do the gap sealing work. The script located the data gaps and filled in an average number obtained from total posts divided by gap size. Full script can be seen in Appendix A.

3.2 Data quick analysis

A rough calculation and analysis of the NFL data set was performed in order to acquire its basic features. When investigating the trends within large data sets, often the first approach is to interpret the data in a simple but understandable way via visualization, since graphical results can provide a basic sense of data behavior.

In general, the NFL season starts in early August with a 4-game exhibition period, and the regular season runs from September to the end of December. After the end of the 16-game regular season, the playoffs occur, with the final game – the Superbowl – occurring in late January or early February. In 2012, it occurred on February 5th.

The first analysis performed was summing the tweet and post numbers for each day and then plotting them in Excel by month. The idea was to see whether Twitter and forum users would be influenced by the NFL match season.

The NFL data was collected from May 2, 2011 until March 14, 2012. The Twitter data was almost complete and covered every single day. However the forum data was missing for all of June and was partly missing in May, October and January.

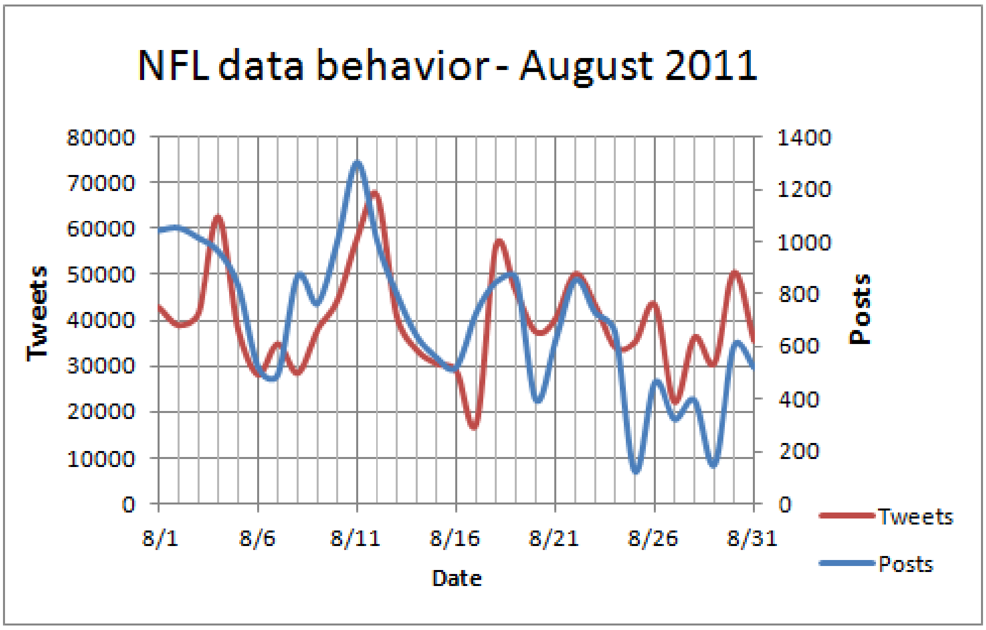


Figure 3.3: NFL data behavior in August 2011

Figure 3.3 illustrates the trends of tweets and posts in August 2011. August was the month when the NFL pre-season exihibition games were held. Exhibition games are also known as preparation matches, so they can be regarded as the kickoff for the NFL. According to the 2011-12 NFL schedule, games started on August 7th and occurred on the 11th, 12th and 13th for week 1; on the 15th, 18th, 19th and 20th for week 2; on the 21st, 22nd, 25th, 26th and 27th for week 3; and on the 28th, 29th for week 4. Figure 3.3 demonstrates that all of the peak points for either tweets or posts exactly matched the days when games occurred. Twitter data and forum data followed almost the same pattern of increases and decreases, although sometimes posts peaked ahead of tweets. In most cases, they increased to their maximum point on the same day.

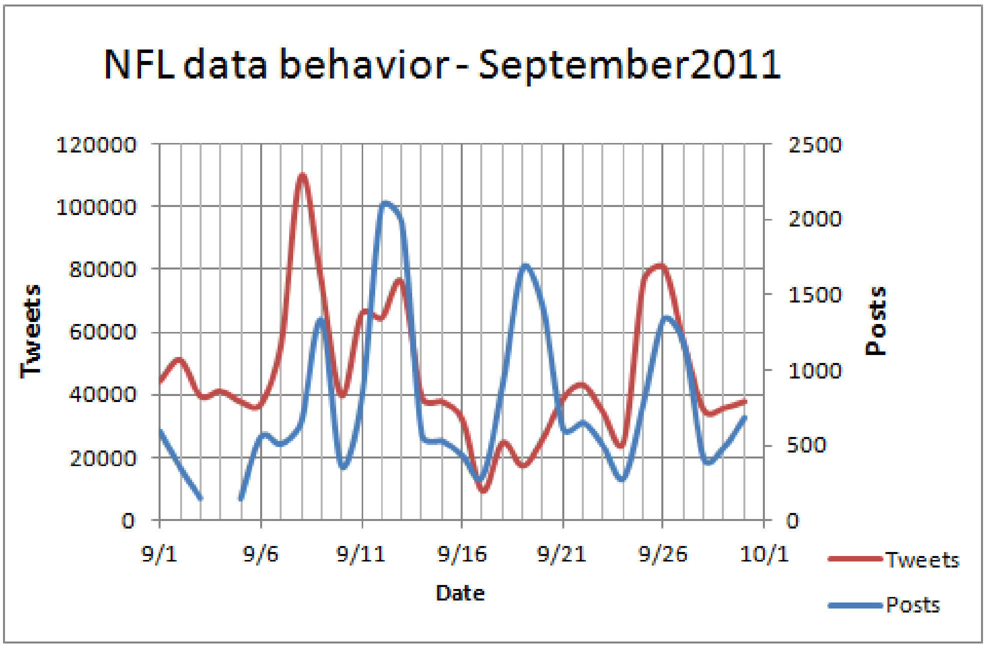


Figure 3.4: NFL data behavior in September 2011

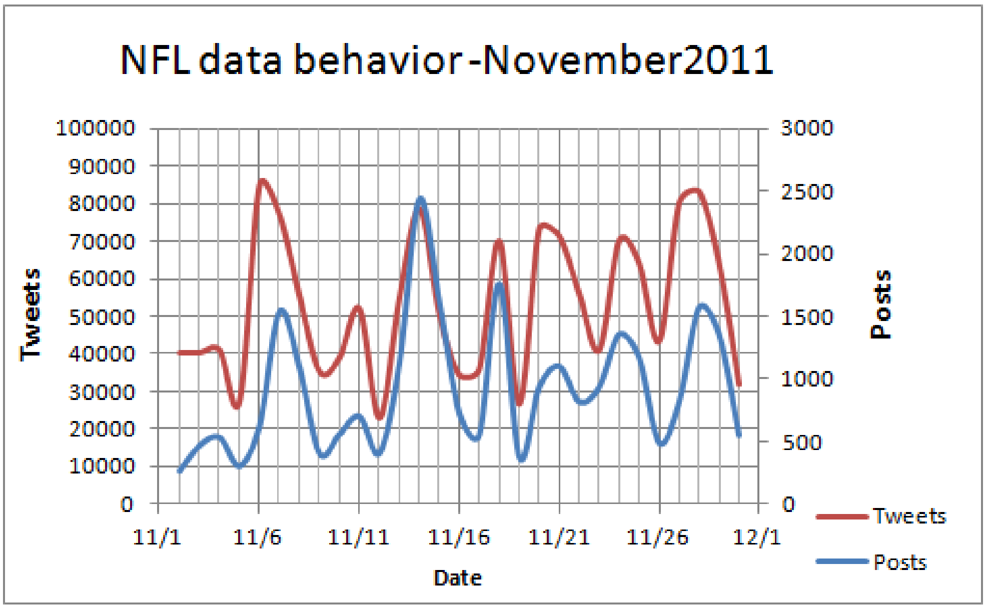


Figure 3.5: NFL data behavior in November 2011

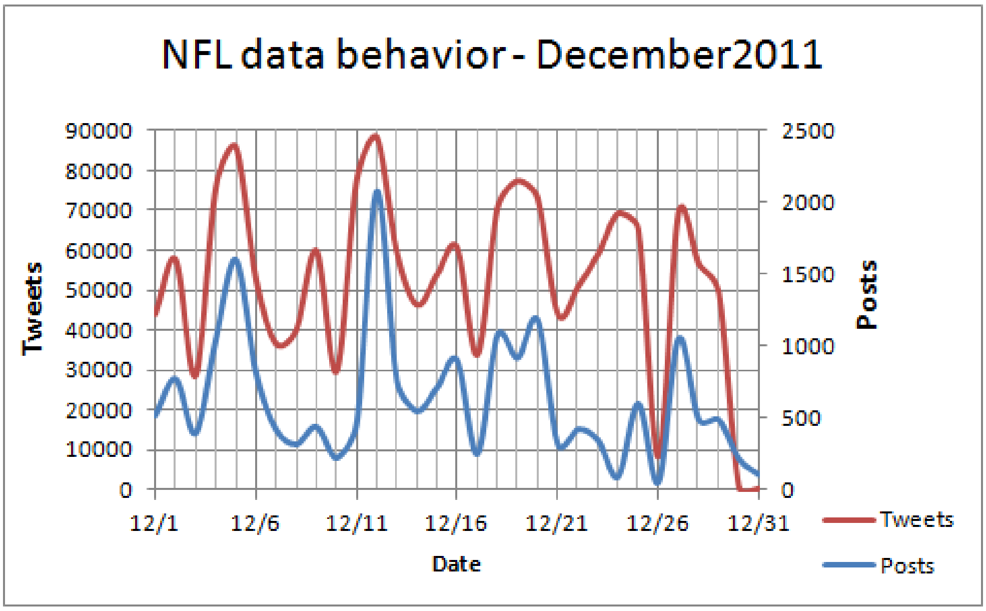


Figure 3.6: NFL data behavior in December 2011

Figures 3.4, 3.5, and 3.6 give examples of tweets and posts behaviors during the NFL regular match season. Generally the games were played on Monday and Sunday, although sometimes also on Saturday and Thursday. All matches have corresponding tweets and posts peaks in the figures. In addition, Twitter and forum traffic spikes' timing was even more interesting: tweets always reached its maximum value ahead of or at least the on the same day as posts. The data appears similar to human emotional behaviors, and it supports the idea that human enthusiasm for the NFL indeed turns into traffic burst on corresponding web sites.

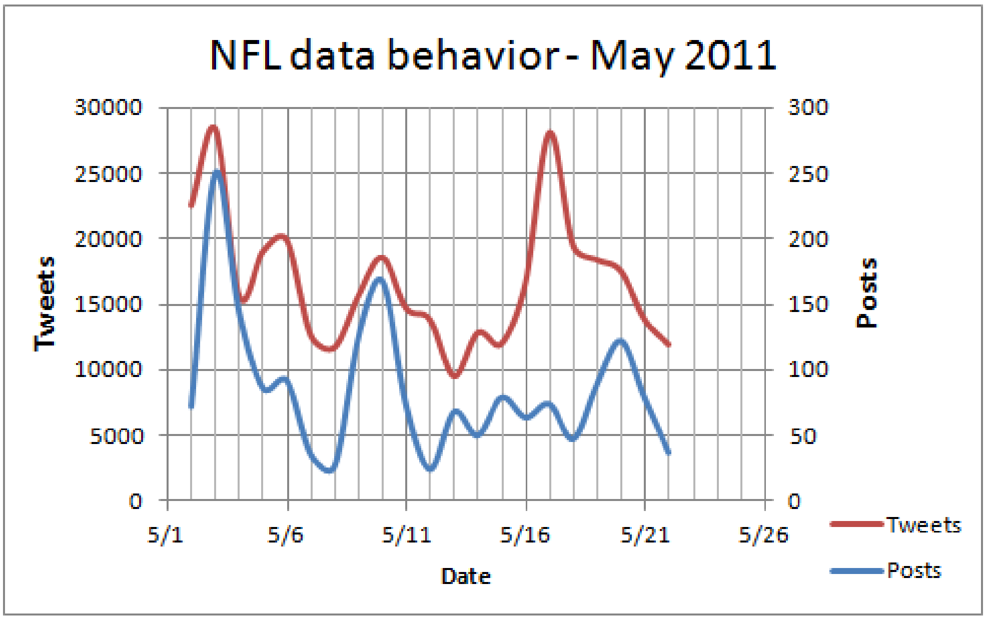


Figure 3.7: NFL data behavior in May 2011

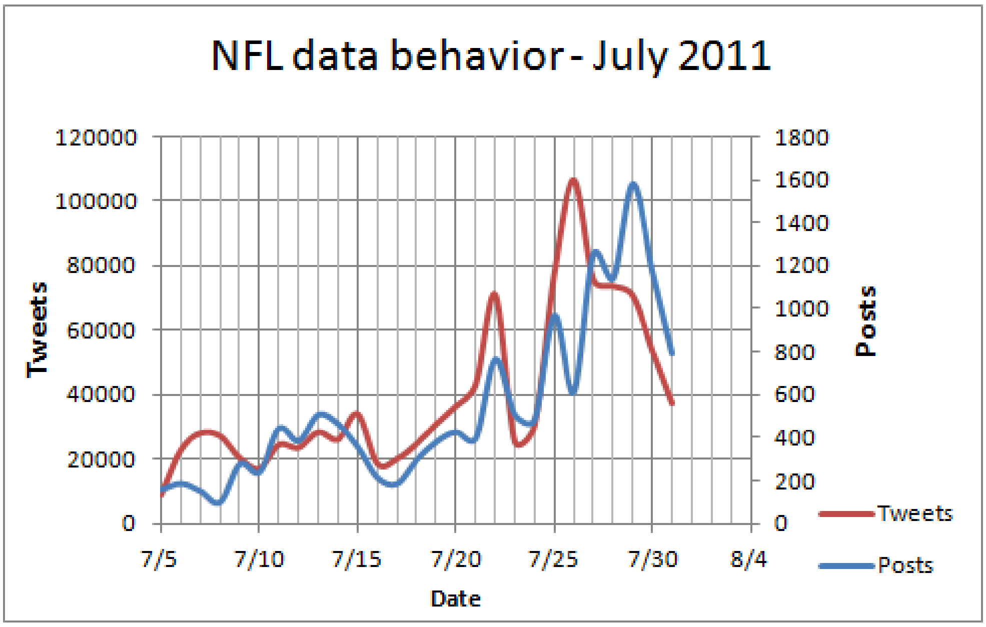


Figure 3.8: NFL data behavior in July 2011

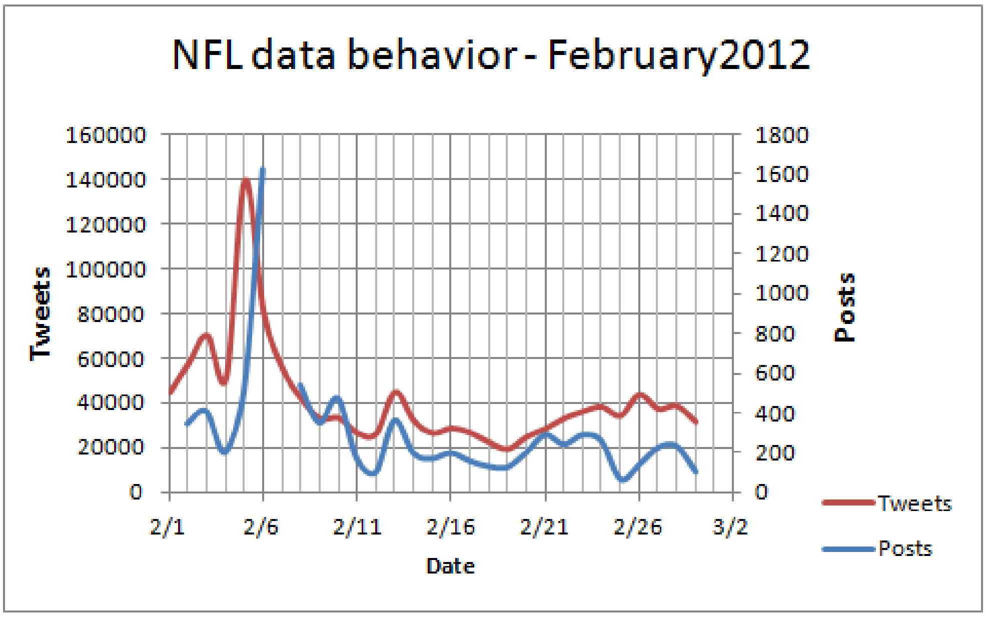


Figure 3.9: NFL data behavior in February 2012

Figures 3.7, 3.8, and 3.9 give examples of tweets and posts behaviors outside NFL season. The Twitter data and forum posts still follow some similar trends of increasing and decreasing, however their traffic spikes no longer appear in a one-to-one match. February 5, 2012 was the final NFL game of the 2011-12 season. Figure 3.9 illustrats clearly how the final match impacted both Twitter and forum data. This was the only data spike in February, and the maximum value was more than 3 times larger than the other days.

The preliminary investigation of Twitter and forum data was accomplished with very satisfying findings. The results indicate that Twitter and the topic-related forum traffic can be largely affected by current events and public passions. Tweets increased slightly prior to posts at times when they both reacted to NFL sports events, and this results suggest a possible correlation of traffic surges between twitter and the corresponding web site.

3.3 Event Identification

This section describes the efforts of event definition and filtering. The purpose of this project is to investigate the correlation between sudden dramatic traffic increases on Twitter and its topic-related websites, which are difficult to forecast by any long-term predictive algorithm. Thus, the primary problem is to identify the unexpected traffic surges on each site. Considered as the modeling part in the data mining process, the idea here is to define a unusual traffic spike as a event, and identify all the events on both websites by a common method. There is no existing systematic approach or modle for locating such spikes in this data traffic. The following subsections explain the different attempts to do so and the final resulting method.

3.3.1 Twitter Events

Defining events

The cleaned data is difficult to comprehend in its original format. Visualizing data points has been shown to be advantages in last section. However, plotting so much data by hand is obviously time consuming. RRDtool provide an easier and more efficient way to handle and visualize the data, allowing the storage of time-series data and providing the ability to generate graphs according to changing demands. The following commands create a round-robin database (RRD) called twitter.rrd designed to sample and store data points starting from timestamp 1302040900 (Unix epoch time).

```
rrdtool create twitter.rrd −−step 300 \
 −−start 1302040900 \
 DS:tw:GAUGE:600:0:U \
 RRA:AVERAGE:0.5:1:100000 \
 RRA:AVERAGE:0.5:6:17000 \
 RRA:AVERAGE:0.5:48:2100 \
 RRA:AVERAGE:0.5:288:350 \
 RRA:MIN:0.5:1:100000 \
 RRA:MIN:0.5:6:17000 \
 RRA:MIN:0.5:48:2100 \
 RRA:MIN:0.5:288:350 \
 RRA:MAX:0.5:1:100000 \
 RRA:MAX:0.5:6:17000 \
 RRA:MAX:0.5:48:2100 \
 RRA:MAX:0.5:288:350
```

This database accepts data values every 300 seconds. If no new data is supplied for more than 600 seconds, the tweets value will be considered unknown. The RRA lines define various archive areas. The first RRA line stores 100000 5-minute twitter data points (i.e., raw collected data). The other RRA lines stored tweet values averages over every 30 minutes (300 seconds * 6 intervals), every 4 hours (300 seconds * 48) and every 24 hours (300 seconds * 288). The MIN lines store the minimum tweets value and the MAX lines store the maximum tweets value with the same time periods. Since the original data is fetched every 300 seconds, this means RRD stored all the data points by the first RRA line. Data in other forms, like AVERAGE or MAX values over 30 minutes, 4 hours, or 24 hours, are also computed and stored for the case of further retrieval.

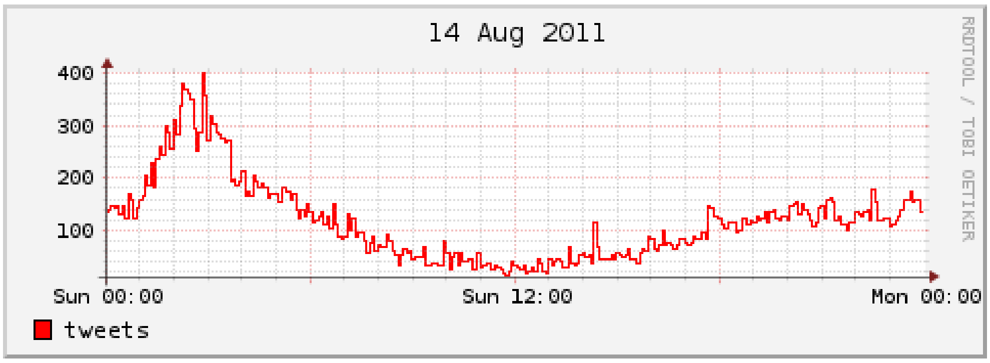


Figure 3.10: Tweets Variation August 14th 2011

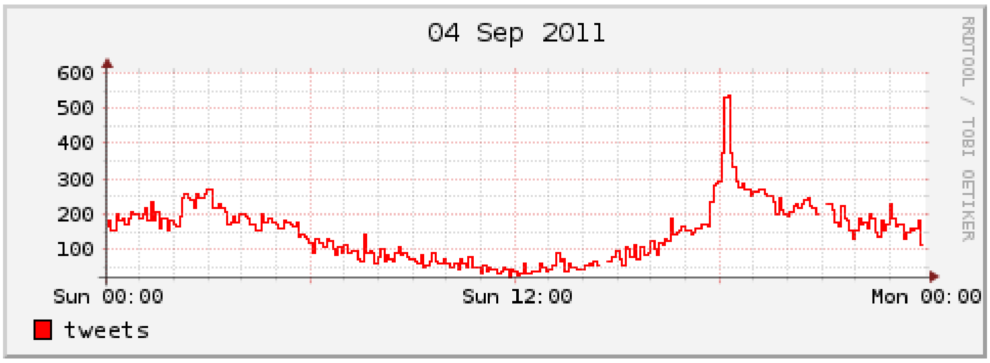


Figure 3.11: Tweets Variation September 4th 2011

After storing data points into its database, RRDtool is capable of generating graphs with all stored values. Figures 3.10 and 3.11 illustrate the tweet amounts varying in a single day. For example, in figure 3.10, the X-axis shows hour of the day, starting from Thursday (August 11th) 00:00 to Friday (August 12th) 00:00, divided by hour. The Y-axis plots the number of tweets. From the graphs, it is clear that there is sufficient data to show the Twitter traffic's normal behavior and as well as sudden changes over time.

Due to their magnitude, tweets spikes are relatively easy to spot. Careful examination and comparison indicates that there are two major traffic increase patterns in all of the Twitter data. One is illustrated by what happens in Figure 3.11: tweets go up all of a sudden around 1:30 a.m, and no other huge variation occurs in the rest of the day. The other pattern is illustrated in Figure 3.10: the tweets amount increases gradually from 1:00 a.m and peaks around 2:00 a.m, while in the other hours of the day, the variation seems quite smooth. Both two kinds of traffic increases should be counted as an event, and all events in the NFL Twitter dataset need to be identified and extracted by a common method.

Event extraction

As mentioned earlier, an event in Twitter is defined as an unexpected significant increase in tweets, and in order to analyze the correlation between tweets and posts unexpected traffic increases, events must be filtered from the data set in advance. According to the literature, the general logic for identifying spikes in traffic or any data set is to set a valid threshold such that when the incoming data value is higher than the threshold, an event is identified. In addition, since Twitter is expected to be exploited as a predictable tool to forecast forum traffic surges, this requires that Twitter events should be identified in a ongoing process. Taking all of this into consideration, the problem here is to define a valid threshold to identify all tweets spikes in real-time with minimum errors.

The first approach was to compute a fixed threshold based on basic statistic analysis of the tweets data. The reasoning behind this approach is that the number of tweets might be only slightly different across days or months and vary around a certain level. However this idea has been proved wrong after examining some histograms and computing basic statistics.

Month	Mean	Standard Deviation
April	115.12	124.00
May	67.00	45.00
June	58.69	60.24
July	137.47	143.76
Aug	147.86	105.70
Sep	188.35	172.77
Oct	189.99	182.23
Nov	192.79	181.33
Dec	209.99	190.57

Table 3.3: Mean and Standard Deviation of 5-minute Tweets Counts

Table 3.3 shows the mean and standard deviation results for the data by month, from April to December. The average tweets number varies without any distinct pattern from month to month, and the standard deviation is very large in all cases. This means the distances between the mean and the various data point fall into a large range, leading to the conclusion that the tweets data is relatively random, and its variation range is uncertain. Reducing the time period to weeks and days does not reveal any obvious connections or patterns in the tweets. Therefore, setting a fixed threshold by a basic statistical computing based on the previous tweets average values is untenable.

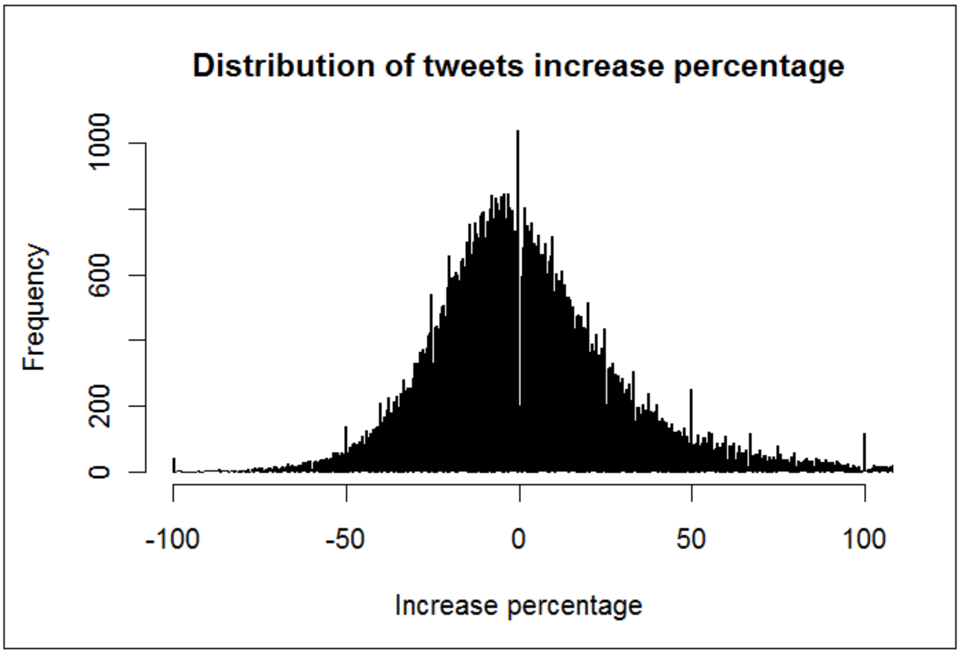


Figure 3.12: Histogram of tweets increase percentage

The next filtering criteria examined was to consider the increase percentage, assuming that a significant data increase will lead to a large increase percentage compared with the previous data point. Figure 3.12 shows the histogram of increase percentage for all NFL tweets data. It is a normal distribution with mean equal to 4.72 and standard deviation equal to 41.92. According to the 95% principle of normal distributions, here 95% of the data is located within the interval (-79.12,88.56), meaning that less than 5% of the increase percentages are larger than 100%.

So the decision was made to filter data points when the increase percentage is over 100%, and this is expected to present data which suddenly increase dramatically. However, this method generates lots of false alerts due to small prior data points. The following listing, which includes the timestamp, number of tweets and percentage increase, gives examples of these false positives:

```

Aug 12 1:20:42 595 −11.72106825
Aug 12 1:25:43 970 63.02521008
Aug 12 1:30:43 924 −4.742268041
Aug 12 1:35:43 921 −0.324675325
Aug 12 1:40:41 1302 41.36807818
Aug 12 1:45:41 1372 5.376344086

Aug 12 9:20:32 64 −23.80952381
Aug 12 9:25:31 109 70.3125
Aug 12 9:30:32 54 −50.4587156
Aug 12 9:35:29 39 −27.77777778
Aug 12 9:40:30 87 123.0769231
```

The last column is what matters most here. Tweets are generated by Twitter users more frequently around 1:00 a.m, with the tweets amount within 5 minutes being as many as 1372, while around 9:00 a.m Twitter activity about the NFL seems to calm down to mostly less than 100 tweets every 5 minutes. However, the increase percentage in the last column shows totally opposite results. If using the 100% increase as a threshold, then only data at 9:40:30 will be identified as an event. However, its 87 tweets is actually a very small value compared to the rest of the day. And the actual large tweet numbers near or over 1000 are ignored by this strategy.

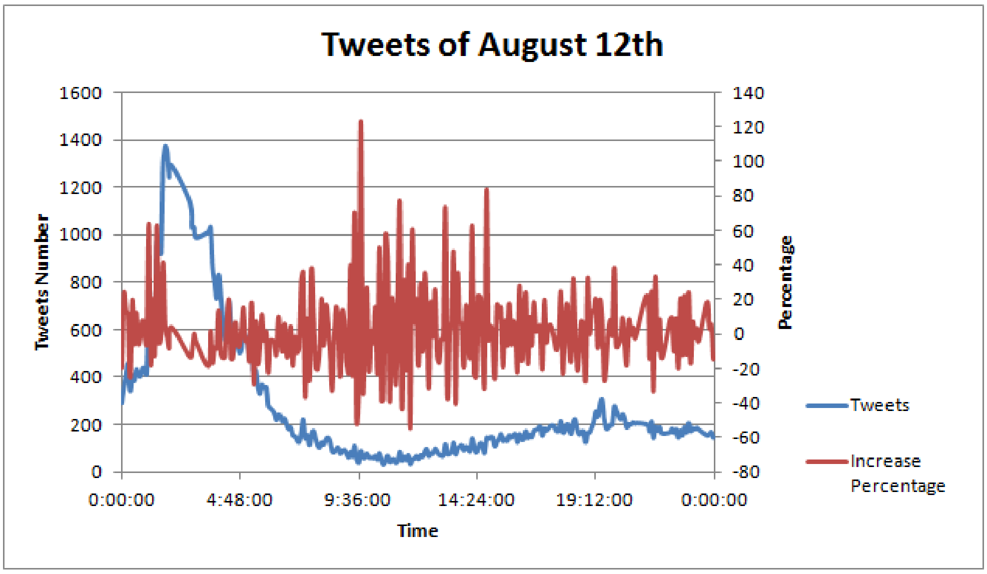


Figure 3.13: Tweets variation and increase percentage August 12th 2011

Figure 3.13 gives a better explanation of the errors of the increase percentage method. The blue line illustrate tweet numbers posted every 5 minutes on August 12th, and the red line, which varies like a signal, presents the increase percentage of each data point. The graph shows that a huge traffic increase occurred in the morning around 2:00 a.m. However the red signal indicates that the largest increase percentage is found around 9:30 a.m. In addition, comparing the actual tweets variation and the data value difference as a percentage makes clear that there is no rational proportionable relationship between tweets spikes and a large increase percentage. Therefore this method is proved improper in this case.

Considering the shape of a spike among normal traffic, the third trial focused on the slope of the increasing traffic over 15 minutes (other intervals were also tried). Generally in mathematics, steeper lines will result in larger slopes, and if the data varies smoothly, the slopes will be nearly flat. For each tweets data point, calculate the slope in last 15 minutes and compare the result with the preceding 24 hours. The mean and standard deviation give a measure of the slope in last 24 hours. Thus, the threshold for the slope is set to be (mean + 3*standard deviation) since only extreme large values are desired. A Perl script was created to implement the algorithm.

The following listing shows the resulting Twitter events identified for the period from August 20th to August 26th (the corresponding graphs can be viewed in the Appendix B):

```
Sat Aug 20 19:05:33 2011 ; 340
Sat Aug 20 19:10:41 2011 ; 475
Sat Aug 20 19:15:43 2011 ; 433
Sun Aug 21 20:20:47 2011 ; 810
Mon Aug 22 02:50:42 2011 ; 676
Mon Aug 22 04:25:45 2011 ; 771
Mon Aug 22 19:40:46 2011 ; 521
Mon Aug 22 19:50:47 2011 ; 813
Tue Aug 23 03:35:44 2011 ; 697
Wed Aug 24 23:55:47 2011 ; 307
Thu Aug 25 00:05:44 2011 ; 182
Thu Aug 25 15:50:42 2011 ; 356
Thu Aug 25 15:55:41 2011 ; 285
Thu Aug 25 16:00:42 2011 ; 242
Fri Aug 26 02:10:41 2011 ; 359
Fri Aug 26 02:40:43 2011 ; 531
Fri Aug 26 05:20:40 2011 ; 378
Fri Aug 26 05:25:42 2011 ; 369
```

However, these results provide only rough points when the tweets number appears to be trending upward. The time point lacks definitive validity since it does not necessarily reveal either the beginning nor the peak of an event. Moreover, the tweets amount sometimes floats up and down around the peak point; this may generate false positive alerts when several time points happen to represent the same event.

Considering the results for August 20th, Figure 3.14 indicates that there is only one sudden traffic increase in the day, but the script included three points in the list (19:05:33, 19:10:41, 19:15:43), which apparently refer to the same event. Similar mistakes occur also on August 22th (19:40:46, 19:50:47), 25th (15:50:42, 15:55:41, 16:00:42) and 26th (02:10:41,02:40:43 and 05:20:40, 05:25:42). In fact, the tweets variation looks messy and unstructured on August 24th, with tweets numbers floating in specific range. Although no distinct event exists, the script gives 23:55:47 as an event timestamp.

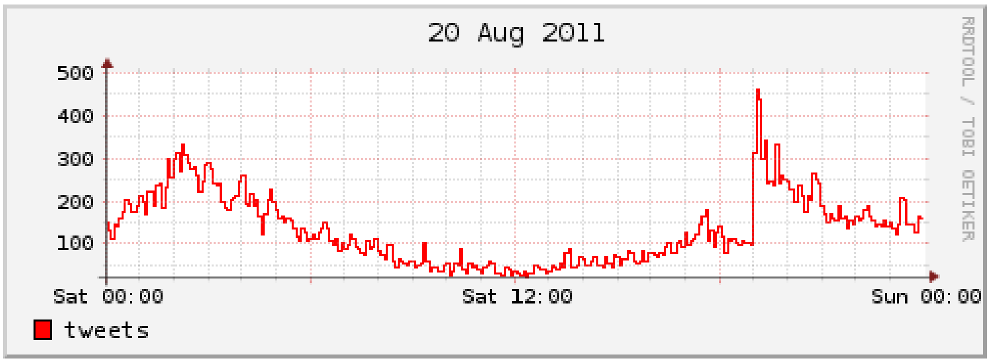


Figure 3.14: NFL Tweets Variation - August 20th

This analysis indicates that events should be identified in a more accurate way. Considering this project aims to analysis traffic spike connections on two different websites, the peak point of Twitter events might be the more significant if it can tell when a traffic peak is coming the on related forum. Accordingly, the following method combines two aspects based on calculating the slope:

- Firstly, the spike detector will compare subsequent data points with the filtered one with the large slope. If the following data values are larger than the current one, which means the tweets amount continued increasing, then the maximum data value will be selected as it is the real peak point. This algorithm also takes drifts with a small time period (15 minutes) into account; if the tweets number goes down first and then goes up again in 15 minutes, the script will only count the point with largest tweets amount.
- Secondly, there should not only be a threshold for identifying a sufficiently steep slope, but also another as a check for the tweets amount. When a spike occurs which actually has only a relatively small number of tweets, it will not be identified as an event; this kind of false alert will be ignored. The specific implementation here is to set a secondary threshold of (mean + 1.5* stdev) of the data values over the last 24 hours and use this statistic result to discard small values.

This second version of the Twitter events detector successfully filters data peak points in traffic spikes, and it provides more accurate results for future analysis. The following listing gives examples of how the developed algorithm works on the same data from August 20th to August 22th. Rough results are corrected and the false alert is removed.

```
Sat Aug 20 19:10:41 2011 ; 475
Sun Aug 21 20:15:44 2011 ; 1317
Mon Aug 22 02:50:42 2011 ; 676
Mon Aug 22 04:25:45 2011 ; 771
Mon Aug 22 19:50:47 2011 ; 813
Tue Aug 23 03:35:44 2011 ; 697
Thu Aug 25 15:50:42 2011 ; 356
Fri Aug 26 02:40:43 2011 ; 531
Fri Aug 26 05:20:40 2011 ; 378
```

In this way, Twitter events definition and filtering was achieved, via an ongoing process of identifying event peak positions within the Twitter data. The full script can be viewed in the Appendix C.

3.3.2 Forum Events

Defining Events

Forum event identification benefited from the experience of the Twitter events processing described above. First of all, posts data was imported and stored in an RRD database, and then plotted as for the Twitter data. The following figures illustrate posts variation on the ProSportsDaily.com forum. Forum posts do have days (e.g., Figure 3.15) in which the data expresses people's surging enthusiasm about American football as for the Twitter data. However, most of graphs are similar to Figure 3.16, indicating that the popularity of this forum is much less than Twitter. Posts amounts over 5 minutes are very small and vary quite frequently, making the forum traffic graph very spiky.

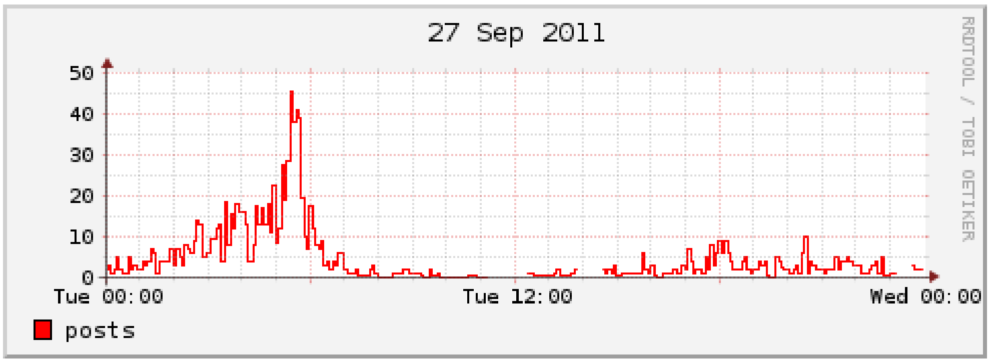


Figure 3.15: Posts Variation ProSportsDaily.com September 27th 2011

In Figure 3.16, for example, shows that on July 31th the largest posts values are only 9 or 10, but each data value is quite distinct, without any smooth upward or downward trend. All of the changes are random in the quantity, they last for a extreme short period (5 or 10 minutes), and spikes appear everywhere, all of which make it particularly difficult to identify events in this graph.

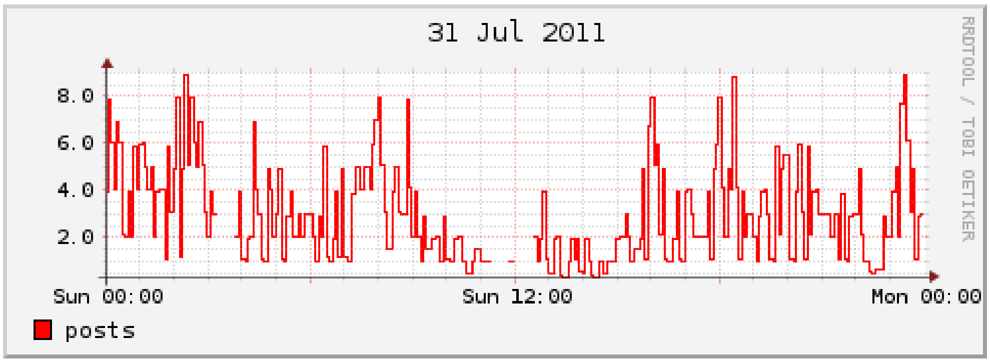


Figure 3.16: Posts Variation ProSportsDaily.com July 31st 2011

Processing forum data directly from the source file seems unfruitful. It seemed a better choice to handle the rough and messy data some way that reduces the noise and identifies the data's potential patterns, or at least give an explicit description of the posts oscillation. Two different smoothing methods for reducing the forum were analyzed:

- **Moving average**: For time-series data, the moving average is the simplest smoothing algorithm that help to reduce the noise by replacing values with the average of a number of consecutive points. In other words, traffic oscillation can be removed or at least reduced by averaging over several data points. Two times scale were used to smooth the forum posts traffic. Figure 3.17 is the result of the floating average over the last 15 minutes, and Figure 3.18 is the result of the running average over the last 30 minutes.
- **Calculate the sum over a short period**: If the change in the data is small and random, it may take time for an increase come into effect. Therefore calculating a data sum for each short time period may be efficient to tell posts increases. Figure 3.19 shows the graph of summing the data every 15 minutes and only keeping the sum results. This method reduces data points amounts from 288 to 94 per day.

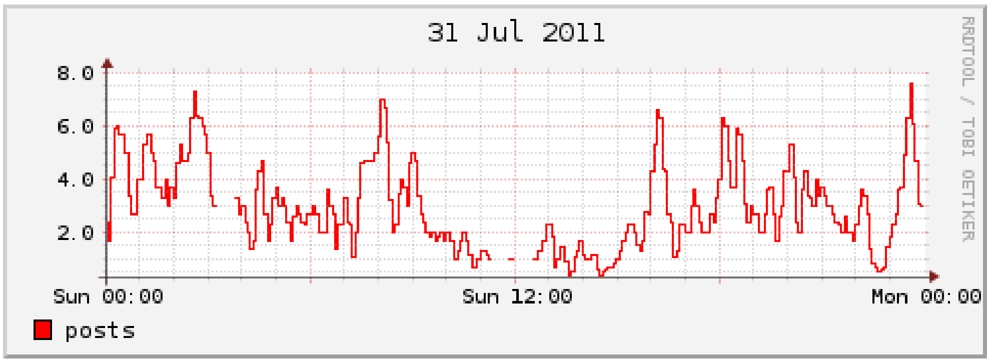


Figure 3.17: Smoothing: Moving Average over last 15 minutes

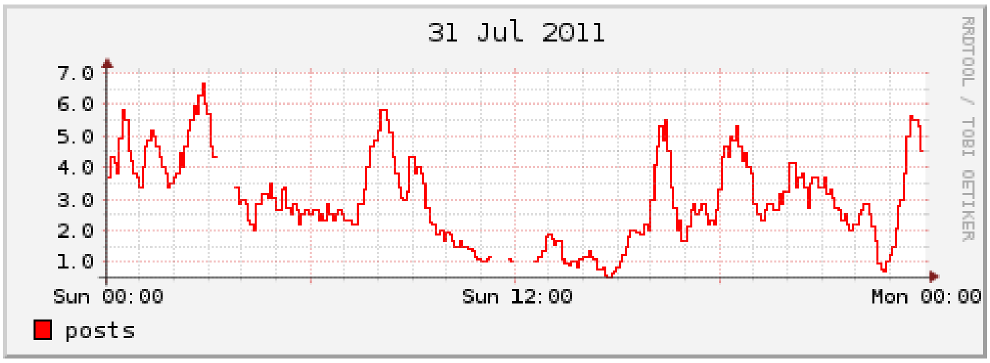


Figure 3.18: Smoothing: Moving Average over last 30 minutes

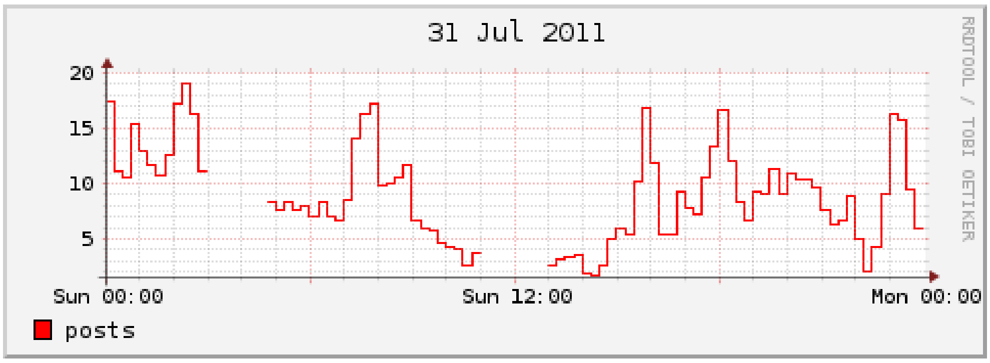


Figure 3.19: Smoothing: Data Summed over 15 minutes

Compare to the original graph, Figure 3.16, the posts traffic shape in Figures 3.17, 3.18 and 3.19 allows events to be identified in any of them. Previous spiky data are smoothed, noise is reduced to an acceptable level so that significant traffic increases are clear to see. The 30 minute moving average method seems to do the best job; the posts curve in Figure 3.18 looks very smooth, and almost all unimportant sudden jumps in the data are removed. But the disadvantage here is that the data peak point will float from the actual location because of averaging too much, and large spike might be turned into a much smaller one as it is affected by the small value data around it.

The 15 minute moving average result remains many annoying tiny spikes, but it is able to identify large increase areas. Figure 3.19 with data summed every 15 minutes looks stilted. It indeed reduces the negative influence from random data change and gives only significant spikes. However, this method is weak in timeliness, which means that filtered events will not be able to report exact peak points since data are summarized.

Taking all of this into consideration, the 15 minutes floating average method seems the best choice since data appears to lack fidelity over 30 minutes average and data details are lost in summed results. So the first round processing here is to smooth all the forum data with the 15 minute moving average method for future events analysis.

Event extraction

Given the previous work for Twitter events, forum events filtering is quite straightforward. The same slope calculating approach can identify all the upward trends in posts traffic and select only the peak data point representing an event.

In addition, it is unnecessary to report events in a continuous ongoing process because hopefully, after correlation analysis, the Twitter data will be able to predict forum events. So the event identifying work here changes slightly: for each posts point, calculate the slope in the last 15 minutes and compare the result with the whole day. The filtering threshold for slope is set to be (mean + 2.4*standard deviation). After experimentation to obtain the most reliable, solidest results, the threshold for the posting amount is set to be (mean + 1.5*standard deviation) since the posts counts are generally small and discard tiny spikes is desirable.

A Perl script was created to perform this job, beginning from the one developed for identifying Twitter events. The following listing shows examples of filtered events from September 17th to September 21st, where the second column is the number of posts. The corresponding graphs can be viewed in the Appendix D.

```
Sat Sep 17 01:50:08 2011;9;
Sat Sep 17 20:25:08 2011;6;
Sat Sep 17 23:35:09 2011;5;
Sun Sep 18 22:00:12 2011;32;
Mon Sep 19 05:50:08 2011;41;
```

```
Tue Sep 20 02:50:10 2011;32;
Tue Sep 20 04:45:09 2011;28;
Tue Sep 20 05:15:08 2011;31;
Wed Sep 21 21:30:09 2011;12;
Wed Sep 21 23:20:08 2011;11;
```

Once the event identification process development was completed for the forum case, the resulting script was only marginally different from the one for the Twitter data. The forum script was then applied to the Twitter data. When the results from the two scripts were seen to be nearly identical, the decision was made to use the same script for both processes. The final scripts can be viewed in the Appendix E.

3.4 Correlation Analysis

This section discusses the process of developing proper methodology for correlation analysis between Twitter and forum events. According to the problem statement in chapter 1, the question of this work is to determine whether Twitter data reacts to current affairs fast enough to be exploited as a predictive tool for corresponding websites. The next major item to consider is investigating if/how sudden traffic surges on Twitter are related to its topic-related forum traffic. Events from both websites are already identified in the previous section, so the next step is to deeply analyze the data of interest. Since correlation can be wide ranging and diverse technique, the following aspects should be taken into consideration:

- **Events time correlation**: First and foremost, this study will focus on the correlation in events' existence, which can be explained as "Do forum events reliably follow tweet events?" When a Twitter event is reported, a forum event is expected within a certain time period. Specific circumstances which support this situation will be analyzed.

- **Events magnitude correlation**: If time correlation of events exists, the expectation would be that huge Twitter spikes will result in larger forum events. So for each Twitter event of a certain size, the analysis of corresponding forum events will focus on the magnitude to see if they relate in some way.

- **Applying linear regression models**: Linear regression is always the first approximation used to model the relationship between a dependent variable Y and one (or more) explanatory variables (X), where the general formula for the mathematical relationship is Y = aX + b (a and b are constants). If correlation exists between Twitter and corresponding forums, a linear relationship is the most probable model. In this case, the single explanatory variable X stands for the Twitter events timestamp or scale, and Y represent forum events values, and simple linear functions are expected to be discovered between Twitter and forum events.

3.4.1 Time correlation

According to results from the event identification section, Twitter events and forum events are not simply one-to-one matched. Instead they have a many-to-many relationship. The graphs above and in the Appendix F show the tweets and posts variation in November. The red line represents tweets change amount while the blue line represents posts variation, over the course a day. The posts number was magnified 30 times so that it is convenient to compare both data sets' trends and patterns on the same graph.

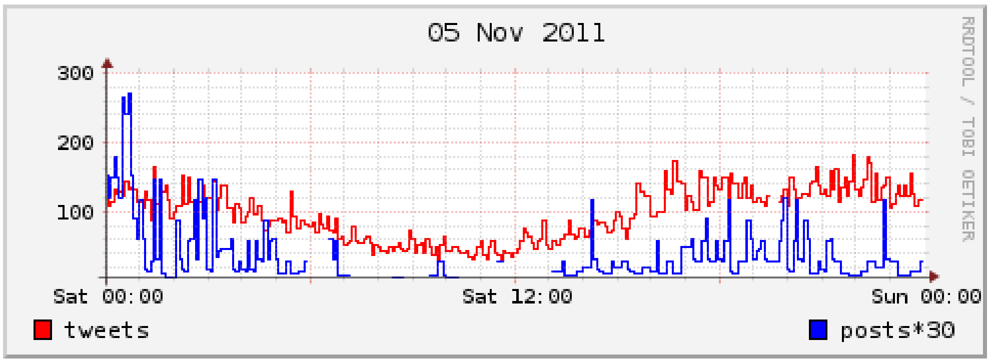


Figure 3.20: NFL Tweets and Posts Variation November 5th 2011

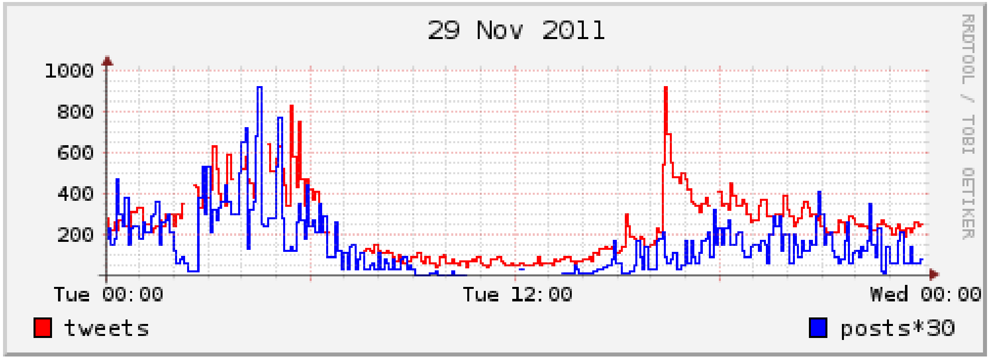


Figure 3.21: NFL Tweets and Posts Variation November 29th 2011

The following points describe the interesting relationships discovered in Twitter and forum events:

- Both Twitter and forum events might occur alone, without matching events from the other data set. Figure 3.20 shows a day where tweets varied smoothly without any spikes, while the posts exhibit a very clear peak around 1:00 a.m. Figure 3.21 shows the opposite situation: a large event occurred around 16:00 in Twitter data but it was not followed by anything of interest within the forum data.

- Forum events are expected to occur after Twitter events. However, as a matter of fact, sometimes forum spikes appears slightly earlier than the tweets peaks, or they occur at almost the same time. Several figures provide examples of this situation: in Figure 3.22, events around 6:00 in tweets and posts both have been reported; however, the exact timestamp from Twitter is ”Mon Nov 7 06:00:50 2011” and the one from the forum is ”Mon Nov 7 05:50:09 2011.” Figure 3.23 illustrates the same situation in which one can see that posts increase ahead of tweets, and they peak almost at the same point.
- As mentioned before, the forum data jumps up and down more frequently than tweets, so spikes exist more within the posts data. This can lead to Twitter events that are followed a two or more forum events.

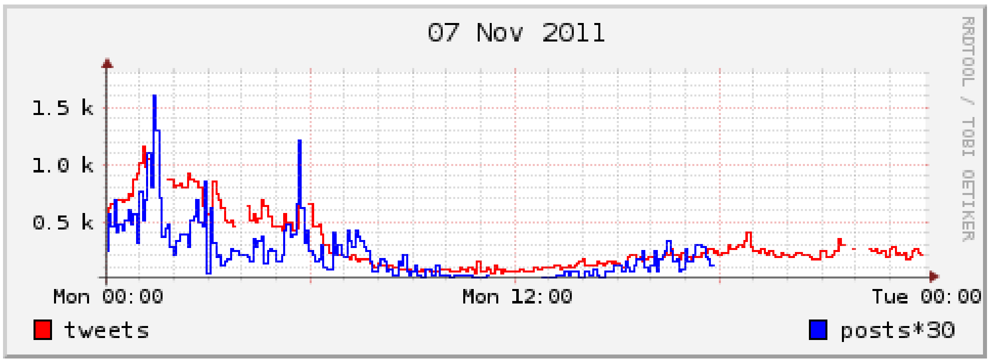


Figure 3.22: NFL Tweets and Posts Variation November 7th 2011

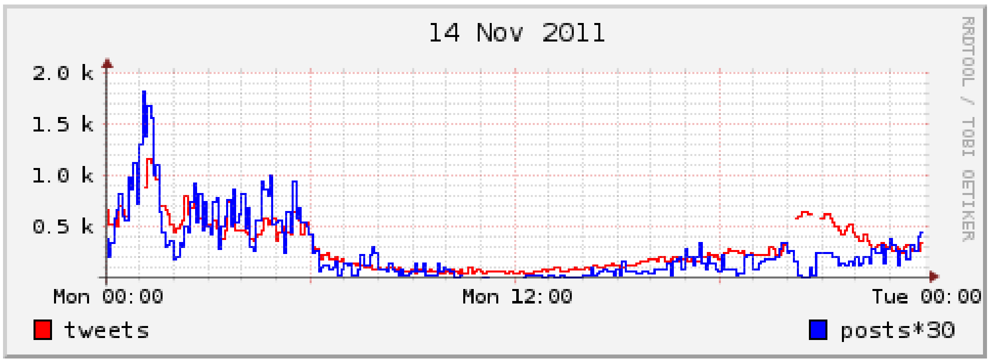


Figure 3.23: NFL Tweets and Posts Variation November 14th 2011

According to the observations above, the question arose as to how many Twitter events have relevant forum events. Do all Twitter spikes reliably predict or hint at increasing forum posts counts? One Twitter event could be surrounded by several forum events and their time sequence is somewhat uncertain under different situations. Accordingly, forum events occurring up to one hour before a Twitter event (but not close to a previous one) were discarded.

With the events filtering process, there were 420 Twitter events and 464 forum events in total. Forum events were inserted into continuous Twitter events to simplify the many-to-many relationship. For each Twitter event, all possible valid forum events should be picked out. 281 Twitter events are followed by one or more forum events within the time criterion. This means 66.9% Twitter events are correlated with forum events.

The time difference between each Twitter event and the following forum event represents the time duration it takes for forum posts increase after a Twitter traffic surge. Figure 3.24 illustrates the distribution of the time difference between Twitter and forum events.

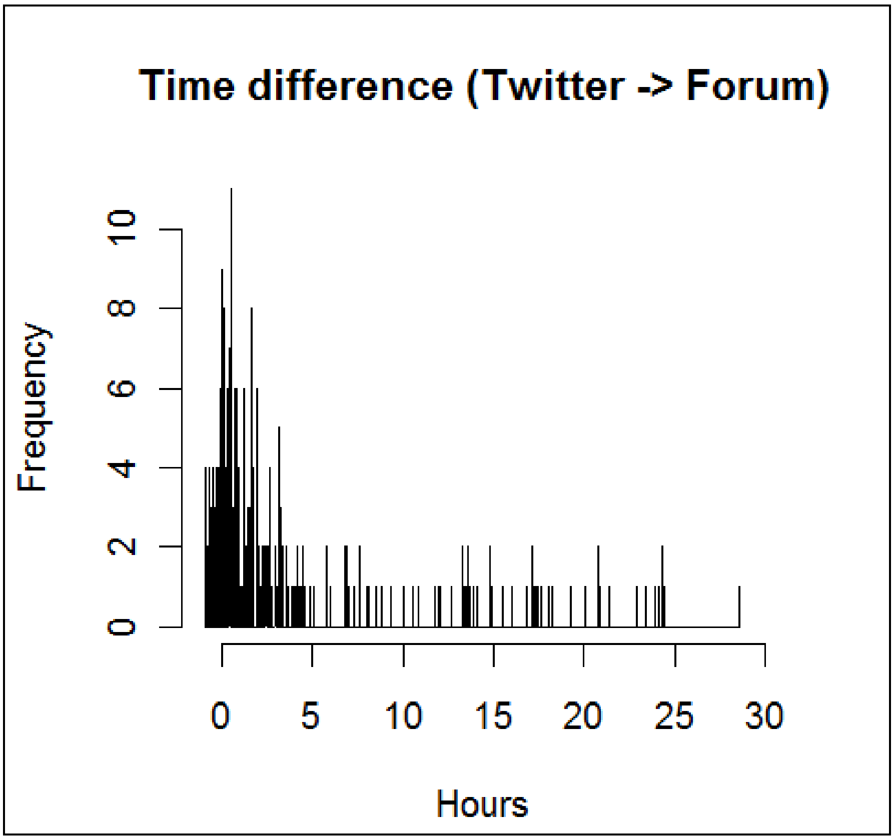


Figure 3.24: Time difference between Twitter events and forum events

The time differences are mainly located within a 6 hour block, but some spread randomly along the timeline. The maximum value of time difference can be as large as 28.57 hours. After investigating the long duration cases manually in the graphical results, these long time differences between Twitter event and forum event are quite meaningless to any correlation analysis, and no reliable relationships can be concluded in such situations. It is possible that these long gaps might be caused by data loss or false negative alerts which the event detector script failed to report. However, the first decision here is to discard all forum events which happened beyond 24 hours. 277 Twitter events remained which could be paired with relevant forum events.

The next step is to apply a linear regression model onto the events data, where events timestamps from Twitter are set to be the independent variables X while posts events timestamps are the dependent variable Y. The expectation here is that the time at which tweets spikes occur is proportional to the time at which posts increases appear. However, the linear regression model works with one-to-one correspondences. Even after handling the many-to-many relationship, some Twitter events are still followed by multiple forum events. So it is necessary to reduce the subsequent forum events to a single one corresponding to each Twitter event. Two different approaches to handling this were attempted:

- Keep the forums event which happened earliest after the Twitter events
- Keep the forum events which have the maximum posts after Twitter events

These two approaches will be discussed individually in separate subsections.

Keep the forum events which happened earliest after Twitter events

The first forum event could have the tightest relationship with its corresponding Twitter event since the time difference is the smallest among all possibilities. Figure 3.25 shows how the timestamps related to each other under this situation:

This was the first attempt to describe potential relationship between Twitter and forum events. R is the sample correlation, a measure of how strong a linear relationship exists between two variables. The more they are linearly relative, the larger the R absolute value will be. If a linear relationship in the sample data sets is perfect, then R is equal to 1 or -1. R^2 is known as the coefficient of determination, and it is used to measure the quality of the linear regression. R^2 indicates the extent to which Y can be explained by the regression line, and how good the regression lines fits the situation. So $R^2 = 1$ indicates the two variables are tightly relevant and the linear regression fits perfectly.

Figure 3.25 suggests that linear regression model suits perfect in our case, the first version result tells that forum events is correlated to Twitter events

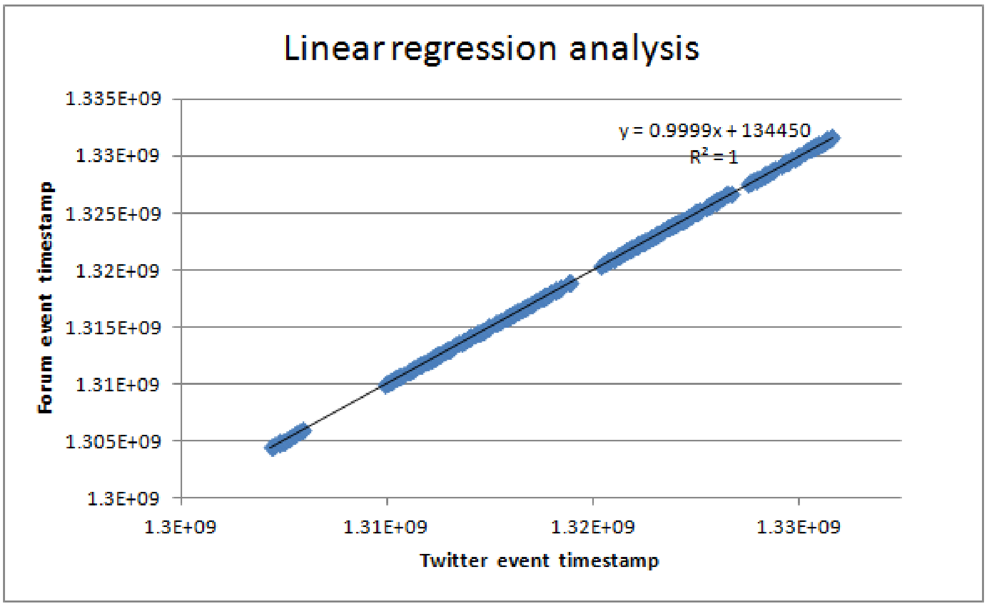


Figure 3.25: Linear regression result of NFL raw time correlation

with a linear function of Y = 0.9999X + 134450 (seconds), and the coefficient of determination $R^2 = 1$ indicates that this linear relationship is able to cover all Y values with given X.

However, this result also means all forum events happened almost 134450 seconds (37.34 hours) after Twitter events. This is completely incompatible with the previously-set valid time criterion of one day (any following forum events beyond 24 hours were filtered out). The problem comes from the large magnitude of the Linux timestamp, calculated as the number of seconds since the midnight of the 1st of January 1970 UTC. Examining the time scale of the X-axis and the Y-axis reveals that both of them have a huge base value of 1300000000 seconds, and any time difference between events is tiny when compared to it. The influence of huge base number must be reduced so that the actual data points distribution and exact relationships can be viewed. Since linear relationship means that two variables are proportional to each other, increasing or decreasing one variable n times will cause the corresponding increase or decrease of n times in the other variable. For example, if Y = a*X + b, then ΔY = a*ΔX.

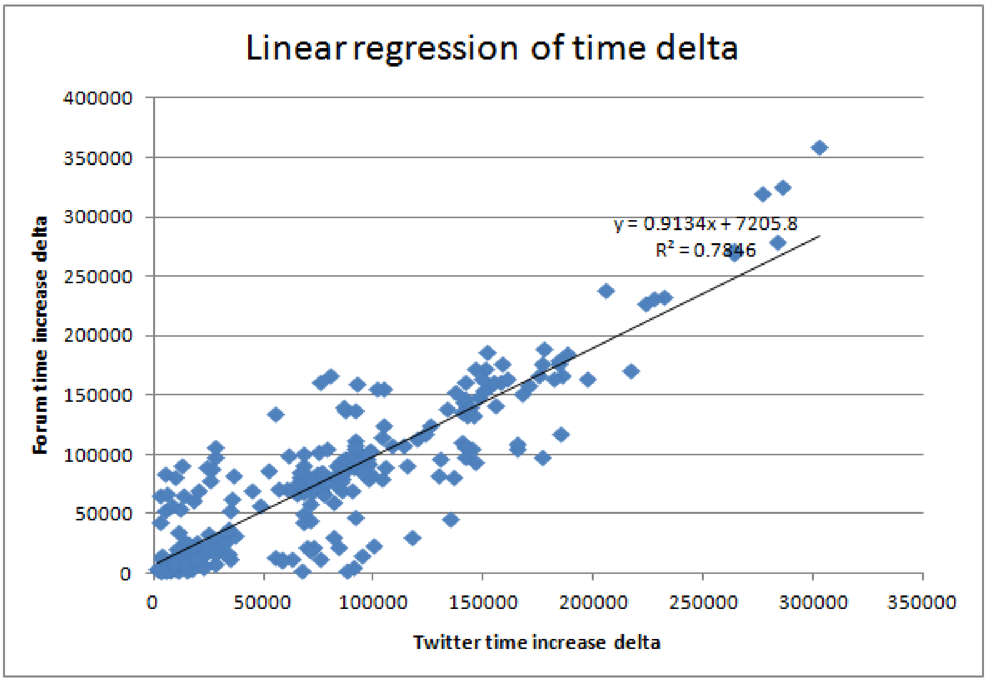


Figure 3.26: Linear regression result of NFL time delta correlation within 24 hours

So calculating the time-delta between each two subsequent Twitter events and forum events and then plotting both delta value sets should reduce the impact of the huge Linux time scale. Figure 3.26 shows the newer correlation with ΔY(Y = Forum event timestamp) and ΔX(X = Twitter event timestamp). This time, time scale of X-axis and Y-axis have been reduced to 400000 seconds and data points are now distributed around the linear trend line with distances from each other, so the findings should be more reliable now.

The result shows there exists a linear relationship between Twitter and forum data and can be interpreted by a linear function (ΔY = 0.9134*ΔX + 7205.8) with R^2 = 0.7846. This relationship is able to roughly explain 78.46% forum events with the given Twitter events, so the majority of the data was covered.

However, according to time difference histogram in Figure 3.24, most of forum events were found within 6 hours after Twitter events. Some events happened in the same day, but their correlation must still be questioned. For example, on August 16th (Figure 3.27), after Twitter event was reported at ”05:25:40”, two forum events were identified at the time points of ”16:15:07” and ”22:35:10,” with time differences of 11 hours and 17 hours respectively. After such a long time duration, the Twitter event has long lost the capability to generate an alert for a posts increase on the forum. In other word, correlations between Twitter and forum across such large time gaps seem unreliable and unhelpful; only tight time-related events can create convincing correlation.

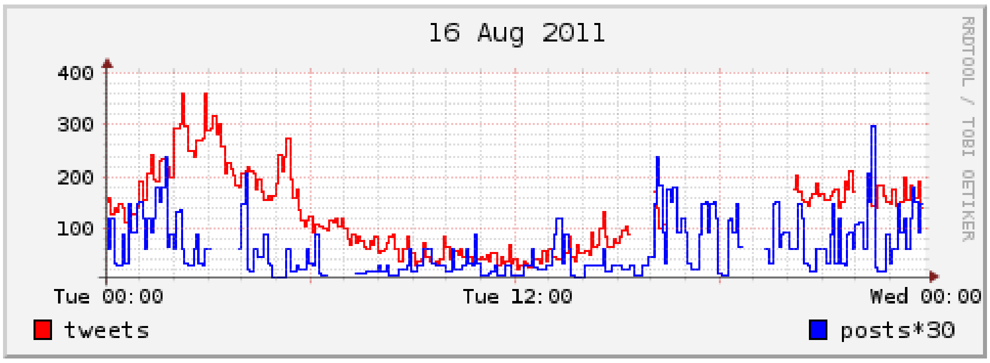


Figure 3.27: Long durations between Twitter and forum events

So a decision is made here, when analyzing events time correlation, to consider only forum events with a time differences within -1 to 6 hours as related events. This decision also considers those forum traffic spikes which happened before or at the same moment with Twitter events (restoring them to active consideration), which should give added reliability to the correlation analysis.

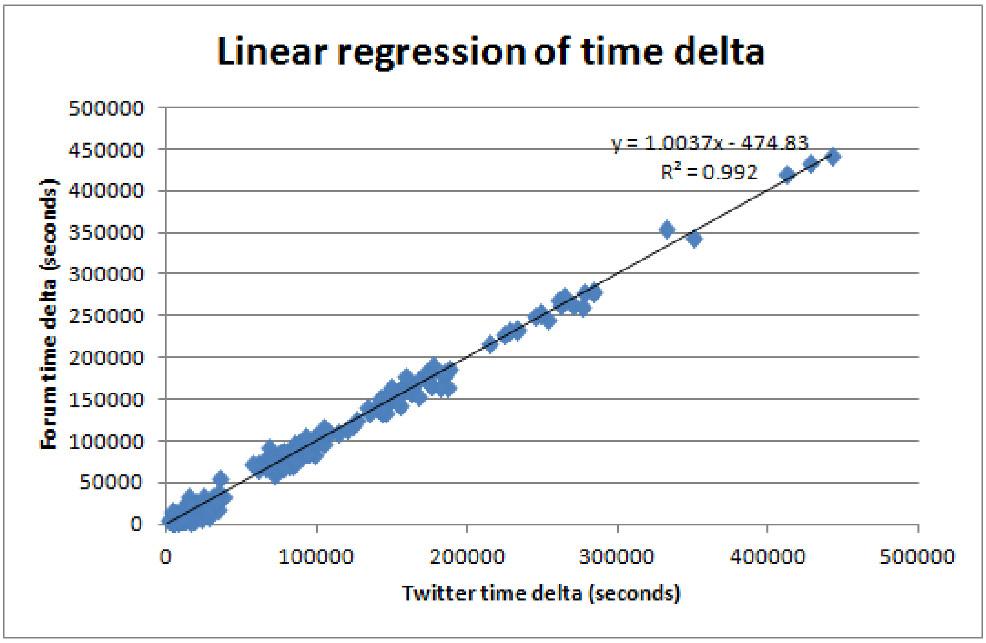


Figure 3.28: Linear regression result of NFL time delta correlation within 6 hours

Figure 3.28 shows linear regression results after modifying the criterion of the events time gap from 24 hours to 6 hours. The results shows there exists a linear relationship between the delta value of Twitter and forum events' timetamps, which can be described with the linear function (ΔY = 1.0037*ΔX - 474.83). R^2 = 0.992 means this function can explain 99.2% of the ΔY values in the data set and therefore confirms a near perfect fit of linear correlation in this situation. The large R value (close to 1) also concludes that the relationship not only exists but is very strong.

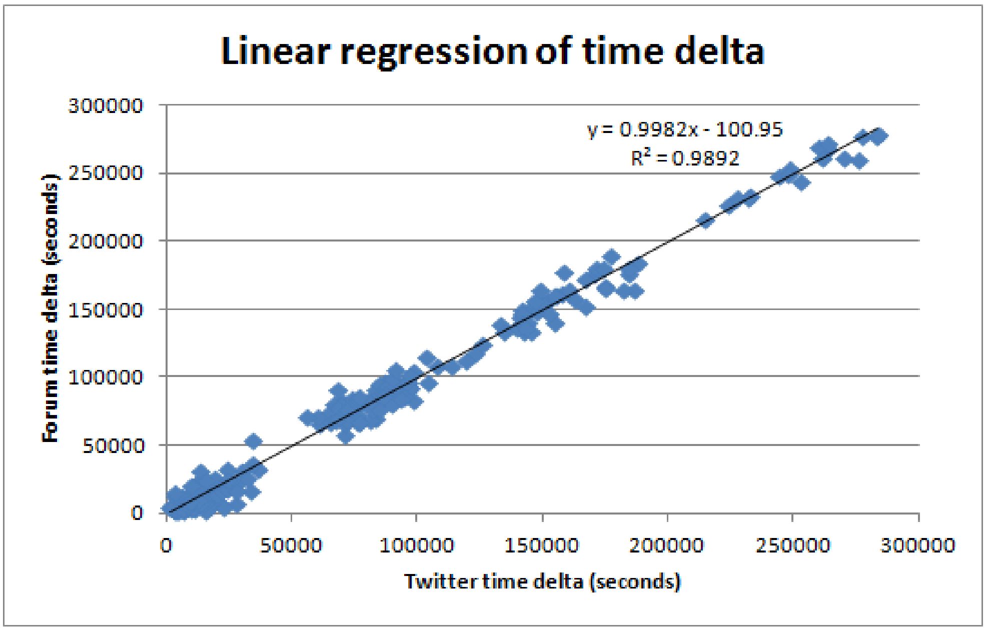


Figure 3.29: Linear relationship in NFL data (first following forum event)

Narrowing down the time scale to 300000 seconds by discarding large data values which might affect the regression result causes the linear function and sample correlation R to look even better. The results in Figure 3.29 indicates that majority of time increase deltas of Twitter events and forum events belong to a linear relationship ΔY = a * ΔX with a acceptable floating range (intercept= 100.95). According to linear theory, it is possible say that Twitter events indeed relate to forum events, although they are exactly correlated is not yet known.

The time difference between each Twitter event and its corresponding forum event can somewhat explain the expected time scope in which forum event might appear after a Twitter event. With the ”earliest forum event counts” approach, there are 212 events pairs in total, and the average of their time gap is 1.16 hours, with the standard deviation equals to 1.51 hours. Considering the 2 delta principle of normal distribution statistics, this result points out that among all events pairs, 95% forum events happened within 4.18 hours after Twitter events, and it just confirm the solid correlation between tweets and posts.

Keep the forum events which have the maximum posts after Twitter events

The forum events with the maximum number of posts after Twitter events might provide a more accurate time point for correlation analysis since the existing forum traffic appears spiky, and one increase might be consist of several consecutive spikes.

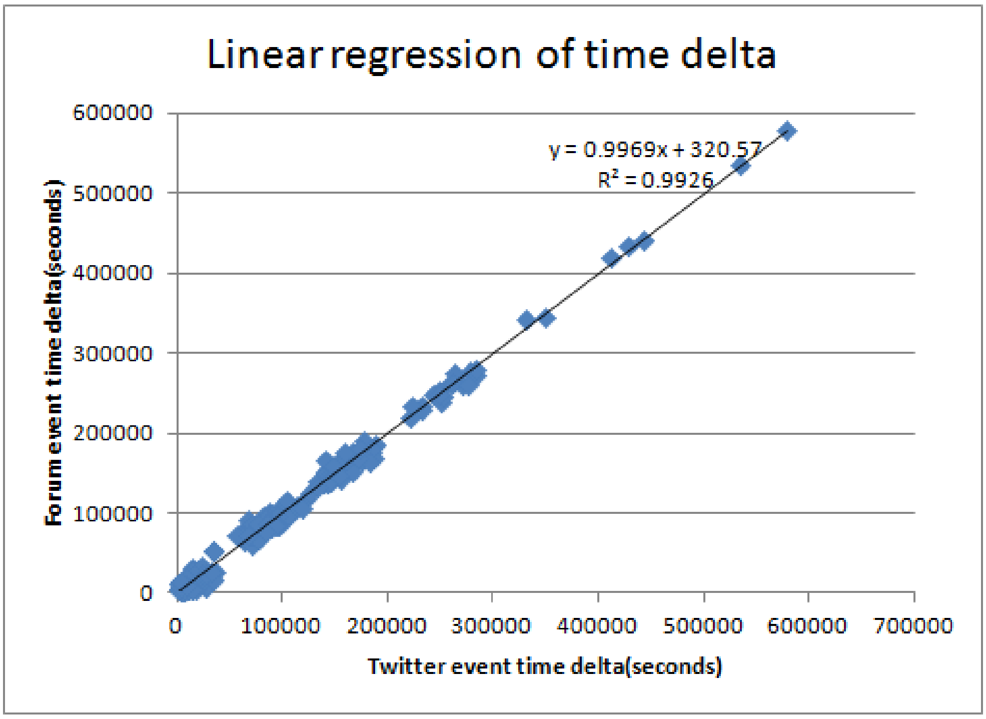


Figure 3.30: Linear relationship in NFL data (maximum value forum event)

Figure 3.30 shows the linear regression result between Twitter events and the forum events with maximum posts. The linear function (ΔY = 0.9969*ΔX + 320.57) with $R^2 = 0.9926$ confirms the solid correlation in this situation. The mean value of all time difference between Twitter events and forum events is equal to 1.67 hours, and the standard deviation is 1.67 hours. Since 95% forum events follow Twitter events within 5.01 hours, this result is slightly looser than the previous one since the selected forum events give priority to the posts value's increase in size instead of closeness in time.

The events correlation analysis was accomplished with two possible results, both of which have advantages. Forum spikes occurred frequently, and sometimes small, closely clustered ones can be combined into a single larger one. The earliest forum events indicate the rising time for forum posts rates, and the maximum forum events suggest the peak point of the entire increase period. Figure 3.31 illustrates this situation. Forum events were reported 3 times at ”04:10,””04:35” and ”05:10,” while only one relevant Twitter event was found, at ”5:10.” It is easy to see that all those forum spikes can be merged into a single larger one with the peak at 04:35. According to the events matching rules, if the earliest event was kept, the result should be ”04:10,” and if the biggest event was kept instead, the result will be ”04:35.” Since the time distribution of succeeding forum events and the correlation results obtained above are only slightly different, either situation is acceptable.

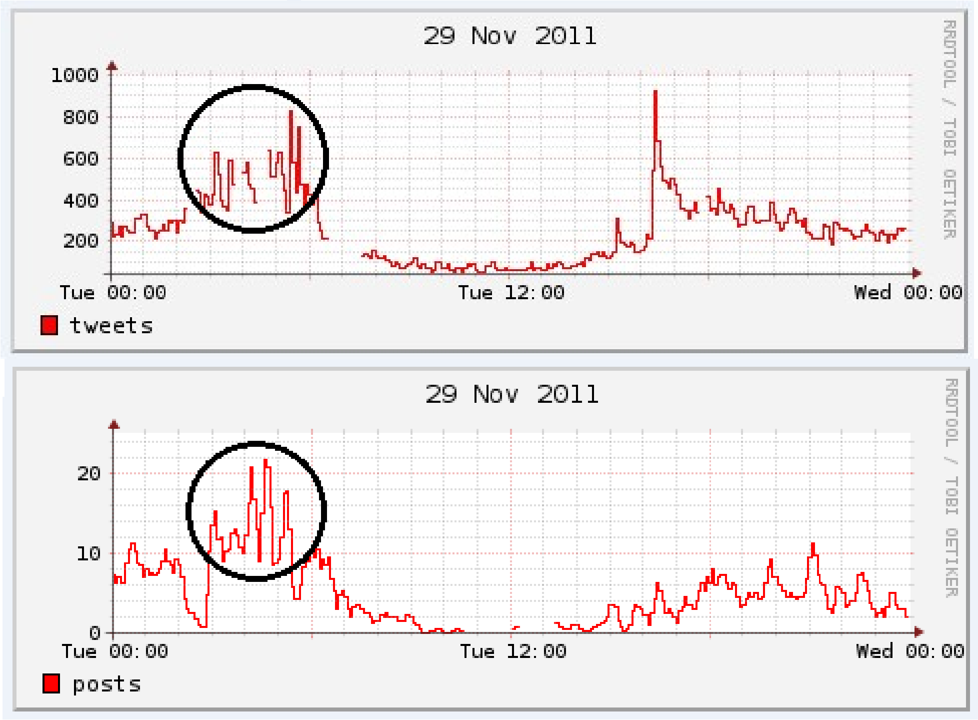


Figure 3.31: Selecting posts events according to different criteria

3.4.2 Magnitude correlation

The magnitude analysis of tweets and posts must based on events pair which already proved to be correlated in time. Huge Twitter spikes are expected to correspond to large posts increases. If tweets and posts amounts varies proportionably, then Twitter traffic is able to provide information about the timing and size of posts spikes.

The method used is to plot the tweets number and posts number to see if they have any kind of relationship in their variation. Figures 3.32 and 3.33 illustrate how tweets and posts magnitudes correlated within the two previous analyzed situations. They make it clear that the data points are spread randomly, and the data range is completed unpredictable. The graphs also illustrate how large tweets values can be in pair with small posts values, for example on Feb 5 2012, when the Twitter event was reported at 17:46 with 1374 tweets, but the following event had only 10 posts. Therefore, according to this analysis, there is no possible correlation in the size of Twitter events and corresponding forum events.

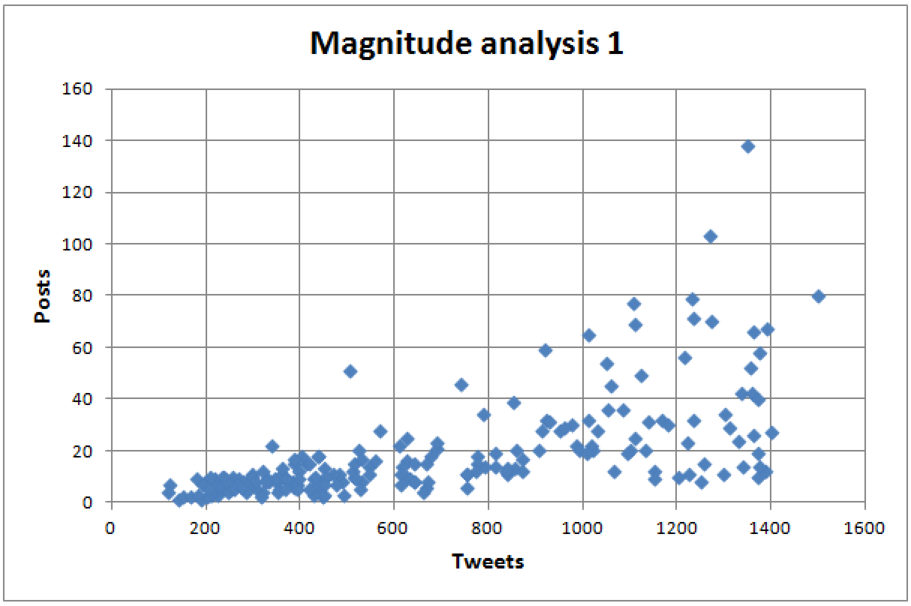


Figure 3.32: Traffic magnitude analysis in NFL with the max forum events

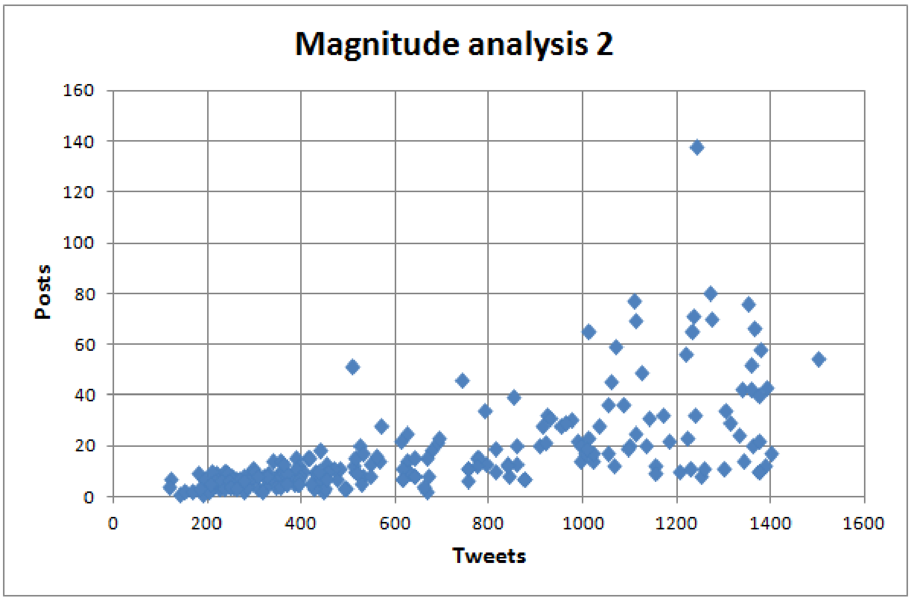


Figure 3.33: Traffic magnitude analysis in NFL with the first forum events

In conclusion, this chapter proposed a data mining methodology to analyse potential traffic correlation between Twitter and topic-related forums. The mining procedure mainly consisted of three parts: data cleaning, events identification and correlation analysis. NFL tweets and posts were used as the training data sets, and corresponding events and correlation results were presented in each section along wit the development of the methodology.

Chapter 4

Results and Analysis

This chapter will present events and correlation results and analysis for the data relating to the NFL (National Football League), the NBA (National Basketball Association) and MMA (Mix Martial Arts). Full analysis was not possible for the data related to the NHL (National Hockey League) and boxing since there are too many gaps inside the original forum data sets.

The data mining methodology was described in the last chapter, and the training process was done using the NFL data sets. In this chapter, these data mining procedures of data handling, event identification and correlation analysis will be employed for the NBA and MMA data sets. All results and analysis are presented in following sections.

4.1 NFL Results Analysis

NFL data was collected from May 2nd 2011 until March 14th 2012. It is clear from the previous results that the NFL Twitter data have the expected pattern of flow with sudden spikes, while the forum data presents complicated and totally unpredictable patterns. Since the NFL season starts from early August with a 4-game exhibition period, and the regular season runs from September 2011 to December 2011, the original data is fortunate to cover the whole match season during which sports fans can act on their burning enthusiasm. This provides a large advantage toward reaching the goal of this project. The results have already shown that data variation has similar behavior to public events and emotion, illustrating that human enthusiasm indeed turns into traffic bursts on corresponding web sites.

The NFL data was analyzed in detail and the correlation results between Twitter and the ProSportsDaily.com forum were presented in Chapter 3. From these results and especially Figures 3.28 and 3.30, it is clear that Twitter traffic surges are related to forum posts spikes in time with a near-perfect linear relationship. Among all events pairs, 95% of forum events appear within 5 hours after Twitter events. So there exists a solid statistical relationship between tweets and posts spikes. These findings and correlation results will now be analysed.

4.1.1 Accuracy of events identification

The first thing to pay attention to when evaluating the NFL results is the accuracy of event identification, since the more precisely the events peaks are reported, the more reliable the correlation results are. Two thresholds were used in the events detection tools for Twitter and the forum: one checks the slope to identify increasing traffic, and the other threshold is set to filter out peaks with small numbers of tweets or posts to reduce unimportant spike alerts. However, event identification is still crude, and there are two possible faults in the results:

- **False positive**: tweets/posts spikes which should not count as an event are reported.
- **False negative**: tweets/posts spikes which should be identified are missing from the results.

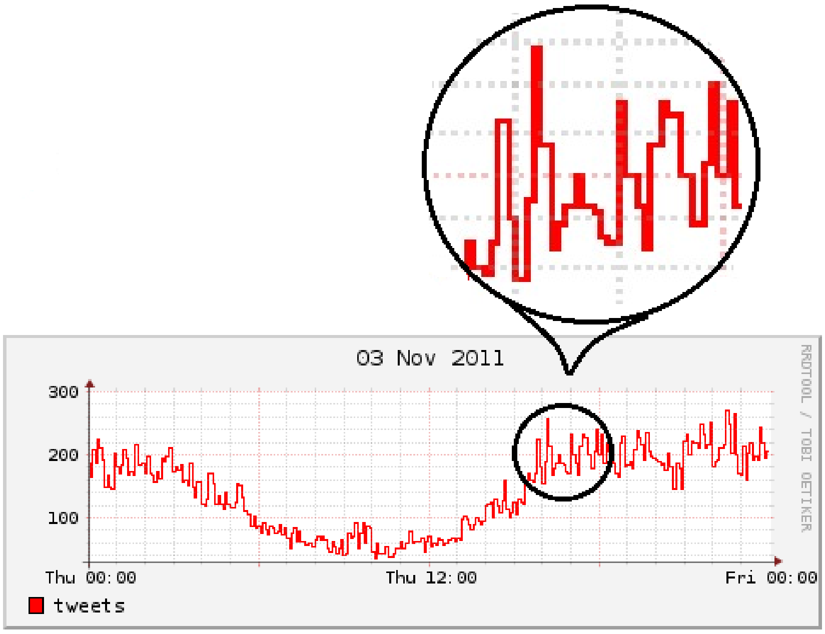


Figure 4.1: Magnification of a false positive tweets increase on November 3rd 2011

Manual inspection of the events identifying results provides a better understanding of how the algorithm actually worked on the NFL data sets. Events were compared manually to the RRDtool graphical results to check the goodness of the event detecting tool results. Examples will be taken from the November Twitter events.

Figure 4.1 illustrates a false positive result. According to the events identification tool, the time point of ″16:15:49″ was reported an event on November 3rd. The subpicture inside Figure 4.1 magnifies the identified increase period in the tweets. It makes it clear that the tweets curve is actually a small one compared to other tweets around it. Twitter traffic appears spiky with frequent variation in the entire day, but none of them is large or distinct enough to be a real events. This kind of false alerts is usually caused by small magnitude oscillating data which makes the ″slope threshold″ small and easy to reach.

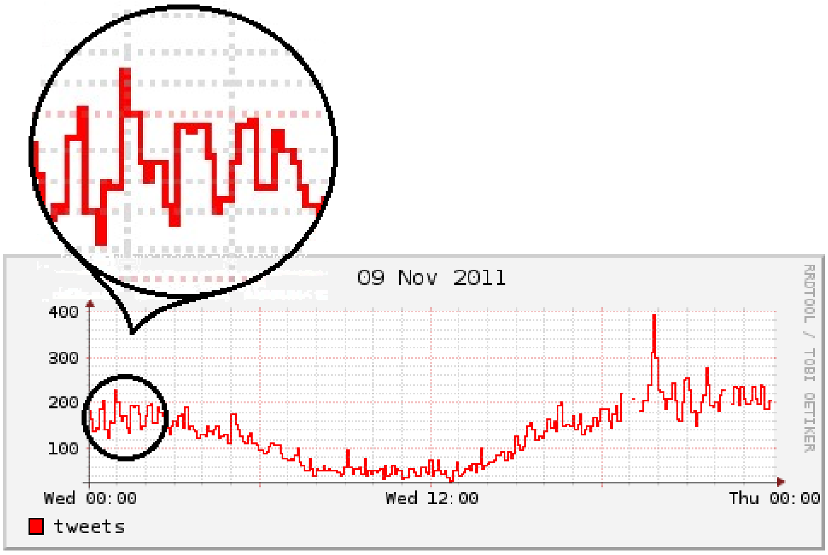


Figure 4.2: Magnification of a false positive tweets increase on November 9th 2011

Figure 4.2 is almost the same situation. The magnification inset shows the tweets variation of a false event alert at ″01:00:54.″ Even though there is a genuine Twitter event at 19:00, the filtering threshold was dragged down a lot by small tweets spikes during that day. Figure 4.3 illustrates a different situation. Tweets increases at ″23:00:52″ and ″23:30:51″ are both identified as events. However, it is clearly seen from the graph that the spikes' times are very close to each other and give a sense of gradual growth. The events detecting strategy only identifies the peak point of an event, which in this situation occurs at ″23:00:52.″ However, it should actually be combined with the larger peak at ″23:30:51″ and treated as one. This kind of false positive is mainly due to the approximations made in the events identification script.

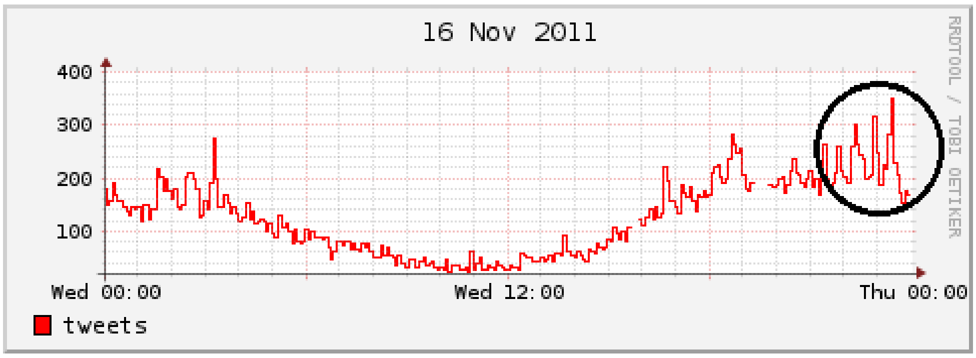


Figure 4.3: A false positive from consecutive peaks on November 16th

False positive errors might be corrected later when matching the Twitter and forum events into pairs. For example, when two Twitter events occurred within a short period of time, as in graph 4.3, but the forum event followed the second one, then the first Twitter event is ignored. In this case, the false positive is easier to handle and won't cause too much negative impact on the final correlation results.

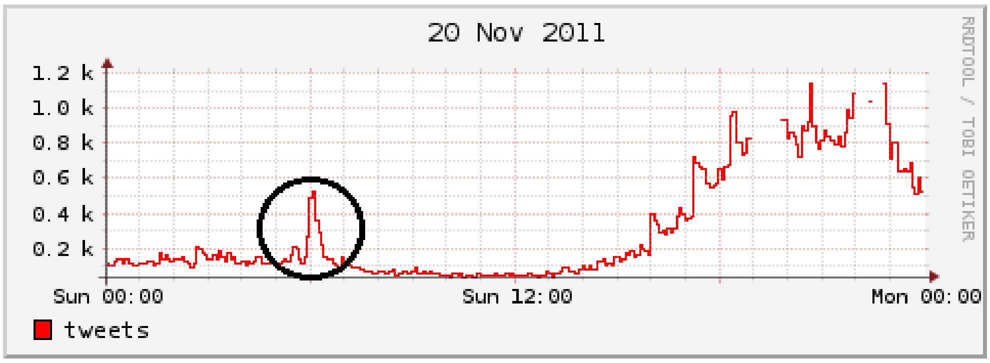


Figure 4.4: A false negative by strict tweets amount criteria on November 20th 2011

Besides false positive errors, false negative errors also exists, and this kind of fault is mainly caused by the strict criteria set by the filtering thresholds. Figure 4.4 illustrates a missing Twitter event due to its tweets count being smaller than the (mean + 1.5*standard deviation) threshold, which was distorted by the very large and long duration increase at the end of the day. On November 7th (Figure 4.5), the script failed to identify a tweets spike at "01:10:49" because the increasing traffic again can't reach the (mean + 3*standard deviation) slope threshold.

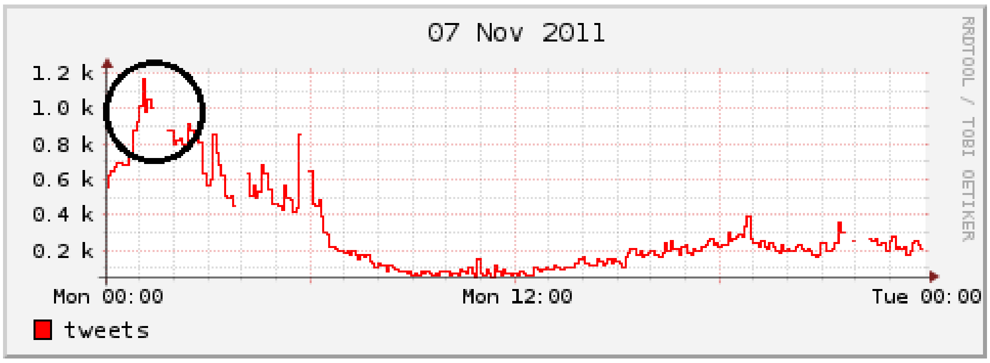


Figure 4.5: A false negative by strict slope criteria on November 7th 2011

Table 4.1 shows all the missing spikes and false positives found in the events results for November 2011. In this month, there are 57 Twitter events identified in total by the events detection tool. the false positive error rate is 8.8% and the false negative error rate is 5.2%. It is reasonable to say that the majority of events identification results is satisfactory and thus provide reliable data to the correlation analysis. In addition, it also suggests Twitter events amount might be larger than the actual number one would really like to see.

False positives	False negatives
Nov 3 16:15:49	Nov 7 01:10:49
Nov 9 01:00:54	Nov 20 06:05:43
Nov 14 03:40:49	Nov 26 10:16:30
Nov 16 23:00:52	–
Nov 23 19:20:48	–

Table 4.1: False alerts in NFL Twitter event identification, November 2011

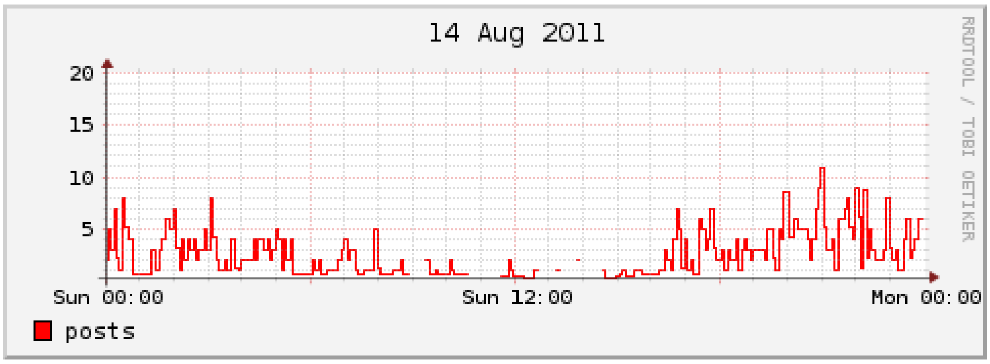


Figure 4.6: Major increase situation in posts

In general, the forum data is smaller in magnitude and contains many more gaps than is desirable, especially during the non-match season. In May 2011, the daily posts average is only 85.80. Assuming the data was collected every 5 minutes without loss, then the average posts every fetch was only 0.30. Even in the best month, January 2012, where daily average posts reach the peak of 1246 a day, the posts amount for every fetch was only 4. Therefore, the forum data was not sufficient to reveal significant increases. Events in forum posts cannot be as strong and obvious as the ones in the tweets, and events identification results are crude. Figure 4.6 illustrates the general situation of event identification within the forum posts data. There are no sudden spikes asare present in the tweets data, and the only increase trend was found around 21:00. Moreover, the posts numbers during the entire day were very small (11 was the maximum). The events detection tool identified ”21:05” as the event.

4.1.2 Event correlation confidence

After the majority events identification have been confirmed as reasonable, the next important step is to to observe the confidence level of correlation results to ensure the relationship discovered from Twitter and ProSportsDaily.com forum is reliable. According to the results, the linear mathematical model can be applied to Twitter and forum events and it indicates the high possibility that a forum posts increase will appear around a tweet spike. However the predictability of tweets traffic with respect to forum resource consumption is still unknown.

Using strict criteria on events filtering, there were available Twitter events and 464 valid forum events in total. 281 (66.9%) Twitter events have time-relevant forum events without any time scope restriction and 277 (98.5%) of the above forum events happened within 24 hours after Twitter events. The overall time difference gives 4.26 hours as average with a standard deviation of 6.41 hours. Since the long duration between Twitter and forum event lacks reliability in correlation, time scope two related events is restricted within -1 to 6 hours. With this criteria, results shows 265 Twitter events have one or more time-relevant forum events, and this number is 63.0% of all Twitter events.

After all Twitter and corresponding forum events have been matched into one-to-one pairs, there are 212 event pairs when the first forum event is used (Figure 3.28 in Chapter 3) and 206 events pair when the largest forum event is used (Figure 3.30 in Chapter 3). This means that only 50.4% of Twitter events are able to give alerts of an traffic increase happening on the related website. Half of the Twitter events are mismatched, which is to say either they failed to find a corresponding increase trend on forum, or they are ignored somehow. By conducting manual investigation on events matching results, several major reasons lead to the small amount of event pairs:

- **No corresponding event exists**: The major loss of events pair comes from missing corresponding posts increases, even though this is the most undesirable situation. The table below indicates the pattern of successful event matches by the different hours in a day:

Time	Twitter E	Forum E	Pairs	Match%
00-03	98	99	62	63.2%
03-06	81	113	33	40.1%
06-09	12	19	6	50.0%
09-12	2	1	0	0
12-15	2	1	0	0
15-18	39	33	19	48.7%
18-21	101	82	54	53.5%
21-24	85	116	38	44.7%

Table 4.2: Events distribution during periods of a day

The start time of NFL match is generally between 1:00 to 8:00 in the afternoon of American Eastern time. Since data was collected by a server placed in Norway, 18:00 in the afternoon until 06:00 in the morning Norwegian time should be the most popular periods for Twitter and forum activities. Results in Table 4.2 verify this. Twitter events and forum events mostly happened during periods of 00-06 and 18-24. Each subperiod included approximately 20% of the events, and the remaining 12 hours had only 10% in total. However, these results cannot give strong evidence that tweets and posts behavior were affected by NFL games, because the 12:00 a.m. to 12:00 p.m was also the time during which humans participate in network activities most. In addition, this table also indicates that the NFL match time does not affect the correlation between Twitter and forum events. The matching percentage is only slightly different during each time periods, and results cannot be improved within current data.

- **Influence from the game season**: It has been shown before that the posts amounts do not exhibit distinct data increases, especially in the non-match season during which forum users have little passion or news to post.

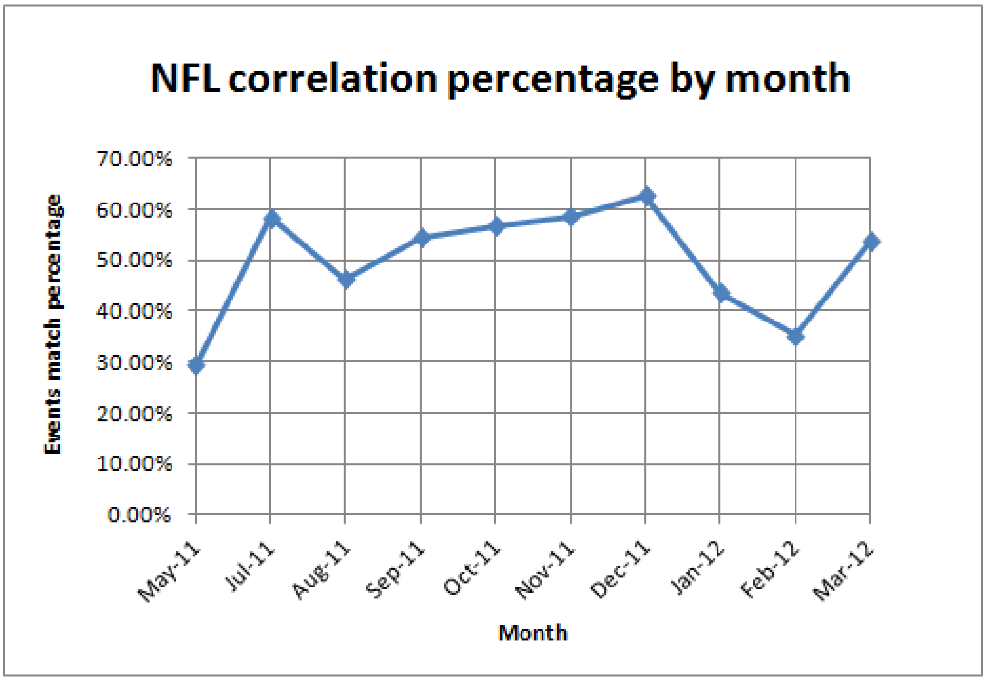


Figure 4.7: NFL events correlation percentage by month

The impact of game season is confirmed by investigating the correlation results. Figure 4.7 shows there are more connections existing between Twitter traffic spikes and forum posts during the match periods. The percentage of correlation goes up from July and reaches the highest point in December. May, which is a completely idle time for the NFL, gives the worst correlation results, with 70% Twitter events missing a match. In addition, surprising decrease in events correlation is also found in January and February 2011 during which the Championship games was played, the the up-going correlation in March can be explained by the NFL draft publishing.

- **Intense Twitter events are ignored by matching**: Redundancy exists in the events identification results, as shown in Figure 4.3, and ”extra” events are filtered out during the matching process in order to achieve one-to-one relationship. Events are identified only with the peak point. Based on this, Twitter events have been proved solid, but forum posts failed to generate distinct and concrete spikes. Therefore small spikes which might describe the same increase in posts might be reported multiple times.

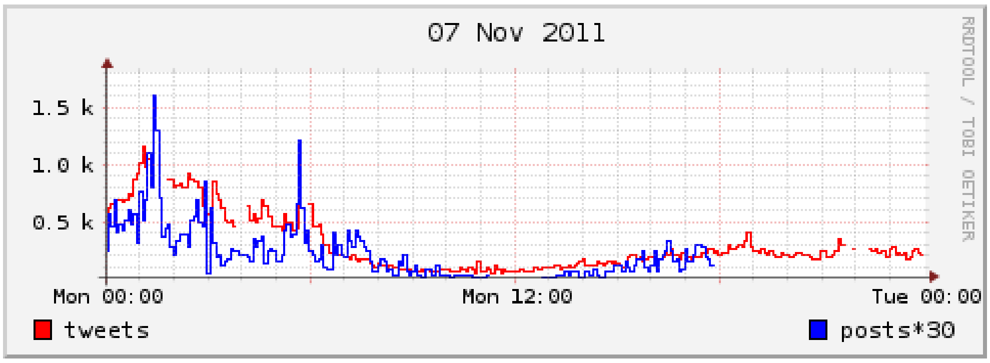


Figure 4.8: Linear regression results of NFL data time correlation

- **Missing events**: These are events which was failed to be identified in both data sets. Referring to Figure 4.8, the Twitter event at 1:10 was not identified, and therefore the forum event at 1:30 is mismatched. This kind of situation has been proved to be very rare in Twitter events, but errors in forum result appears more often.

With respect to those 50.4% valid events pairs, the results indicated a strong linear relationship between tweets and posts increase times. Since the delta value of events timestamps was computed to reduce the negative impact of the large Linux time base, the results suggest proportional varying of Twitter events time points and forum events time points.

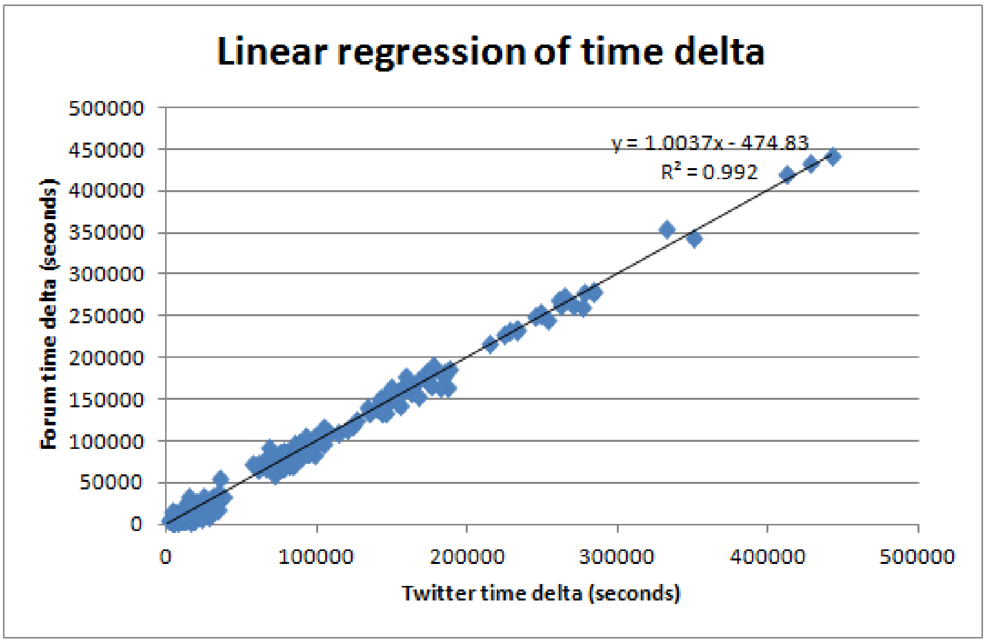


Figure 4.9: Linear relationship in NFL data (first following forum events)

Figure 4.9 proved the existence of equation "ΔY = a * ΔX - 100.95" where "a" equals to 0.9982, and the intercept 100.95 on the Y-axis indicates it is small enough to be acceptable that ΔX and ΔY are proportional related. This trend line suggests a linear correlation "Y = a * X + b" between Twitter events and forum events in time.

Assuming a value for a of 0.9982 is close enough to say ΔX and ΔY are 1:1 proportional related, then b corresponds to the time difference between related events. Accordingly, it is the key element to describe the expected scope for the location of a forum event after a Twitter alert. According to the statistical results in the last chapter (Average(b)=1.16 hours, StDev(b)=1.51 hours) with respect to the time scope in the events correlation, 95% of correlated forum events appear within 4.18 hours after Twitter events. This value goes down to 78.7% when counting only related forum events within 2.00 hours (both cases include ones occurring before tweets spikes). It is clear to see that the majority of forum events happened close to Twitter events in time.

When filtering out the related but not predictive event pairs, and speaking for the predictability of Twitter events, only 188 Twitter events were able to suggest an upward trend in forum posts, that's 44.8% of all Twitter events.

These results found interesting correlations between tweets and posts variations. However, strong evidence is lacking that would enable stating that Twitter traffic spikes are able to predict the increases in posts in the corresponding forum.

4.2 NBA Results

This section will present the results of applying the proposed methodology of data handling, events identification and correlation analysis to the NBA data sets. Results in last chapter provided interesting findings in the football data. Now it is necessary to test the methods on other types of data to see if the hypothesis of this project is supported by other kinds topic or just came by coincidence.

4.2.1 Introduction

The National Basketball Association (NBA) is the men's professional basketball league in America, and its surging popularity around the whole world is just as incredible as football. The regular match season always begins in the last week of October and ends around the middle of April in the following year. After the regular season, the NBA Play-offs begin in late April and run to June.

However in 2011, NBA exhibition game was called of because of the NBA lockout and the regular season was delayed to December 25th, regular season games were reduced from 82 to 66 games. The raw data for the NBA from Twitter and ProSportsDaily.com was collected from May 2011 to the middle of January 2012, so it partly covered the period of the 2011-12 NBA match season.

4.2.2 Results

NBA data behavior by month

The first step dealing with the NBA datasets is again to interpret and understand data behaviors via visualization. Tweets and posts were added together by day and plotted in Excel by month. The initial inspection revealed that the forum data was completed lost from May 23rd to July 4th, and that the data in October and January was also partly missing.

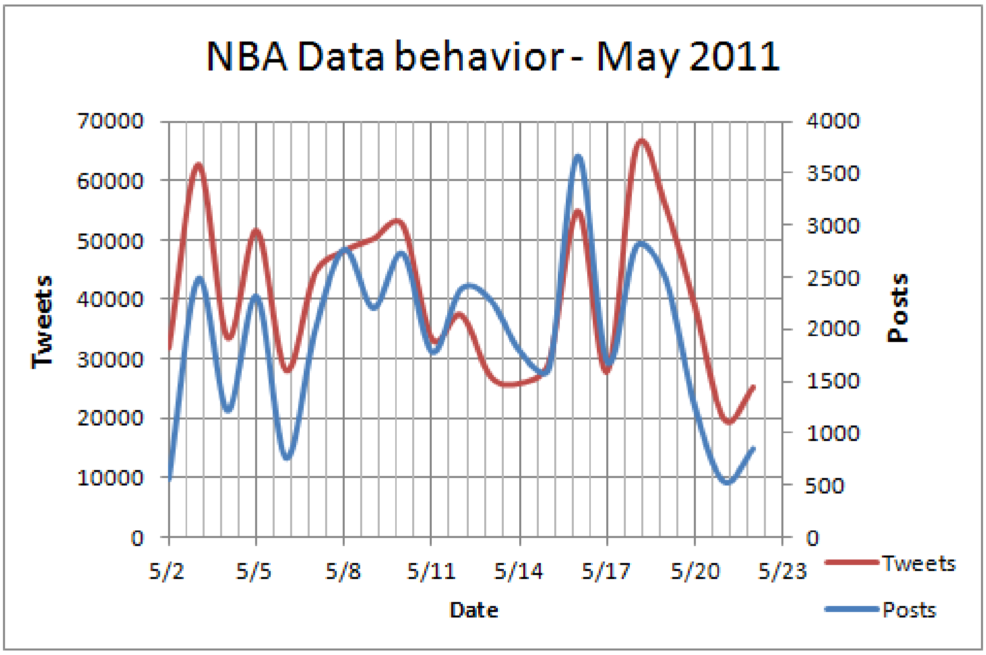


Figure 4.10: NBA data behavior in May 2011

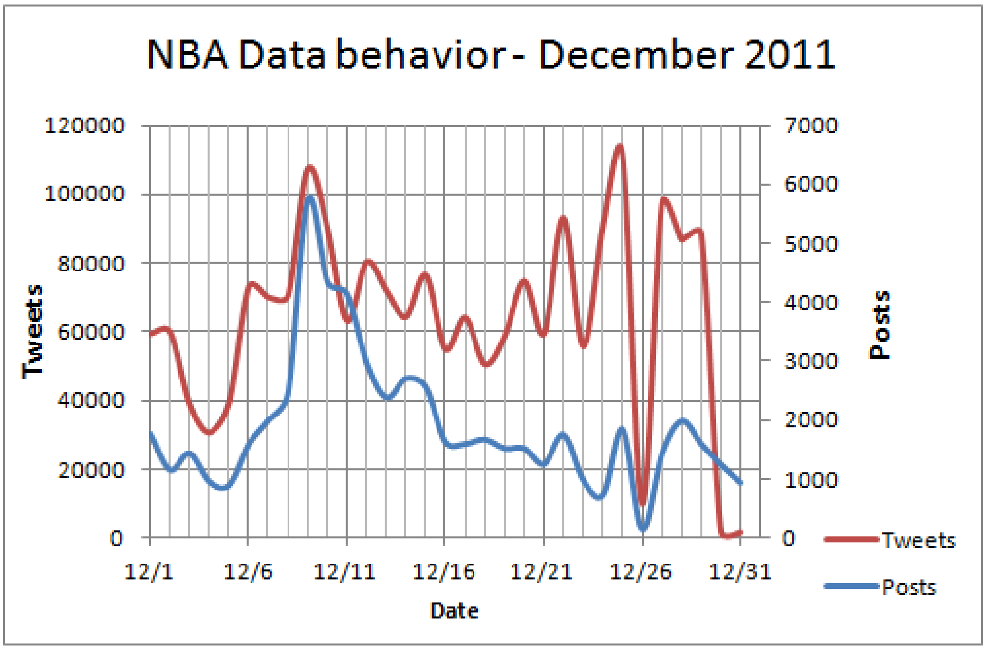


Figure 4.11: NBA data behavior in December 2011

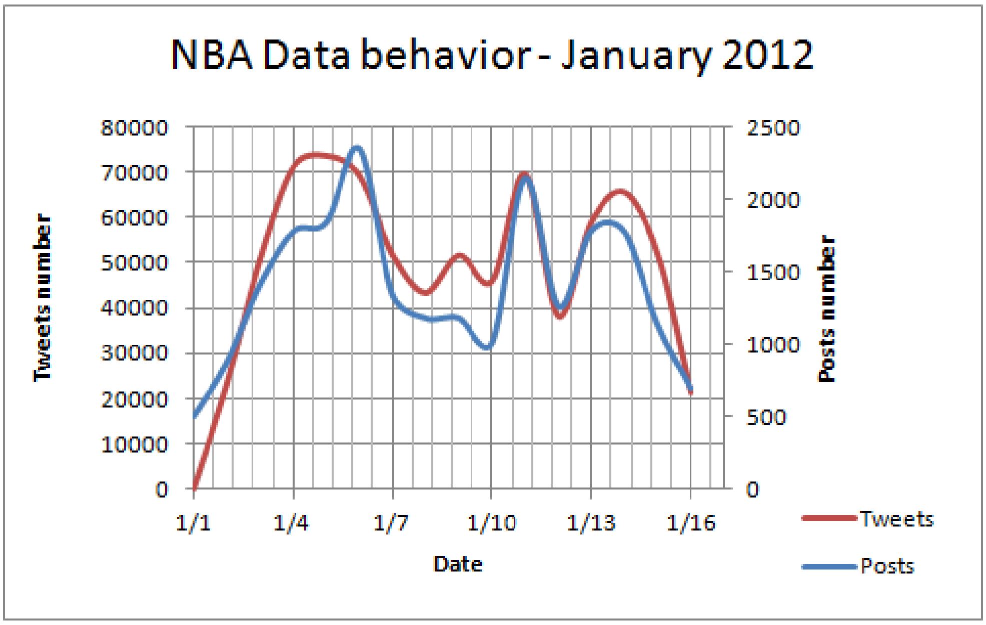


Figure 4.12: NBA data behavior in January 2012

Figures 4.10, 4.11 and 4.12illustrate how the Twitter and forum data behaved during the match season. These figures indicate that the variations of tweets and posts followed almost the same patterns along the timeline. The majority of the time, posts and tweets peaked on the same day. Besides, according to the graphs, even though all these month are within NBA game season, December seems to be the hottest month for fans on both Twitter and Prosportsdaily.con forum. The maximum number of tweets and posts were almost doubled compared to May 2011 and January 2012.

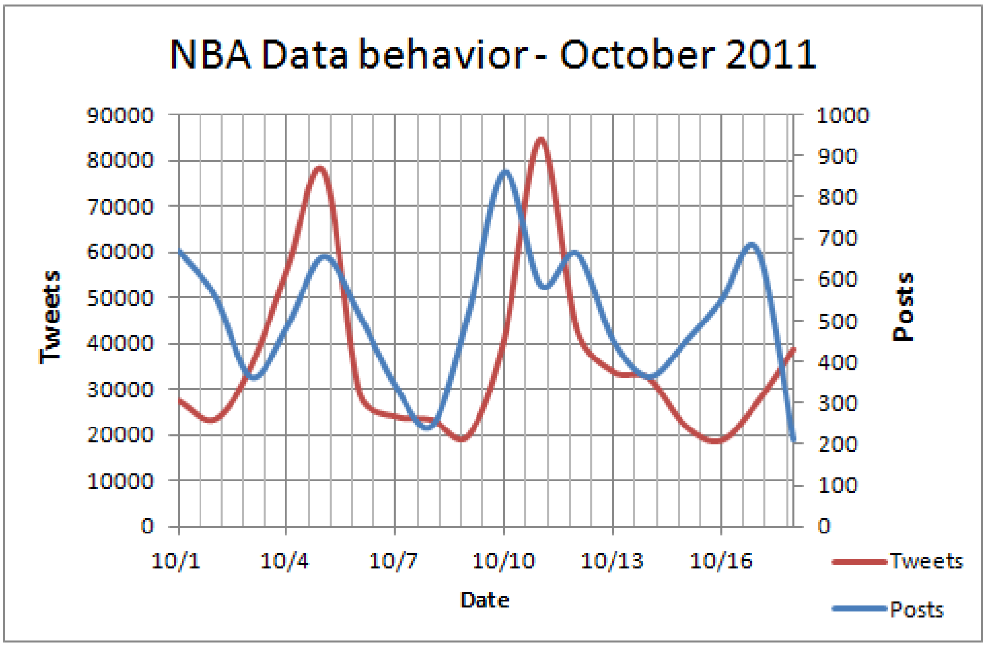


Figure 4.13: NBA data behavior in October 2011

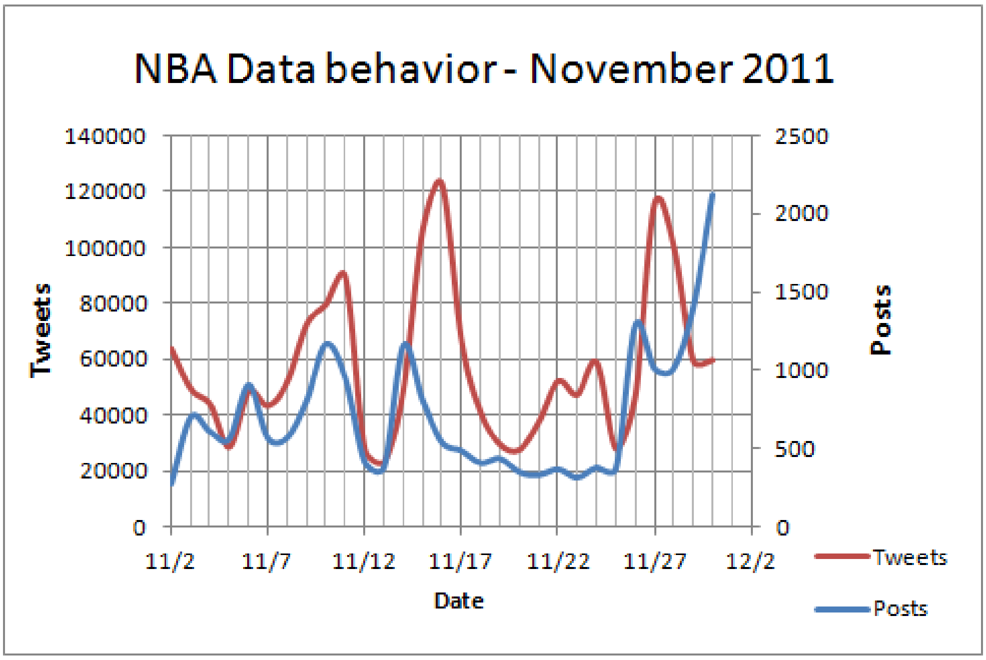


Figure 4.14: NBA data behavior in November 2011

October and November are two particular month which used to be NBA regular season but were not in 2011 because of the NBA lockout. Figures 4.13 and 4.14illustrate how the Twitter and forum data behaved during these two month. Even though no games were played then, it is surprising that the variations of tweets and posts still followed similar patterns. But it's difficult to conclude which one was prior in time. On some days posts reached their peak point earlier than tweets, and some days tweets rose up ahead of posts. It seems that the on-line discussion about NBA was not weaken by the lockout, and fans were still having passion posting NBA related news.

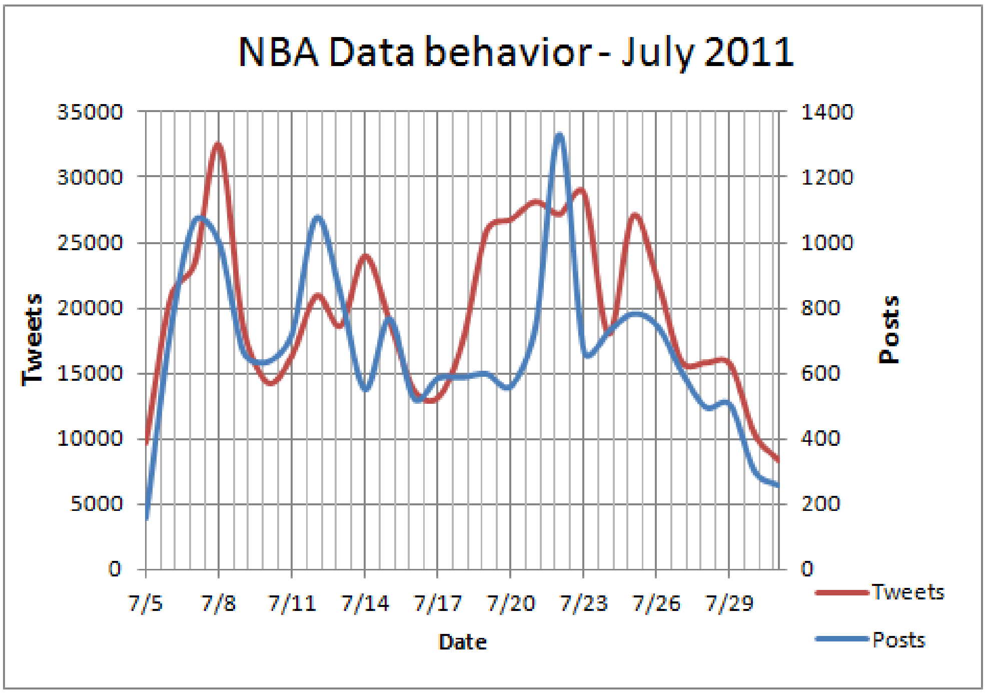


Figure 4.15: NBA data behavior in July 2011

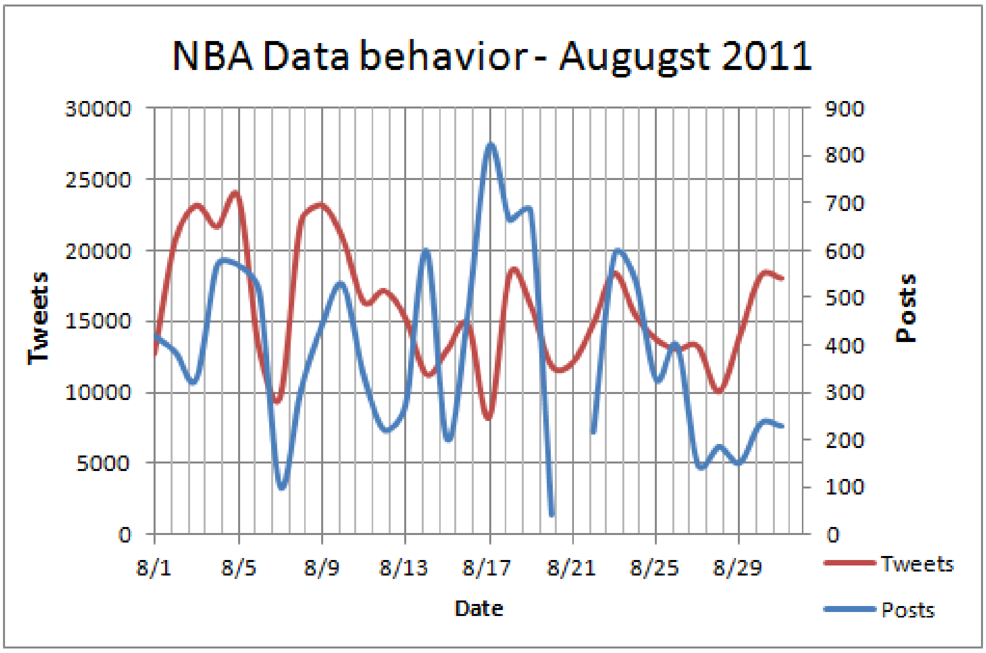


Figure 4.16: NBA data behavior in August 2011

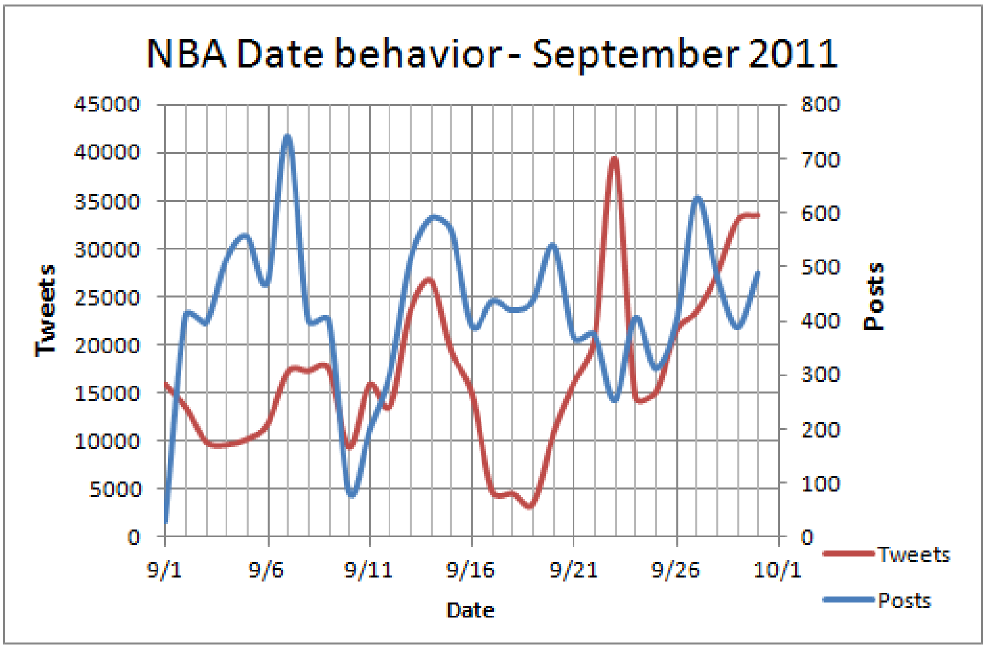


Figure 4.17: NBA data behavior in September 2011

Figures 4.15, 4.16 and 4.17 illustrate how the Twitter and forum data behaved during the absolute non-match season. The data increase and peak points were less correlated in tweets and posts during this period, August and September looked completely messy. In September when posts exhibited a small spike on 20th, tweets were at their lowest point at that time. In addition, Twitter data decreased since July and remained low during the whole of August, and then increased again in September. Forum data kept dropping starting in July but failed to go up in September.

NBA events and correlation analysis

In general, the NBA data records little traffic in the forum. The worst periods were in August and September with 383 and 416 daily average posts, respectively, which was exactly the period of the lowest activity for the NBA. The highest values occur in May and December, with 1900 posts in one day.

After applying the data handling and events identification process on the original data sets, there were 305 Twitter events and 311 forum events in total. Table 4.3 below shows the distribution of NBA Twitter and forum events throughout the day.

Time	Twitter E	Forum E	Pairs	Match%
00-03	64	72	32	50.0%
03-06	94	79	36	38.3%
06-09	31	26	14	45.2%
09-12	3	2	1	33.3
12-15	2	0	0	0
15-18	22	20	4	18.2%
18-21	53	51	14	26.4%
21-24	36	61	23	63.9%

Table 4.3: NBA events distribution during the periods of one day

Generally NBA games starts between 19:00 to 21:00 in American eastern standard time, and each match runs about 90 minutes or longer. Converting to Norwegian time, it means 00:00-06:00 is the period during which there might be the most tweets and posts activity. Table 4.3 illustrates the impact from NBA games on tweets and posts. 50% of Twitter and forum events happened between 00:00 to 06:00; other busy periods are from 18:00 to 24:00, where 30% events were located.

Among 305 Twitter events, 191 (62.6%) events are able to be matched with related forum events within 24 hours, and for 124 (40.6%) Twitter events, the forum events occurred within -1 to 6 hours by using the earliest occurring forum events criteria. This means more than half of the Twitter events failed to find a corresponding forum event under 6 hours' time constrain. Table 4.3

also shows the matching percentage during different time periods in a day. This time matching seems unbalanced. 21:00-03:00 gives the best results, and there is little or no event matchups between 09:00-15:00. It seems NBA match time doesn't have much to do with the events correlation either as NFL.

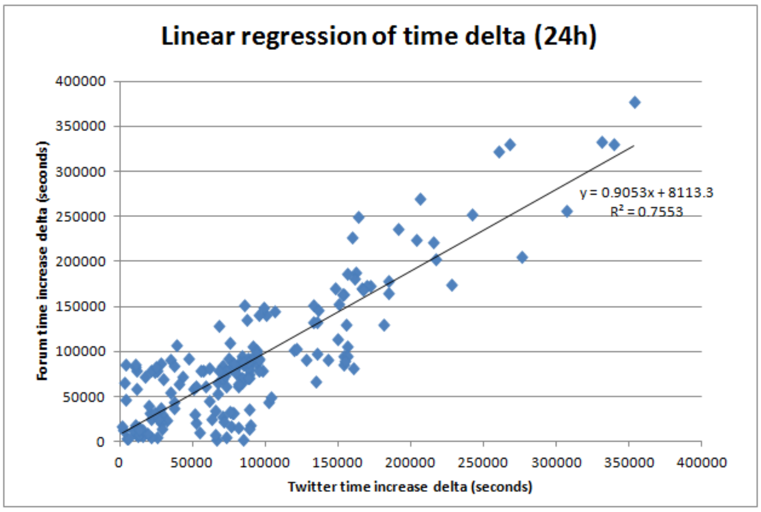


Figure 4.18: Linear regression of NBA events time delta

Figure 4.18 illustrates the investigation of a proportional relationship between the forum event timestamp increase ΔY and the Twitter event timestamp increase ΔX. Data points are discretely distributed around the trend line, and this linear function can only cover 75.5% of all data, with a large intercept of 8113.3. This result includes all possible correlated forum events within 24 hours, and it gives a sense of how forum increases follow tweets spikes. Compared to NFL data, time differences within the NBA events pairs here are larger, with the mean equal to 6.27 hours and the standard deviation equals to 7.45 hours.

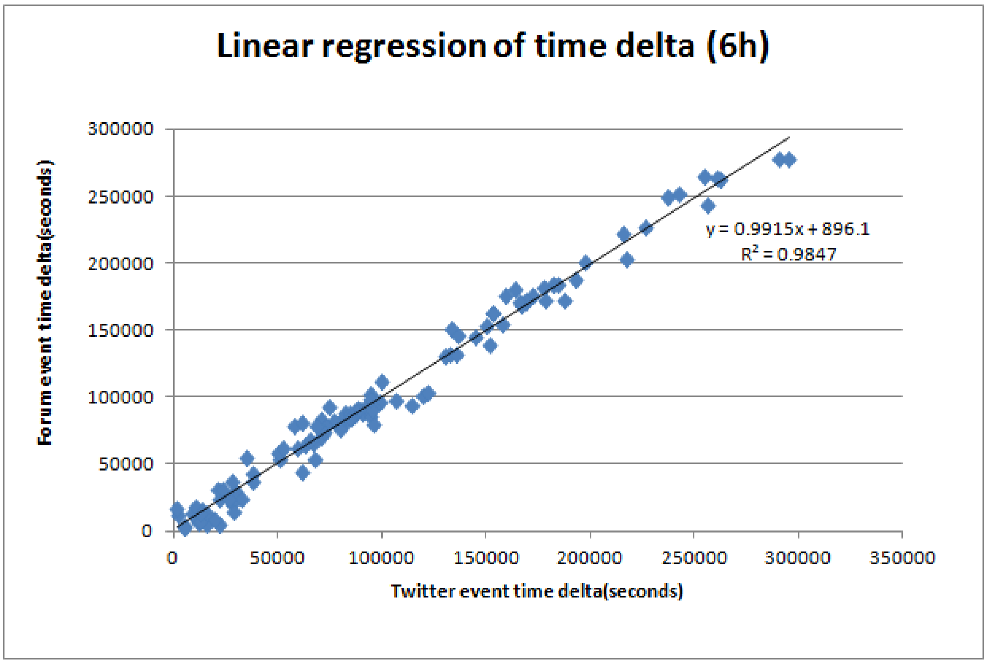


Figure 4.19: Linear regression of NBA events time delta (first following forum event)

Figure 4.19 shows the linear regression result of time delta when the earliest forum events are kept and the time scope between two related events is reduced to 6 hours. The results indicates a near-perfect linear relationship here with the function ΔY = 0.9915*ΔX + 896.1. R^2 is equal to 0.9847, meaning that the trend line covers 98.4% data. With an acceptable intercept of 896.1, ΔY and ΔX can be regarded proportional related. Therefore a linear correlation ″Y = a*X + b″ where X represents Twitter event timestamp and Y represents forum events timestamp is proved existing in NBA events pair.

Again assuming a value of a of 0.9915 is sufficient to conclude that ΔX and ΔY are 1:1 proportional related, then b is the time difference between two successive events. NBA events have 1.19 hours as the average duration it takes for a forum events to occur after a Twitter event, where the standard deviation is 1.79 hours. 95% of correlated forum events appear within the 4.77 hours′ time scope.

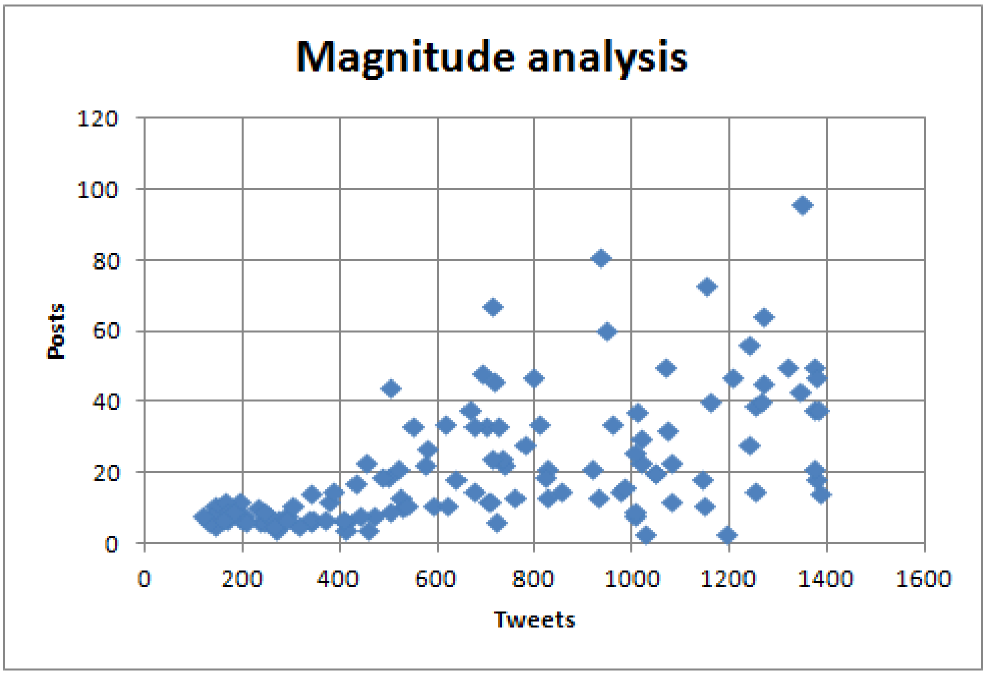


Figure 4.20: Magnitude analysis between tweets and posts (first following forum event)

Figure 4.20 shows how posts and tweets amounts are correlated. It is clear that the data is just as messy as the NFL results. Posts numbers remain pretty low although tweets exceed 1000. No obvious relationship is able to be found between tweets and posts increases.

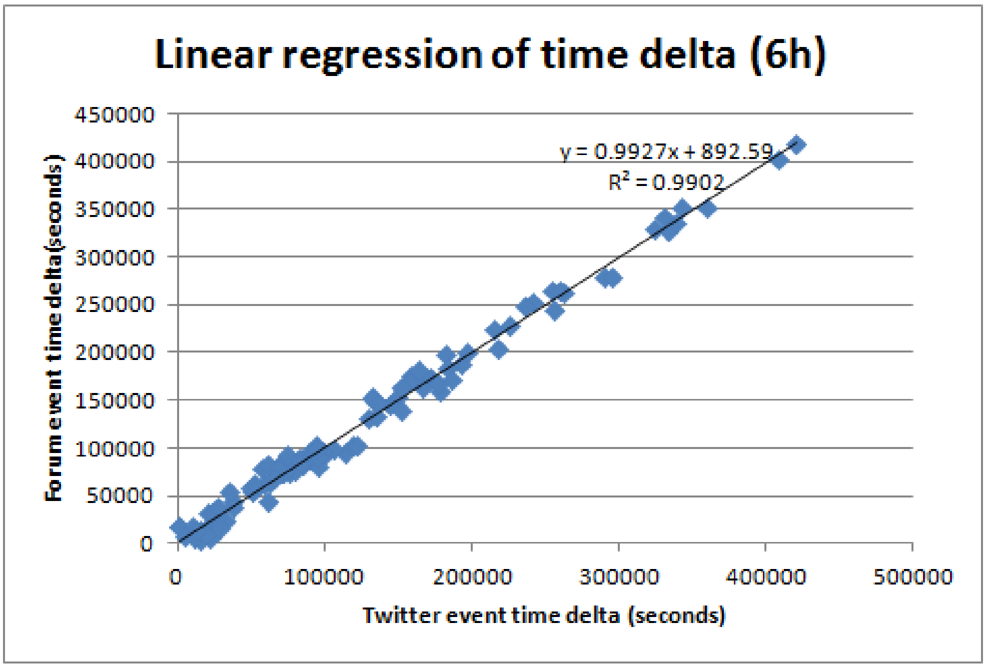


Figure 4.21: Linear regression of NBA events time delta (maximum value forum event)

If the largest forum events are kept, and the time scope between two related events is reduced to 6 hours, Figure 4.21 shows the linear regression result of the time delta when giving matching priority to posts amount. The result is not very different from Figure 4.19. The linear regression here gives the function ΔY = 0.9927*ΔX + 892.59, and this trend line covers 99% data, with an intercept of 892.59.

The time increases between Twitter events and forum events can be viewed as proportional. The average time difference between Twitter and forum events is 1.48 hours with a standard deviation of 1.88 hours. 95% of the correlated forum events can be found within 5.25 hours after Twitter events under this scenario with the NBA data.

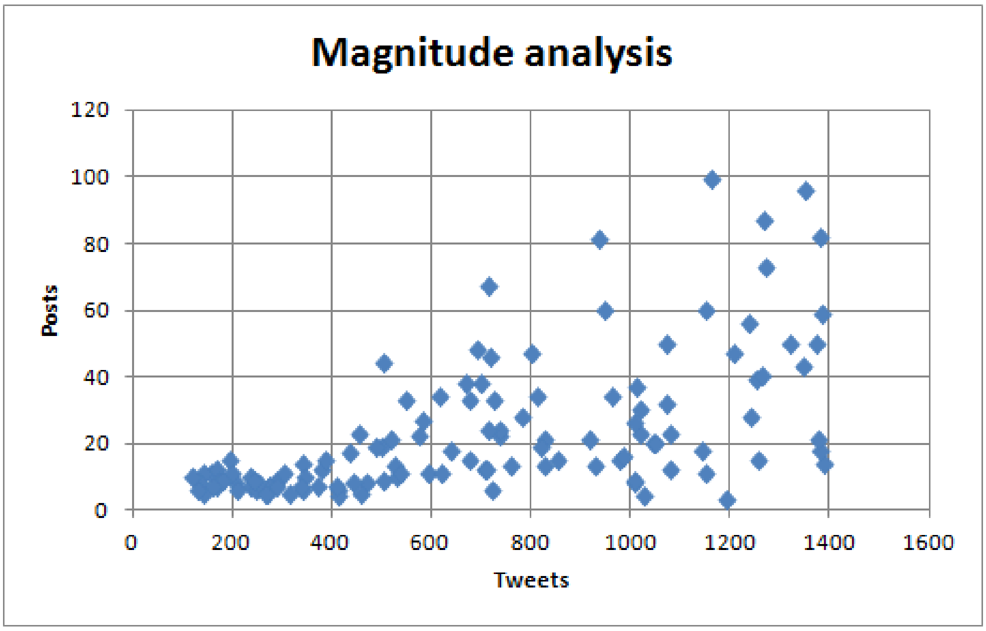


Figure 4.22: Magnitude analysis between tweets and posts (largest forum event)

The magnitude correlation between tweets and posts might be expected to be better in this situation, since the maximum posts were picked out when the tweets peaked. However, Figure 4.22 illustrates again that no obvious relationship is able to be found between tweets and posts increases.

4.2.3 NBA Results Analysis

The NBA has same features with as the NFL in terms of match season structure: they both have an exhibition season, a regular season and playoffs during a competition year, and both the NBA and NFL seasons lasts for 8 months. According to this study, the NBA data gives similar results to the NFL data. Linear regression fits with a high sample correlation R, and most correlated forum events occurred within 5 hours after Twitter events.

Compared to the NFL results, the NBA results shows less correlations in tweets and posts in the earlier months of data collection since a large off-season period is included within the data in 2011. Monthly tweets and posts variations exhibited poor correlation in July, August and September. Only 124(40.6%)Twitter spikes are followed by corresponding forum increases. The negative effect from of the non-match month was somewhat verified within NBA and will be discussed later.

The linear relationship in the NBA data appears looser than the NFL data.

Figure 4.18 and Figure 4.19 show the trend line performance with 24 hours' and 6 hours' constraints. Although the sample correlation R is more or less the same compared to the NFL, the intercepts are lager in NBA and this suggests the proportional relationship between ΔY and ΔX is stronger in NFL events. In addition, speaking for the predictability of Twitter events, only 97 tweets events were able to suggest an upward trend in forum posts, that's 31.8% of all Twitter events.

Analysis of the NBA data is accomplished in this section, and the results and analysis shows the successful process with the data mining approach. The NBA data exhibits correlation between tweets and posts, however the predictability of Twitter events remains unknown.

4.3 MMA Specific Results

This section will present the results and analysis of testing and running the data mining approach on MMA data.

4.3.1 Introduction

MMA (Mixed Martial Arts) is a full contact combat sport between two individuals and allows the participants to use both striking and grappling techniques, both while standing and on the ground. Mixed martial arts includes a variety of fighting and martial artists styles, including boxing, wrestling, muay Thai, Taekwondo, Karate and others. MMA doesn't have a regular season format to determine a champion like the NFL and the NBA. Matches start from the beginning of the year and runs to the end. The most basic rule of MMA is the fighter must fight within his weight class. In addition,there is no particular elimination rules as in other professional sports.

The MMA Twitter data was collected from April 6th 2011 to Jan 16th 2012, but forum data from Sherdog.net forum only covers April 8th, 9th, and 10th and from July 5th to September 26th, with unexpected data loss. Since MMA is a non-season sport, data variation and patterns should not be affected by different periods of the year, and the existing data should be able to provide general trends in MMA tweets and posts.

4.3.2 Results

MMA data behavior by month

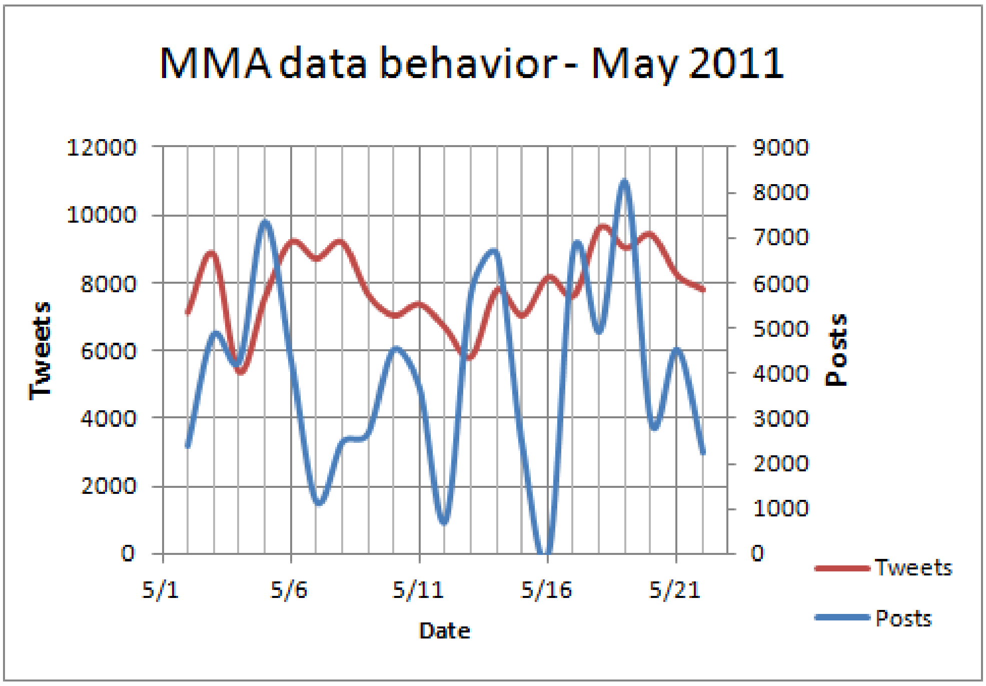


Figure 4.23: MMA tweets and posts behavior in May 2011

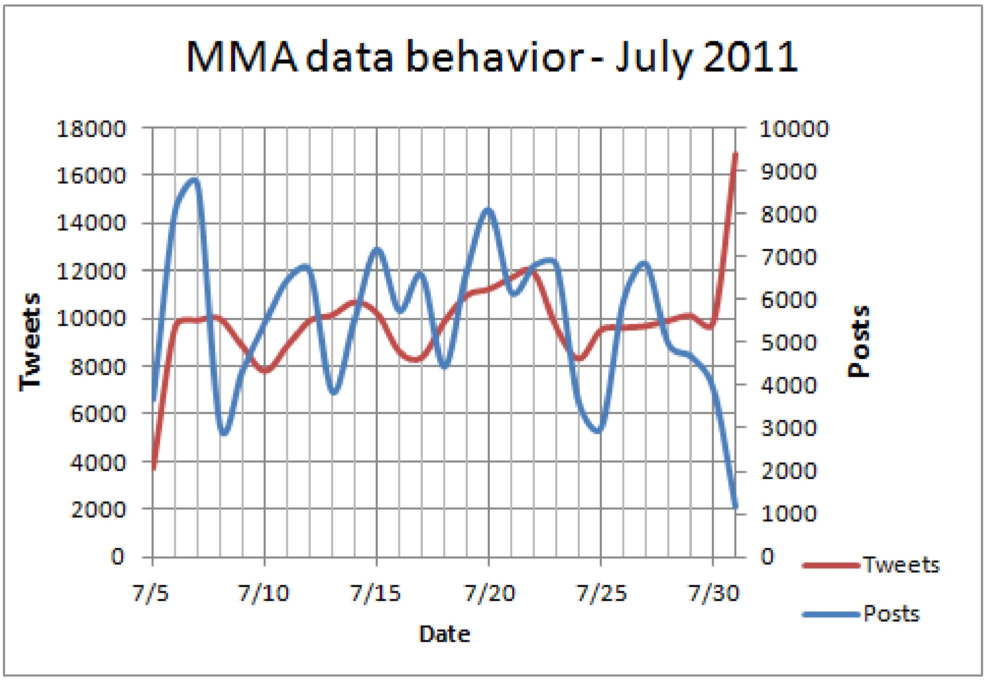


Figure 4.24: MMA tweets and posts behavior in July 2011

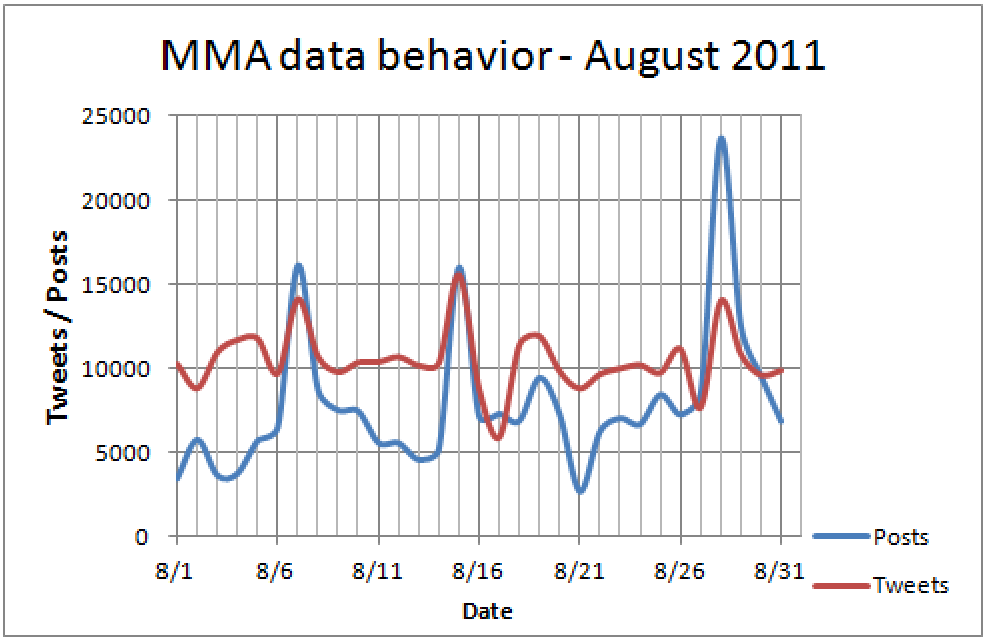


Figure 4.25: MMA tweets and posts behavior in August 2011

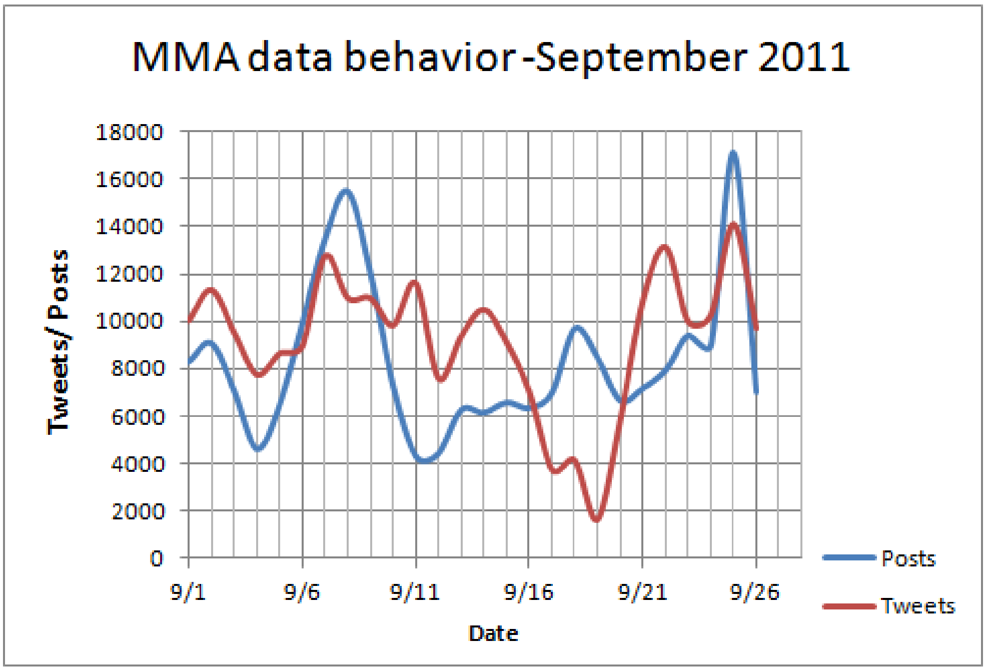


Figure 4.26: MMA tweets and posts behavior in September 2011

This first step is always to sum the tweets and posts amounts by day and plot them by month, to provide a first sense of how the data behaves and operates. Figures 4.23, 4.24, 4.25, 4.26 illustrate the tweets and posts variations in different months. The first thing to notice here is that posts amounts are much larger than for the previous data sets for the NFL and the NBA. The average daily posts values were over 8300 in September, and even in the least active month of May had almost 4000 posts every day. It seems Sherdog.net forum is able to provide sufficient posts for further analysis.

Since there is no regular match season in MMA, the data is expected to behave similarly in different months. In addition, month patterns of data values rising and falling are also expected from previous experience with sports-related tweets and posts. However, the figures above shows completely different results, and no trends can be identified in either tweets nor posts. Their varying pattern differs from month to month. The Twitter data flow up and down without large jumps while the posts data appears more volatile. Unlike the NFL and NBA data, MMA shows little obvious correlation between tweets and posts. August is the only month exhibiting corresponding increases and spikes in both data sets; in other months, the tweets and posts variations just appear unmatched.

Although there is no particularly ”hot” month in MMA fights, a stable increase can be found in the posts numbers: its daily average rises from 3931 in May to 8377 in September. Twitter data goes up from a daily average of 7867 in May, peaking at 10445 in August, and then decreases slightly to 9213 in September. These daily sum and monthly variation results shows the unstable and unpredictable trends in tweets and posts of MMA.

Events and correlation analysis

Although Sherdog.det forum provides larger posts number which indicate varying intensity of online discussion activities, the data preparation and cleaning work is extremely unpleasant for the MMA data. Data was partly missing on almost every day, with hours long gap durations. Most days in May cannot illustrate any expected variation since the data was interrupted too much. Moreover, even the existing data behaves erratically very often, with data jumping around within an unexpected range and rarely showing a stable increase.

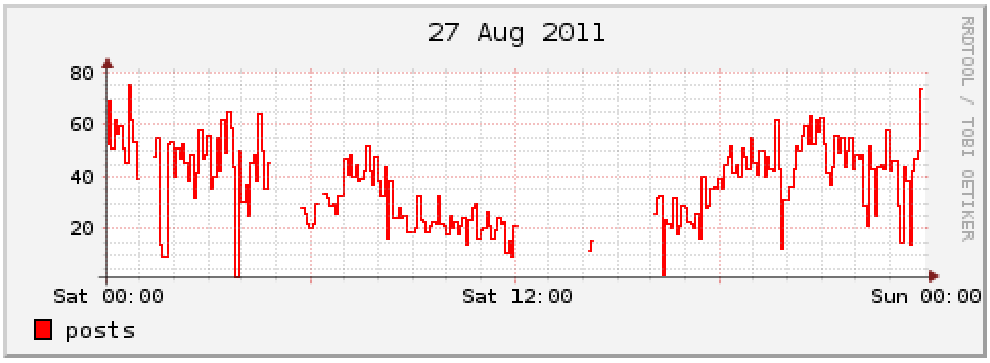


Figure 4.27: Unexpected data jump in forum posts

Figure 4.27 illustrates the general situation appearing in MMA forum data collecting. Posts numbers fell suddenly from 44 to 1 at 03:55 and then went back to 51 at 04:00. Reasons for this kind of sudden change in the data are unclear, and this data may very well be suspect. 15 minute averages method do not help reduce this error very much as the data continually oscillates. In addition, averaging values would also erase some large data points since the increases are not strong enough. Therefore, identifying events in MMA forum data is quite difficult, and the results are likely to contain more faults.

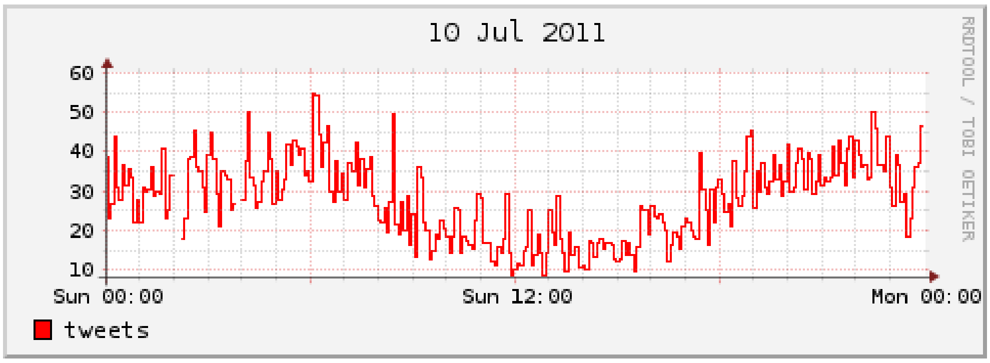


Figure 4.28: Spiky MMA data in tweets

MMA Twitter events detection is also more difficult than previously, due to the lack of strong tweets increases of significant duration. Figure 4.28 shows how the tweets numbers oscillate without any stable surges. In this graph, none of the small spikes are identified as events since the data is just flowing up and down around a certain level. A manual check was done on both Twitter and forum events results before correlation analysis, to make sure the final results are reliable. In the end, there are 153 Twitter events and 81 forum events in total.

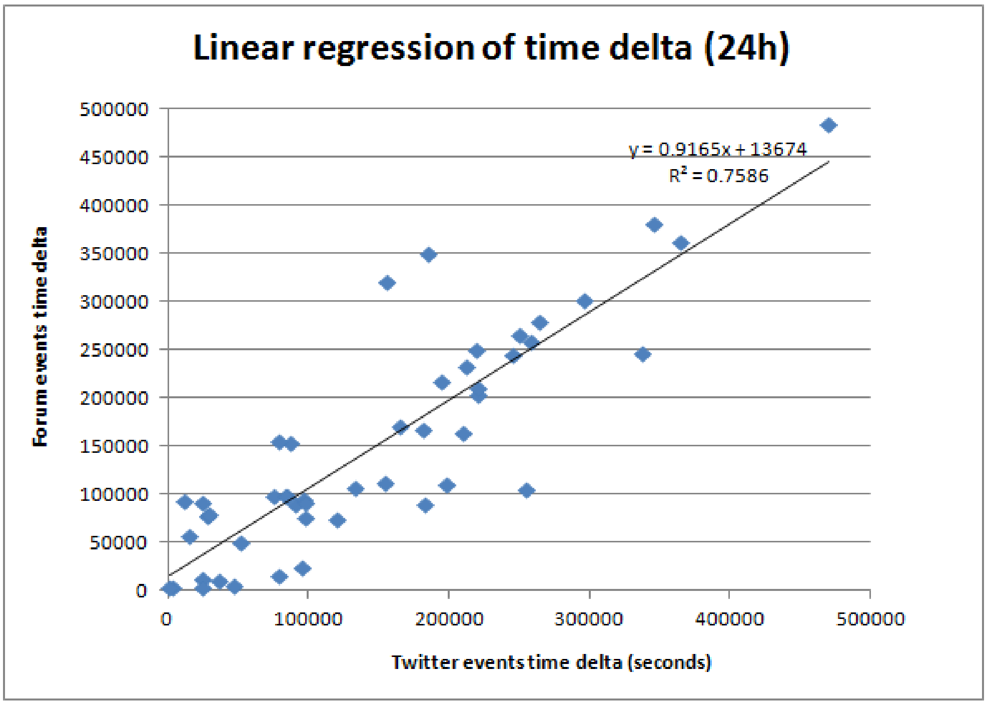


Figure 4.29: Linear regression of time delta within 24 hours

Figure 4.29 illustrates how events were correlated within 24 hours. Results gives only 61 Twitter events which are followed by forum events, that is to say 39.8% tweets spikes are related to posts variations. Even if the linear regression succeeds in this situation with a good sample correlation R, the confidence in the MMA results will be low.

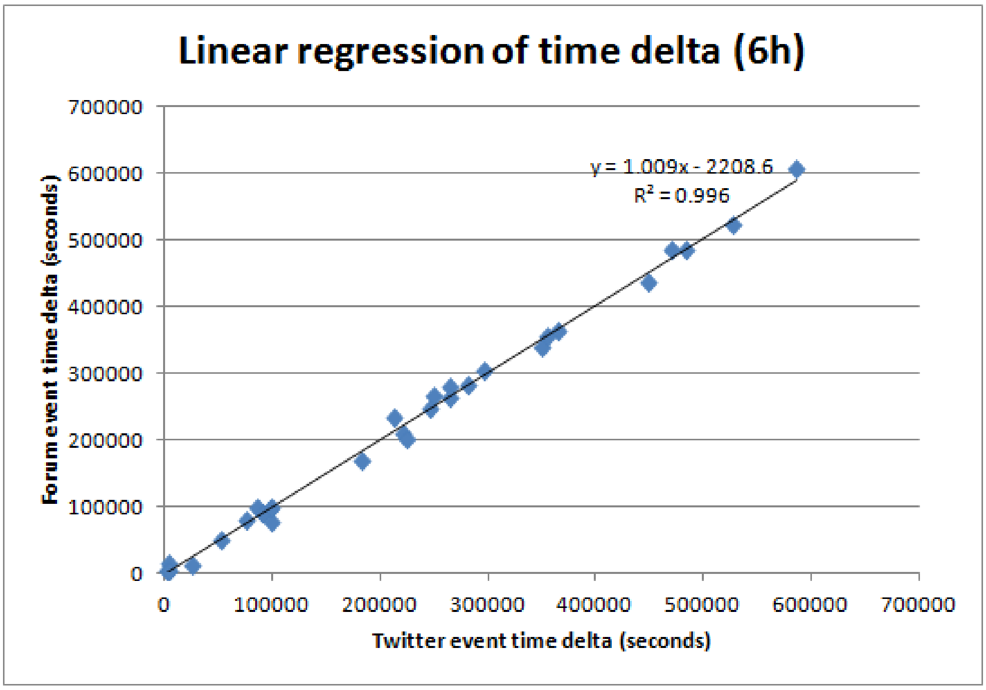


Figure 4.30: Linear regression of time delta within 6 hours

Figure 4.30 shows time delta correlation when time constraint is reduced to 6 hours and the earliest forum events are kept. Only 22.9% pairs exist among all the reported Twitter and forum events. Even though the event time increase ΔX and ΔY between Twitter and forum events are proved proportional, the correlation between tweets and posts performs incredible poorly.

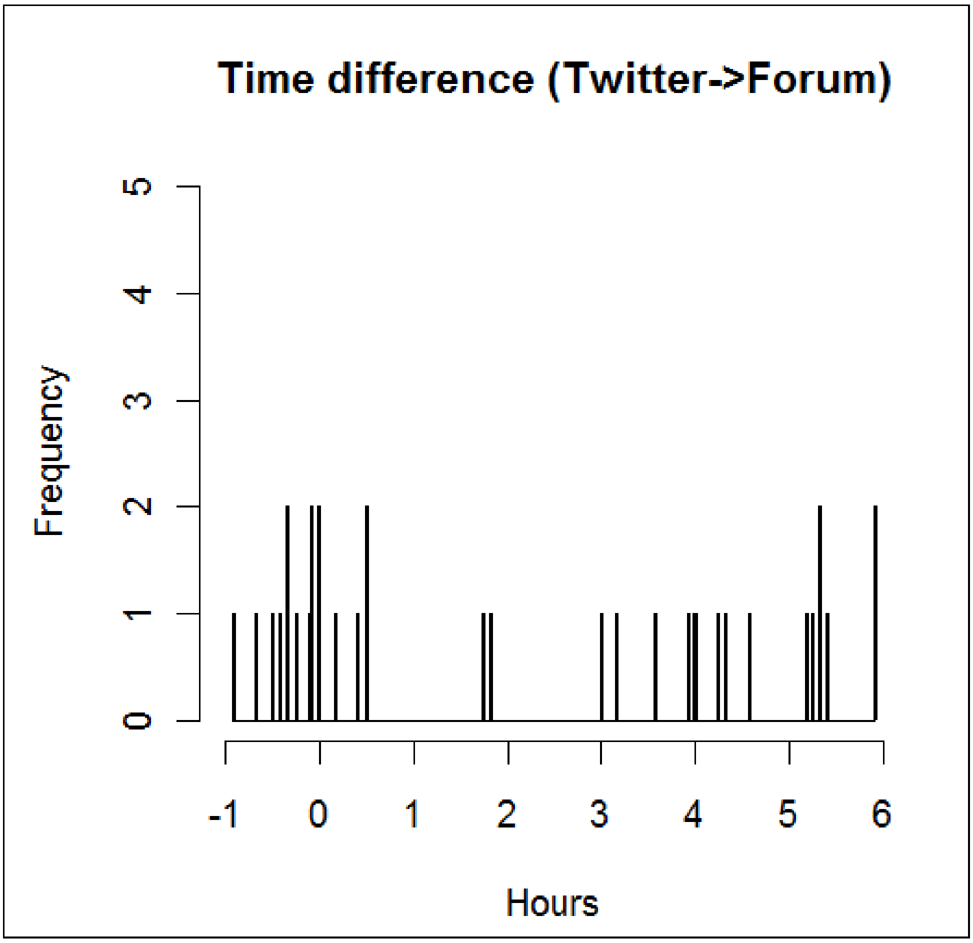


Figure 4.31: Time difference between between MMA Twitter events and forum events within 6 hours

The time differences between Twitter events and forum events are not normal distributed in this case. Figure 4.31 shows the random distribution of time duration between a forum event and the corresponding Twitter event. Mean and standard deviation results are very large, at 2.13 hours and 2.41 hours respectively. Taken all into consideration, the correlation between tweets and posts is believed to be weak, so no further analysis needs to be done.

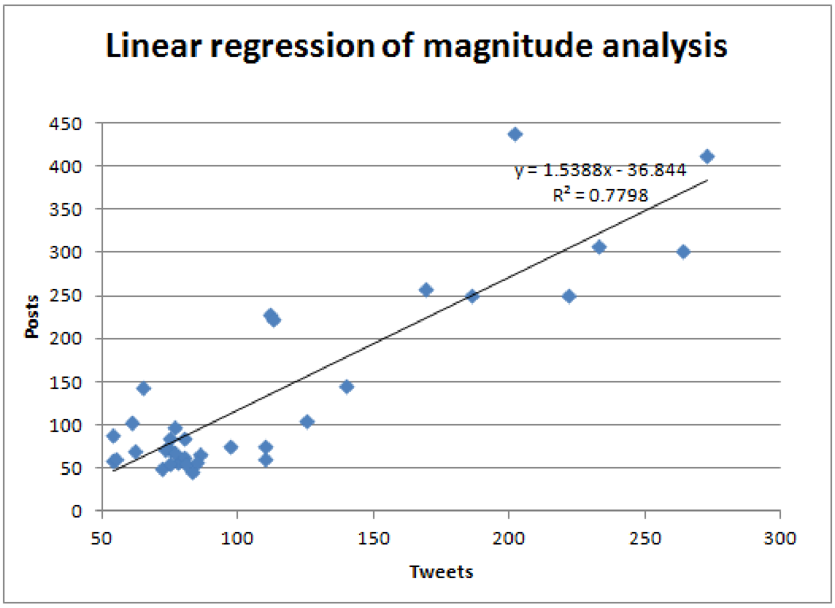


Figure 4.32: Traffic magnitude analysis in MMA with the first following forum events

Interestingly, in the MMA data, Twitter events and forum events failed to shows sufficient connection in time scale. However, correlation was found between the tweets and posts magnitudes. Figure 4.32 demonstrates the linear relationship in tweets and posts amounts for events pairs within 6 hours of one another. It seems that large tweets amounts correspond to large posts values in MMA, and the linear function Y = 1.5388*X - 36.844 is able to cover 77.98% of the posts results. Therefore the MMA data is capable of showing the scale of Twitter events and forum events are correlated, and tweets and posts amounts can be computed out from one set of values.

4.3.3 MMA Results Analysis

This section shows the results of testing and running the methodology on MMA data. Martial arts matches and competitions are totally different from American football and basketball, and the martial arts fights consists of variety types and can start any time of a year. There is no regular match season to be explored in MMA. Viewing from the perspective of Twitter, human attention to martial arts is relatively low compared to major sports like football and basketball. Even though Sherdog.net forums provide sufficient original posts data, the lack of stable large increases in tweets and posts traffic makes the event identification work difficult for the MMA data.

Among all 135 Twitter events, only 35 events that can be matched in pairs, which means only 22.9% Twitter events have corresponding forum events within

6 hours. Although linear regression fits well in all events pairs, this result can conclude nothing but a very weak connection between Twitter and Sherdog.net forum, and certainly not any predictability relationship. However, it is worth mentioning that, in the MMA data, the magnitudes of tweets and posts are correlated with a linear relationship, in contrast to the NFL and NBA data where no such relationship exists. This suggests that considering a possible connection in Twitter and forum event scales is a good topic for future study.

The analysis of the MMA data is accomplished in this section, and the results and analysis shows how mining process has problems with the sort of data present in these data sets. The MMA data shows very little correlation between Twitter and the Sherdog.net forum, and the predictability of Twitter events appears low possibility in this case.

Chapter 5

Discussion

The goal of this project is to investigate the predictability of Twitter traffic for resource consumption of topic-related websites. This project proposes a new methodology for data mining for traffic correlation between Twitter events and corresponding forum postings. Experiments are conducted by examining data variation on pre-collected data sets focusing on sports topics such as NFL, NBA and MMA. Key results are presented in Table 5.1 below.

This chapter will discuss the interesting findings observed in the correlations in the project, and suggestions about future design and research in this field will also be put forward.

Theme	Correlation%	Timespan Avg(hours)	Timespan StDev(hours)
NFL	50.4%	1.69	1.51
NBA	40.6%	1.19	1.79
MMA	22.9%	2.13	2.41

Table 5.1: Key results of tweets/posts correlation analysis

5.1 The predictive ability of Twitter for corresponding forums

Correlation is analyzed by studying the traffic variations on Twitter and the relevant forums. According to the hypothesis at the beginning of this project, for each chosen topic, tweets surges are giving advance warning of subsequent posts surges, ideally within a couple of hours. However, as a matter of fact, the relationship between tweets and corresponding posts is not as anticipated. The following sections discuss key points from investigating several of the various data sets.

Correlation is not causality

The major findings from the graphical and statistical results indicate correlation between Twitter and the chosen corresponding website, but this cannot be interpreted to imply causality. Neither 100% of the Twitter traffic surges are followed by forum posting increases, nor are 100% of the forum events able to be paired with tweets spikes. Both tweets and posts events can occur alone, without connections to the other.

Mining results for the NFL demonstrate that 50.4% Twitter events have time-related forum events with a 6 hours time scope constrain. However, NBA results show looser correlations between tweets and posts with the same constraints; only 40.6% of the tweets spikes are able to be paired with time-related forum increases. Finally, the MMA data gives the worst results. Here, only 22.9% of Twitter events have corresponding forum events ,which suggests that any correlation is too weak to be confirmed in MMA tweets and posts.

Timespans between related events are large

Statistic analysis results of the timespans between events indicates that the duration between a Twitter event and its correlated forum event are larger than expected, although linear regression fits perfectly in analyzing time correlation and gives linear functions which are capable of accounting for 99% of the events pairs under all circumstances for the NFL, NBA and MMA data. Setting the MMA events aside, since the correlation appears so weak, the average timespans is around 1.40 hours with approximately 1.65 hours' standard deviation for both the NFL and NBA. Long durations between tweets and posts events unfortunately lack the reliability needed to confirm strong connections between Twitter and forums traffic.

Not all Twitter events are predictive

According to correlation results, not 100% Twitter events happened ahead of forum events. Forum posting increases can appear in front of or almost at the same time as tweets spikes. The NBA graphical results in May illustrate both situations.

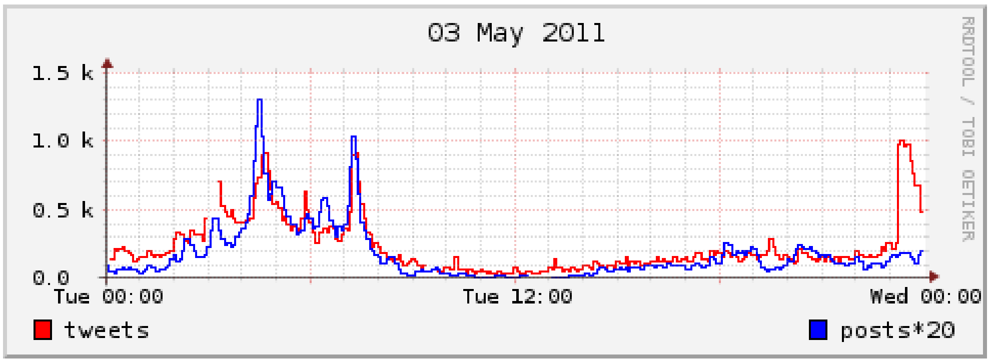


Figure 5.1: Sudden increase in posts at the same time as tweets surge

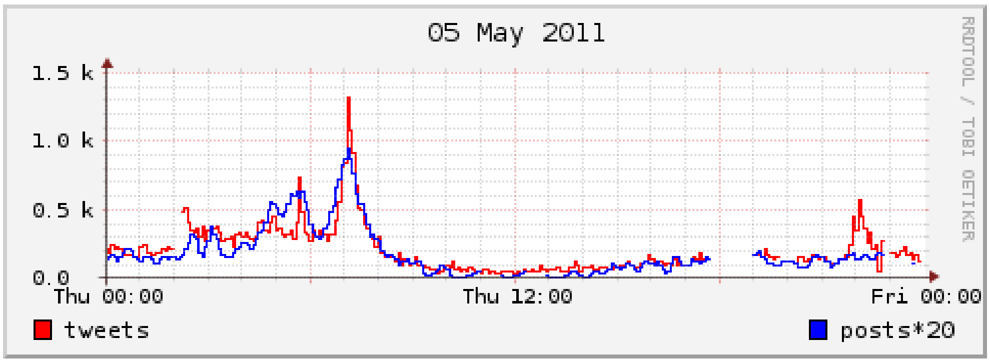


Figure 5.2: Sudden increase in posts at the same time as tweets surge

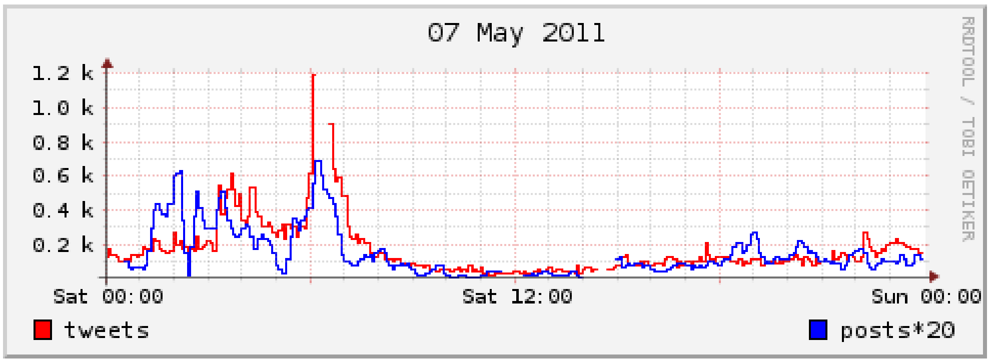


Figure 5.3: Sudden increase in posts ahead of tweets surge

Figures 5.1, 5.2, and 5.3 show how Twitter traffic is correlated with forum posts increase. Posts amounts are magnified 20 times to be better compared with tweets in scale. In Figures 5.1 and 5.2, it is clear that tweets and posts go up and peak without any time difference; both Twitter and forum events occur at the same time point. Figure 5.3 demonstrates the other situation; it is easy to see here that posts increase and reach the highest points before tweets spikes occur. Therefore, not all Twitter traffic are predictive to forum events.

Statistic calculations provide an overview of the predictability of Twitter events. After matching up all possible events pairs from Twitter and the corresponding forums, NFL results shows 88% Twitter events have the predictability to suggest a forum events within 6 hours, while NBA results have 78% of Twitter events as predictive.

Data variation matters the most to correlation analysis

If the answer to ″Can Twitter predict associate resource coming surges?″ remains ″MAYBE″ with NFL and NBA data, the response given by the MMA results is definitely ″NO.″ Even though the Sherdog.net forums provides more sufficient data for indicating posts variations, oscillating data and the lack of stable increases results in the MMA data yielding the worst mining results for the events correlation analysis. Only 22.9% Twitter events can be paired with time-related forum events within 6 hours, and the correlation percentage is too low to even mention how many of them are predictive.

ProSportDaily.com forums failed to provide sufficient posts events because they focus on various professional sports topics. The NFL and NBA lacks special significance within this arena. Sherdog.net is one of the most popular forums which focuses on martial arts affairs. Thus, it provides larger posts amount compared to the ProSportDaily.com forums. However, the most sufficient posts data turns out to give the worst events correlation. Even if the original data collection work had problems and contained many data gaps, it is the MMA data's natural behaviors that is the real reason for its terrible results. Stable and distinct variations rarely show up inside the MMA data sets, causing difficulty for event identification, with only 81 forum events reported in the end.

The unexpected results from the MMA data suggest the most important factor supporting the hypothesis of this project. Rather than needing a sufficient amount of traffic, stable variation and strongly increasing trends matter the most to the events correlation analysis. Only when Twitter and the chosen websites are able to show their traffic surges clearly will any correlation between them be observable and meaningful.

Taking all of these results into consideration, the predictability of Twitter traffic for resource consumption on topic-related websites remains unclear. Correlations between tweets surge and posts increases are found. However, with the current data and analysis results, the answer to ″Can Twitter predict corresponding resource coming surges″ is still ″MAYBE.″ More investigation needs to be done, and the predictive ability of Twitter cannot be affirmed or denied as yet.

5.2 The possibility of exploring traffic correlations by indirect measures of public interest

The idea of this project comes from taking advantage of indirect indications of significant human. The assumption is that real life events are likely to turn into traffic bursts on related websites. As one of the most popular social networking sites, Twitter is believed to have the capability of reflecting the world's latest news in a timely way via tweet rates. Forums generally focus on specific trends, and initiate online discussions among a group of people with similar interests. These two type of websites have tight connections with current world and are able to show popularity surges for particular events by users' postings.

However, there are two aspects which need to be concerned seriously to in the context of this project:

- Whether the traffic correlation can be impacted by public emotion and passion.
- The reliability of exploring traffic correlation by simply considering the common discussion topic.

5.2.1 Positive correlation impact from the match season

Previous analysis of the NFL and NBA data demonstrate that Twitter and ProSportsDaily.com forums have the potential to reveal variation of public enthusiasm during different periods in a year. Since both football and basketball have competition seasons and millions fans and followers, the match season can be expected to most easily cause sports fans to express their burning enthusiasm, represented as increasing tweets and posts during those periods. Figures 5.4 and 5.5 demonstrate the influences on online activities in Twitter and the ProSportsDaily.com forums from NFL and NBA games in real world.

NFL games run from early August 2011 until early February 2012. Figure 5.4 shows that the tweets and posts events have an increasing correlation from August to December. However, the correlation percentage goes down from January to February in 2012, which is exactly the play-offs and championship season. So the online activities are not fully controlled by the real world in this case.

The regular NBA match season in 2011 began on December 25th and ran to April 26th, 2012. Since it was shortened by the player's strike, the preseason games were called off. Figure 5.5 illustrates the connection between Twitter and the related forums. A correlation increase can be seen from September 2011 to January 2012. Apparently, NBA games have the expected impact on

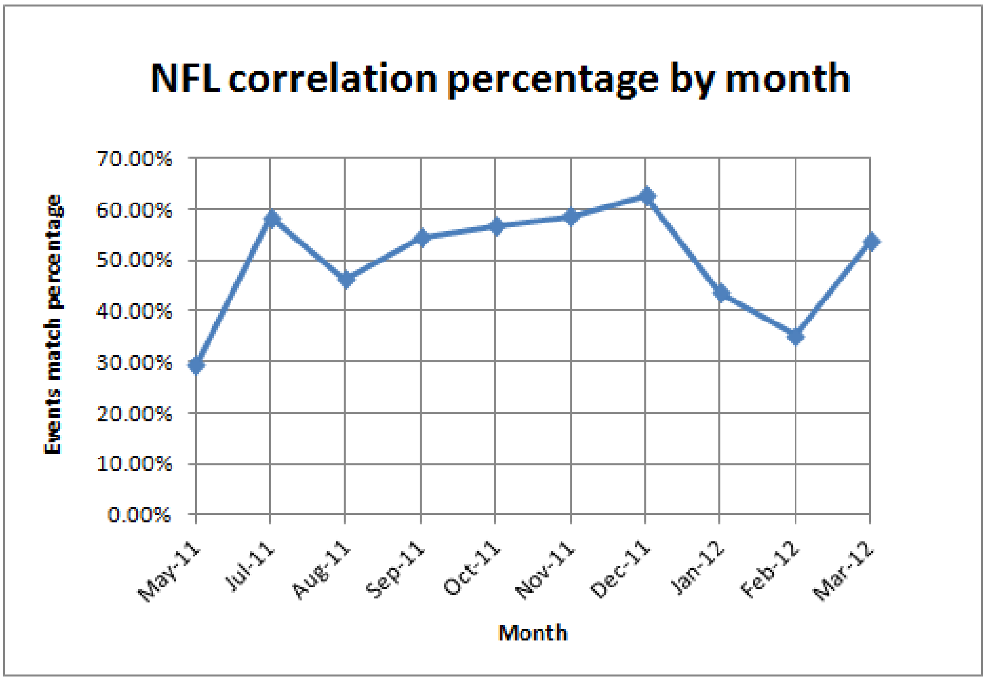


Figure 5.4: Correlation percentage in NFL data sets by month

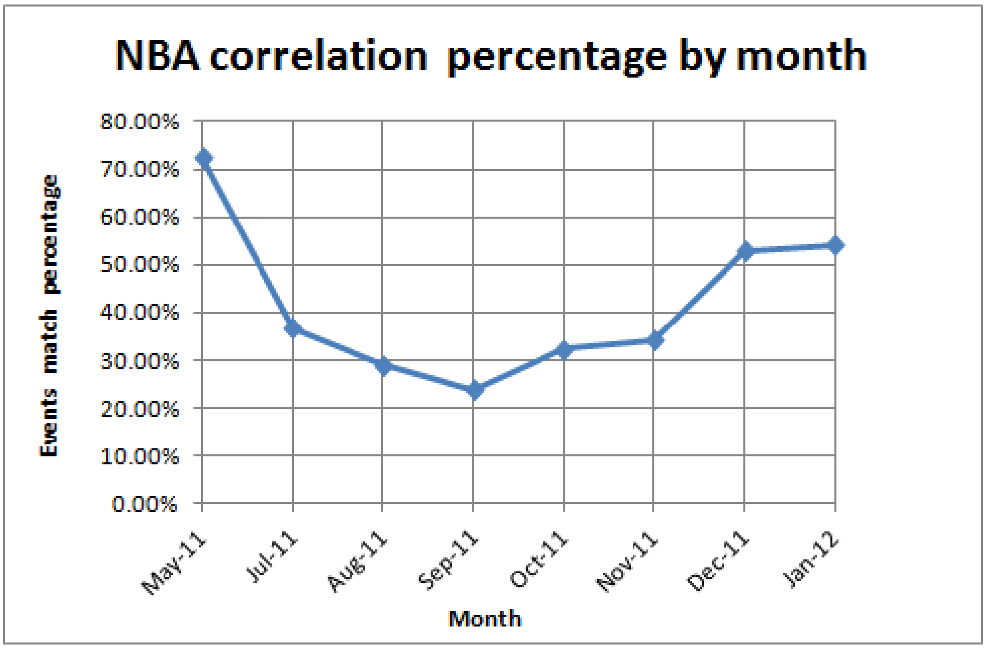


Figure 5.5: Correlation percentage in NBA data sets by month

tweet and post behavior. The correlation percentage in May 2011 is much higher the rest of month, and the data shows that 72.4% of Twitter events are able to correlated with forum events, and this is the very best situation which supports the predictability of Twitter. One possible reason for this result might be that May is the play-offs period for the NBA games during which the best teams fight for the championship.

Even though there are exceptions, the two situations above do indeed suggest that the match seasons provide advantages to traffic correlation between Twitter and the ProSportsDaily.com forums. Connections between tweets traffic surges and posts increases are more likely to appear when more human passions are involved. On the other hand, the MMA data gives poor correlation among all tweets and posts although it provides the most sufficient posts data compared to the NFL and NBA. One possible reason to explain this phenomenon is that there is no particular structured match rules and regular game season in MMA. The human attention is discrete, and no surging popularity can be found during specific parts of the year.

5.2.2 Is the correlation just appearing by chance?

The idea of this project is based on the natural connection between two topic-related websites. Identical themes and topics in which online users are interested are assumed to be the key point which associates tweets surges with posting increases. Results from the last chapter demonstrates a strong linear relationship between tweets and posts in the NFL and NBA data. However, one question that arises here is: "Is the topic-related traffic supposed to have stronger correlation? What if this correlation is just appearing by chance?"

The previous correlation analysis was fully conducted on data sets with the same theme: e.g., tweets about the NFL were connected with NFL posts. In the following sections, interesting experiments have been run and tested here with data sets coming from different fields in order to see how correlation works when things are merely correlated manually.

Beer+NFL and Beer+NBA

Twitter data for the keyword "beer" was collected from May 2011 until January 2012. It is quite a popular habit for sports fans to enjoy a glass of beer while watching professional games. If one searches manually on the Twitter search web page for the keyword combinations "beer" and "nfl" and "beer" and "nba," the results indicate that there are a collection of tweets mentioning both beer and the NBA or beer and the NFL. Beer seems to keep close company with sports. Therefore, it is reasonable to see whether there is a correlation between beer tweets and sports posts, and the strength of their connection if one exists.

Beer tweets resulted in 206 events in total. 133 of these events can be correlated to NFL posts within 24 hours. However, the time duration between two related events is very large, with the mean equal to 8.54 hours and standard deviation equal to 8.28 hours. Results from the NFL tweet data give 4.26 hours as the average and 6.41 hours as the standard deviation. Therefore the mean value is twice as large, and the standard deviation is increased by almost 2 hours as well.

Similarly, 41.3% of the beer Twitter events can be paired with a subsequent NFL forum event with the 6 hours' time constraint. Compared to the former results with 50.4% of Twitter events correlated within the NFL data, the connection between beer tweets and NFL posts can be proved looser in this case.

Losser correlation results from correlating beer tweets and NBA posts. With the strictest time constraint, the average of timespan between beer and NBA events (2.47 hours) is twice as large as in original NBA results (1.19 hours). In the same way, the standard deviation is also 0.43 hours bigger, which means that NBA posts increases are more solidly matched up with NBA tweets spikes.

Asparagus+NBA

The discussion above confirms the correlation is stronger when data is within the same topic than coming from two related fields. But will the correlation still be there if the chosen two topics are completely irrelevant? In order to dig deeply into how the traffic relationship works, data of a totally strange topic to professional sports was collected for two weeks. Asparagus, a spring vegetable, truly has nothing to do with sports. It is rare to hear people mention asparagus while talking sports like football or basketball and the vice versa. So analysis between asparagus tweets and sports posts can definitely explain the reliability of traffic correlation using topic-based keywords.

Asparagus tweets and NBA posts were chosen for study since they both have strong variations for events identified during the data collecting period. Data collection started from May 1st, 2012 and lasted for two weeks. In addition, tweets corresponding to a combination search of "nba" plus "asparagus" as keywords were also recorded in order to see whether anyone actually mention the two topics at the same time.

It turns out that people indeed do talk about asparagus and NBA at the same time. during the first two week in May, there are 8 tweets which contain both asparagus and nba. Since tweets are not recorded, the detail of how they were discussed together is unknown. One clue is found via a later Internet search that which revealed that a few people tweeted about a snack for NBA Finals called "Asparagus Prosciutto Wraps"[40]. Although this tweet was created in June 2011, it nevertheless might explain how asparagus and the NBA might be related to each other. However, this connection is small enough to be ignored, and asparagus and the NBA are still regarded as unrelated topics.

According to the events identification and correlation analysis process, there are 24 asparagus Twitter events and 30 NBA posts events. Within 6 hours, 8 of each can be matched up to be faux-"pairs". This means that completely arbitrary topics can be correlated in traffic. The Time difference average for these "pairs" is 2.93 hours. Comparing these results to the same periods in 2011 when the investigation was conducted only with NBA topic data, this time offset is 10 times larger than the real correlation in the NBA in May 2011, which was only 0.27 hours. In addition, these events "pairs" comprise only 33.3% of the events while the real correlation percentage is 72.4%.

As a conclusion, a mathematic relationship can be found in both relevant and completely irrelevant tweets and posts if the data is good to show enough traffic spikes. However, the strength of the correlation is based on how tightly the chosen topics are connected, and for totally arbitrary topic coupling, faux events "pairs" can not compete with the true correlation in either quantity nor quality.

5.3 Future data collection design

This project proposed a novel data mining process to analyze traffic correlation on Twitter and corresponding forums. It provides a successful methodology including techniques for data cleaning, events identification and correlation analysis. The mining process has been proved sound with high confidence by testing and running on different groups of data sets. According to these analytical results, no solid predictability of Twitter for topic-related websites can be concluded with existing data. But the primary reason giving difficulties in mining traffic correlation is the forum data.

First and foremost, the data from ProSportsDaily.com is poor. The NFL and NBA are the only analyzable data sets. The other data for the NHL and boxing are totally broken because they contain an overwhelming percentage of data gaps. The NFL and NBA data sets also contain plenty of holes, and lots of details of the data collection job are unknown and not available.

In addition, it has been found that the data collection script failed to record 0 post counts. This causes considerable confusion because there is no way to distinguish missing data from periods without any new posts. In addition, the broken data shows less traffic than expected, posts variations are difficult to identify in most of the days, and only a small amount data is able to show the expected traffic surge in posts. Forum data from Sherdog.net is unsatisfying as well. Although the data amount is more sufficient than those from ProSportsDaily.com, due to less popularity of MMA fights, both tweets and posts lack stable large increases in traffic, and this makes the event identification work difficult.

In addition, bad fundamental design of data collection must also be noted. In the existing data sets, the data covers only a single keyword for tweets, and only a single sub-forum for each topic is used for post data collection. In fact, both ProSportsDaily.com and Sherdog.net forum have many related sub-forums under each larger generalized topic. For example, on the ProSportsDaily.con website, the NFL sub-forum is coordinated with the NFL Draft, NFL Comparisons and many other team specific sub-forums under the general theme of Football, but none of data from those sub-forums listed above has been taken into consideration despite the fact they indeed belong to NFL topic. Besides, an overarching and generalized topic adds confusion to data observation since the area covered by the generalized topic is so vast.

Thus, the following suggestions and recommendations for data collection for future research in this field are made:

- Tweet searches should be more specific. A generalized searching keyword during data collection might become pointless, since no details are available for explaining the phenomena. For example, when the NFL data presented the surprising result that event correlation actually decreased during the championship, it is difficult to give reasonable explanations because no clues can be found in the original data. So it would be helpful if the NFL data tracked more information about the popularity of particular teams or matches or even the players. Twitter provides advanced search capabilities with multiple keywords and operators, so information gathering can be more focused with specified time or space constrain or even with negative or positive emotion. Influence from irrelevant message can also be reduced or eliminated this way.

- Forum posts needs to be both more widely gathered and also more specific. Learning from previous posts, which failed in both quantity and quality, the most desirable posts data should not only have a stable base amount but also be able to show traffic surges for events identification. Therefore popularity and concentration of the chosen forums matters the most. Results will benefit if the forum has high activity of online discussion on focused topics in the way that Twitter does. In addition, if a topic contains many sub-forums with different focuses like the ones ProSportsDaily.com has, then data from related sub-forums also needs to be taken into account.

- Richer data sources should be concerned. For each topic in Twitter, data should be collected from multiple corresponding websites to observe the best correlation fit among all possibilities. This project choose only forums because posts are believed to be a transparent measurement of resource consumption one can easily get from a web page. However, posts may not reveal the accurate relationship between posting amount and

> server performance. There are other factors. For instance, the owner of a website might pay attention to its visitation rate instead it is more capable of revealing public concern about certain events as well as indicating when the web servers are under pressure.

With the existing flawed data, results failed to conclude a strong correlation between Twitter traffic and forum posts. The results from the best data collecting situation, the NBA in May 2011, 70% of Twitter traffic surges were followed by forum posts increases within 16.2 minutes. In addition, the MMA data suggests a possible relationship between the magnitudes of tweets and posts. Therefore, it is reasonable to believe that with a sound data collection method, the best case scenario might be capable of showing a tight correlation between tweets and corresponding websites' traffic in both time scope and increasing scale. Optimistically speaking for the expectations for future research, 90% of tweets events might be able to predict a coming post surge within 30 minutes. In other words, although the current data prevents reaching definitive conclusions, the predictability of Twitter for resource consumption or traffic on related websites remains a promising area which need to be explored more.

Chapter 6

Conclusion

The goal of this thesis is to analyze the predictive ability of Twitter traffic for topic-related website resource requirements by examining the data variation correlation between Twitter events and corresponding web forum postings in order to develop a predictive algorithm. Major tasks have been accomplished during this project:

- Specific data cleaning procedures have been developed for both tweets and posts data.
- A data mining methodology has been developed.
- Event identification algorithms have been defined for both the Twitter data and the forum data.
- A generalized traffic events identification tool has been developed and implemented.
- Predictive algorithms of traffic correlation between Twitter and the topic-related forums have been proposed and analyzed.
- The traffic correlation mining methodology has been verified by repetitive testing on different kinds of data sets.
- The predictability of Twitter for corresponding websites has been fully analyzed and discussed based on existing data. Recommendations and suggestions about data collection for future research have been proposed.

With the existing flawed data, this data mining methodology failed to conclude the predictability of Twitter traffic. But the mining process for examining potential correlations between tweets and topic-related website resource demands has been clearly demonstrated. Viewed from the best case scenario in these experiments, with a sound data collection method, future research has a high potential to exploit Twitter to predict future traffic surges on corresponding websites within a short time period. In conclusion, the findings in this paper suggest a promising area of research in exploring the predictability of Twitter.

Bibliography

[1] "Raspberry pi? buying frenzy crashed websites," *http://www.zdnet.com/blog/igeneration/raspberry-pi-buying-frenzy-crashes-website/15463*.

[2] "Coca-cola, acura websites crashed during super bowl," *http://mashable.com/2012/02/06/coca-cola-acura-websites-crashed-during-super-bowl/*.

[3] "How many tweets per day are there on twitter," *http://www.quora.com/Twitter-1/How-many-tweets-per-day-are-there-on-Twitter*.

[4] J. Bollen and H. Mao, "Twitter mood as a stock market predictor," *Indiana University Bloomington*, 2011.

[5] M. O. Joshua Ritterman and E. Klein, "Using prediction markets and twitter to predict a swine flu pandemic," *University of Edinburgh*, 2009.

[6] M. Skoric, N. Poor, P. Achananuparp, E. Lim, and J. Jiang, "Tweets and votes: A study of the 2011 singapore general election," in *2012 45th Hawaii International Conference on System Sciences*, pp. 2583–2591, IEEE, 2012.

[7] G. Mao, "Real-time network traffic prediction based on a multiscale decomposition," *The University of Sydney*.

[8] M. F. Zhani and H. Elbiaze, "Analysis and prediction of real network traffic," *Journal of Networks*, vol. 4, November 2009.

[9] A. Sang and S. Li, "A predictability analysis of network traffic," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1, pp. 342–351, IEEE, 2000.

[10] A. A.Erramilli, O.Narayan and I.Saniee, "Performance impacts of multiscaling in wide area tcp/ip traffic," *INFOCOM 2000,TelAviv*, vol. 1, pp. 352–359, 2000.

[11] W. A.Feldmann, Agilbert and T.Kurtz, "The changing nature of network traffic: Scaling phenomena," *Computer Communication Review*, vol. 28, no. 2, pp. 5–29, 1988.

[12] J.-E. Tyvann, "On the predictability of server resources in online games, an investigative approach," *Oslo University College*, 2011.

[13] D. Larose, "An introduction to data mining," *Traduction et adaptation de Thierry Vallaud*, 2005.

[14] R. Nisbet, J. Elder, J. Elder, and G. Miner, *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.

[15] J. Jackson, "Data mining: a conceptual overview," *Communication of the Association for Information System*, vol. 8, pp. 267–296, 2002.

[16] J. H.Friedman, "Data mining and statistics: What's the connection ?," *Stanford University*.

[17] danah m. boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, pp. 210–230, 2007.

[18] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, pp. 591–600, ACM, 2010.

[19] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29–42, ACM, 2007.

[20] "How much time does the average user spend on social media," *http://www.the3dtechnologies.com/2012/03/03/how-much-time-does-the-average-user-spend-on-social-media-heres-the-answer/*.

[21] M. A. H. Adnan Rashid Hussain and N. P.Hegde, "Mining twitter using cloud computing," 2011.

[22] J. Ghannam, "Social media in the arab world: Leading up to the uprisings of 2011," *Center for International Media Assistance/National Endowment for Democracy*, vol. 3, 2011.

[23] Z. Harb, "Arab revolutions and the social media effect," *M/C Journal*, vol. 14, no. 2, 2011.

[24] S. Wasserman and J. Galaskiewicz, *Advances in social network analysis: Research in the social and behavioral sciences*. Sage Publications, Inc, 1994.

[25] J. Surma and A. Furmanek, "Data mining in on-line social network for marketing response analysis," in *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Confernece on Social Computing (SocialCom)*, pp. 537–540, IEEE, 2011.

[26] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, "Multiple kernel learning on time series data and social networks for stock price prediction," in *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, vol. 2, pp. 228–234, IEEE, 2011.

[27] F. Abel, E. Diaz-Aviles, N. Henze, D. Krause, and P. Siehndel, "Analyzing the blogosphere for predicting the success of music and movie products," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pp. 276–280, IEEE, 2010.

[28] "Most popular social media websites in the united states in february 2012, based on share of visits," *http://www.statista.com/statistics/71336/top-10-social-media-websites-in-the-us-by-market-share/*.

[29] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65, ACM, 2007.

[30] "Twitter: Often first, not always right," *http://edition.cnn.com/2012/02/13/tech/social-media/twitter-not-always-right/index.html*.

[31] R. O'Connor, "Global: Facebook and twitter'reshaping journalism as we know it'," 2009.

[32] M. Bryant, "Could facebook become a better news reporting tool than twitter?," *http://thenextweb.com/facebook/2011/02/25/could-facebook-become-a-better-news-reporting-tool-than-twitter/*, 2011.

[33] "Social media growth 2006-2011," *http://dstevenwhite.com/2011/12/29/social-media-growth-2006-2011/*.

[34] A. G.-T. A. Consultant, "Aggregating tweets: Search api vs. streaming api," *http://140dev.com/twitter-api-programming-tutorials/aggregating-tweets-search-api-vs-streaming-api/*, 2010.

[35] Y. Lai, X. Zheng, K. Chow, L. Hui, and S. Yiu, "Automatic online monitoring and data-mining internet forums," in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2011 Seventh International Conference on*, pp. 384–387, IEEE, 2011.

[36] D. S. Daniele Quercia, Michal Kosinski and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," *IEEE International Conference on Social Computing*, pp. 180–185, 2011.

[37] M. Krauthammer and P. Evans, "private communication," *Yale University School of Medicine*, 2012.

[38] G. Eysenbach, "Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact," *Journal of medical Internet research*, vol. 13, no. 4, 2011.

[39] "Using the twitter search api," *https://dev.twitter.com/docs/using-search*.

[40] "Tweets of asparagus prosciutto wraps as nba final snack," *http://www.mobypicture.com/user/FredCuellar/view/9791163*.

Appendix A

Data gap analysis and fixing script

```
#!/usr/bin/perl
# This script seals the small data gap in posts and store data in a new file
use Getopt::Std;
use strict ″vars″;
my $opt_string = ′f:′;
getopts(″$opt_string″,\my %opt) or usage() and exit 1;

my $file = $opt {′f′};
die ″Error: −f is mandatory\n″ unless $file;

my $newfile = $file . ″.new″;
if( −f $newfile) {
    warn ″Warining: $newfile exists, will be overwritten!\n″;
}
my $timestamp;
my $posts=0;
my $counter=0;
my $print = 1;

open(FILE,″$file″) or die ″Error: could not open $file, $!\n″;
open (NEWFILE,″>$newfile″) or die ″Failed to open $newfile $!\n″;
while (my $line=<FILE> ) {
    my $timediff;
    if ($line =~ /(\d+)\s+(\d+)/){

        if(not $timestamp){
            $timestamp = $1;
            $timediff = 0;
        }else{
            $timediff = $1 - $timestamp;

            my $sdinterval = 300;
            my $i = $timediff/$sdinterval;
            my $gap = int($i + 0.5);
            my $datagap = $gap -1;

            $posts = $2 - $posts;
            my $temptime = $timestamp;
            my $avgpost = $posts;
            if ($gap > 5){
                # ignore big gap
            }elsif($gap > 1){
                # seal the data loss
                $avgpost = $posts/$gap;
                while($gap > 1){
                    $temptime = $temptime + $sdinterval;
                    print NEWFILE "$temptime ; $avgpost ; $sdinterval \n";
                    $gap--;
                }

            }
            $timediff = $1 - $temptime;
            $timestamp = $1;
            print NEWFILE "$timestamp ; $avgpost ; $timediff \n";
            $posts = $2;

        }
    }

}
close(NEWFILE);
close(FILE);
```

Appendix B

NFL Twitter data behaviors during August 20th to 26th

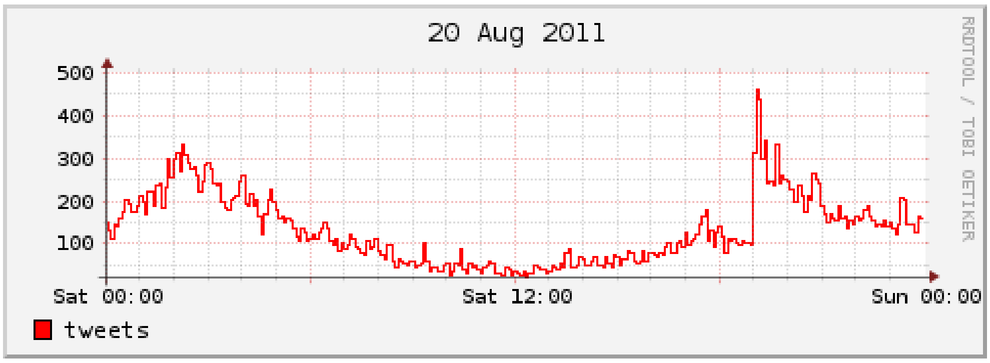


Figure B.1: NFL Tweets Variation - August 20th

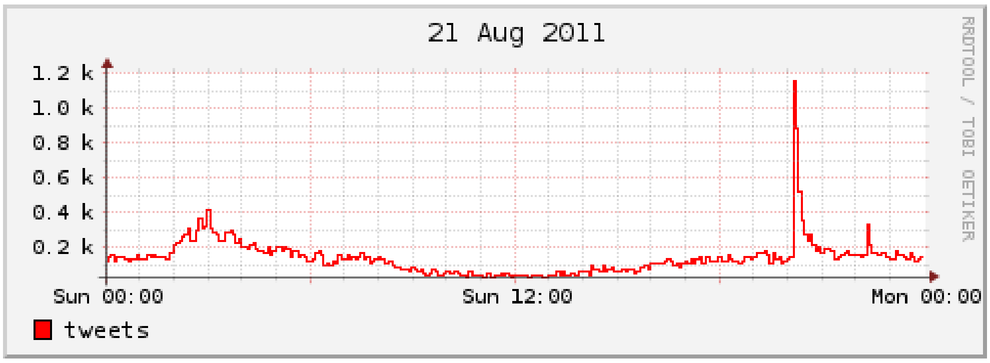


Figure B.2: NFL Tweets Variation - August 21th

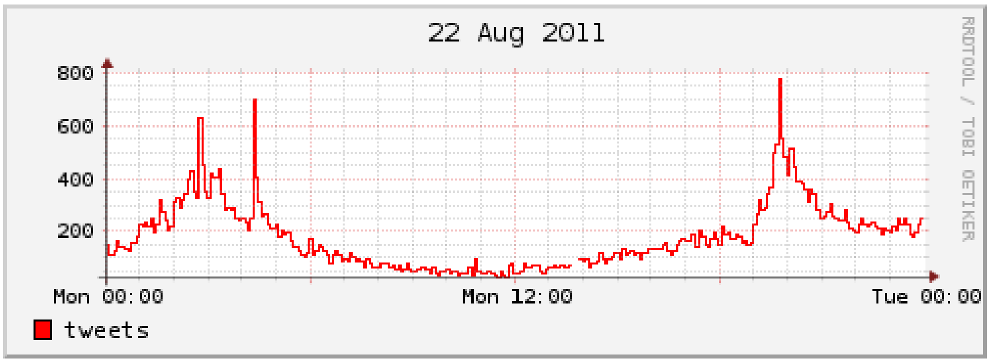


Figure B.3: NFL Tweets Variation - August 22th

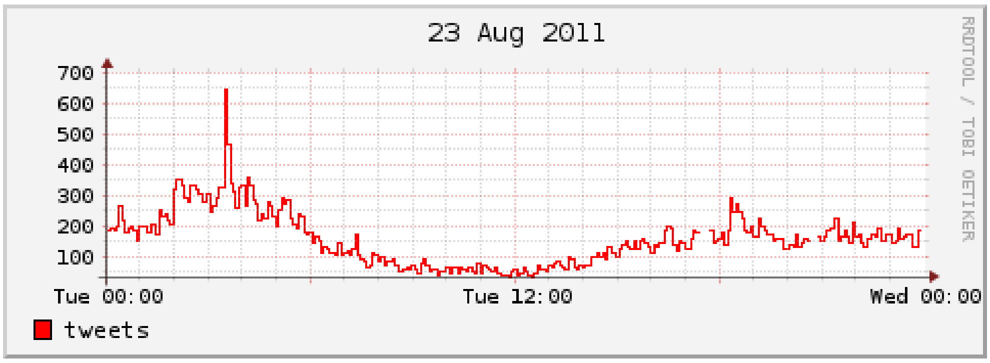


Figure B.4: NFL Tweets Variation - August 23th

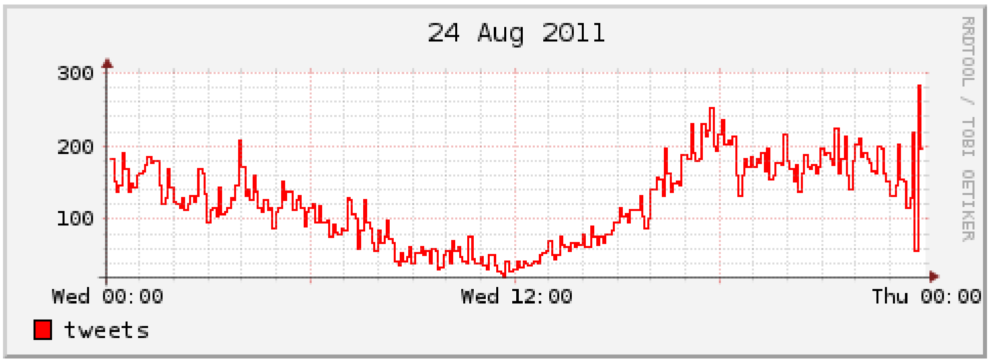


Figure B.5: NFL Tweets Variation - August 24th

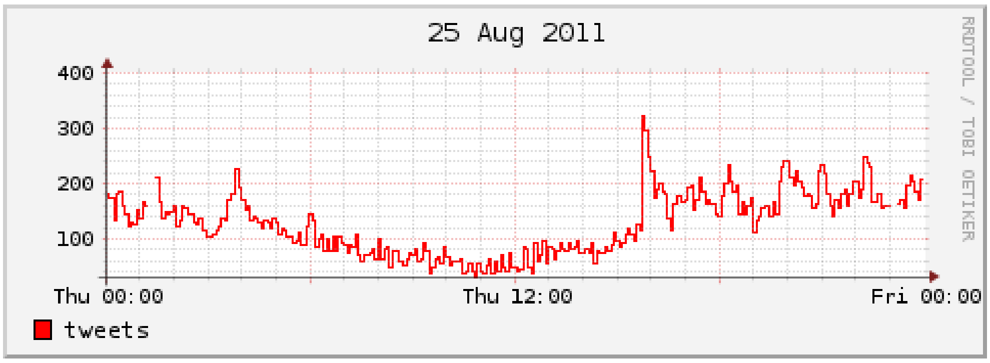


Figure B.6: NFL Tweets Variation - August 25th

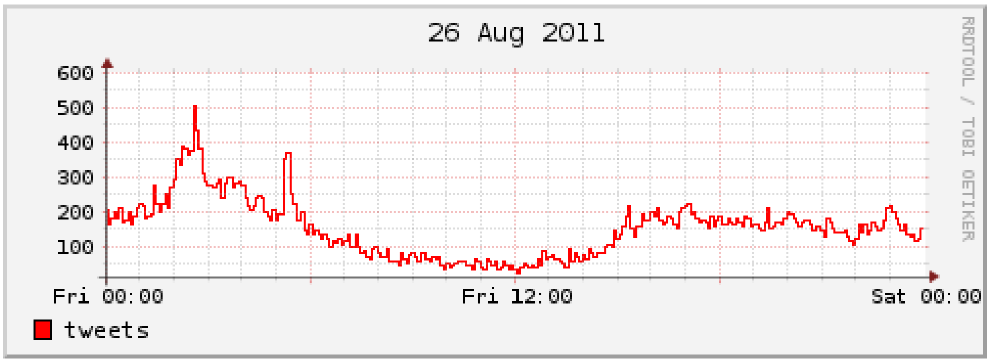


Figure B.7: NFL Tweets Variation - August 26th

Appendix C

Twitter events extraction tool: TwEventdetector.pl

```
#!/usr/bin/perl
###This script implements a detector for extracting Twitter events###
# Needed packages
use Getopt::Std;
use strict "vars";
use Statistics::Descriptive;
use Math::NumberCruncher;
use Time::Local;


my $DEBUG = 0;

my $opt_string = 'dhf:';
getopts("$opt_string",\my %opt) or usage() and exit 1;

if( $opt{'h'} ){
    usage();
    exit 0;
}

$DEBUG = 1 if $opt {'d'};
my $file = $opt {'f'};
die "Error: -f is mandatory\n" unless $file;

###Main script content
debug("Debug is enabled\n");
debug("Checking for the existance of $file\n");
if(not -f $file){
    die "Error:$file does not exist\n";
```

```perl
}
debug ("File exists!\n");

open(FILE,"$file");
my ($post1,$post2,$post3,$interval1,$interval2,$interval3,$diff15,$peakpoint,$peaktime);
my $tweetspeak = 0;
my (@slopes,@tweets);
while(my $line = <FILE> ){
    if($line =~ /(.*)\;\s+(.*)\s+\;.*\;\s+(.*)\s+\;.*\;.*/ ){
        debug("Reading lines from $file:$line\n");

        my $timestamp = $1;
        my $tweets = $2;
        my $slopein15;
        my $unixtimestamp = getUnix($timestamp);

        # Calculate tweets slope in last 15 minutes
        if (not $post1) {
            $post1 = $2;
        }elsif(not $post2){
            $post2 = $2;
            $interval1 = $3;
        }elsif(not $post3){
            $post3 = $2;
            $interval2 = $3;
        }else{
            $interval3 = $3;
            my $intervalsum = $interval1 + $interval2 + $interval3;
            $slopein15 = ($tweets - $post1)/$intervalsum;

            my $size = scalar(@slopes);
            if( $size < 288 ){
                push(@slopes,$slopein15);
                push(@tweets,$tweets);
            }else{

                shift(@slopes);
                push(@slopes,$slopein15);
                shift(@tweets);
                push(@tweets,$tweets);

                # calculate the slope threshold in last 24 hours

                my $stat1 = Statistics::Descriptive::Full->new();
                $stat1 -> add_data(@slopes);
                my $mean1 = $stat1-> mean();
                my $stdev1 = $stat1 -> standard_deviation();
                my $threshold1 = $mean1 + 3*$stdev1;

                # calculate the tweets threshold in last 24 hours
                my $stat2 = Statistics::Descriptive::Full->new();
                $stat2 -> add_data(@tweets);
                my $mean2 = $stat2-> mean();
                my $stdev2 = $stat2 -> standard_deviation();
                my $threshold2 = $mean2 + 1.5*$stdev2;

                # Spot the peak point of a tweets spike
                if (($slopein15 > $threshold1)&&($tweets > $threshold2)){
                    if ($tweetspeak eq 0){
                        $tweetspeak = $tweets;
                        $peakpoint = $line;
                        $peaktime = $unixtimestamp;
                    }else{
                        if ($tweets > $tweetspeak){
                            $tweetspeak = $tweets;
                            $peakpoint = $line;
                            $peaktime = $unixtimestamp;
                        }else{
                            # ignore smaller peak
                        }
                    }
                }
            }
            # Spot the tweets peak point
            if ($tweetspeak){
                if($tweets >= $tweetspeak){
                    $tweetspeak = $tweets;
                    $peakpoint = $line;
                    $peaktime = $unixtimestamp;
                }else{

                    if($unixtimestamp - $peaktime <= 950 ){
                        # ignore tweets draft in about 15 minutes
                    }else{
                        print $peakpoint;
                        $tweetspeak = 0;
                    }
                }



            }

            $diff15 = $tweets - $post1;
```

```
            $post1 = $post2;
            $post2 = $post3;
            $post3 = $tweets;
            $interval1 = $interval2;
            $interval2 = $interval3;
        }
    }
}
close (FILE);

###########
sub getUnix {

    my $date = $_[0];
    my %MONTHS = ("Jan" => 0, "Feb" => 1, "Mar" => 2,"Apr" => 3, "May" => 4, "Jun" =>5, "Jul" =>6, "Aug" => 7, "Sep" => 8,"Oct" => 9, "Nov" => 10, "Dec" =>11 );

    if ($date=~ /^\w{3}\s+(\w{3})\s+(\d+)\s+(\d\d):(\d\d):(\d\d)\s+(\d\d\d\d)/){
        my $month = $1;
        my $day = $2;
        my $hour = $3;
        my $minute = $4;
        my $second = $5;
        my $year = $6;
        my $unixtime = timelocal($second,$minute,$hour,$day,$MONTHS{$month},$year);

        return $unixtime;
    }
}

# prints the correct use of this script
sub usage {
    print "Usage:\n";
    print "-h Usage\n";
    print "-d Debug\n";
    print "-f Filename\n";
    print "./script [-d] [-h] -f filename \n";
}

sub debug {
    print "DEBUG: " . $_[0] if $DEBUG;
}
```

Appendix D

NFL forum data behaviors during September 17th to 21st

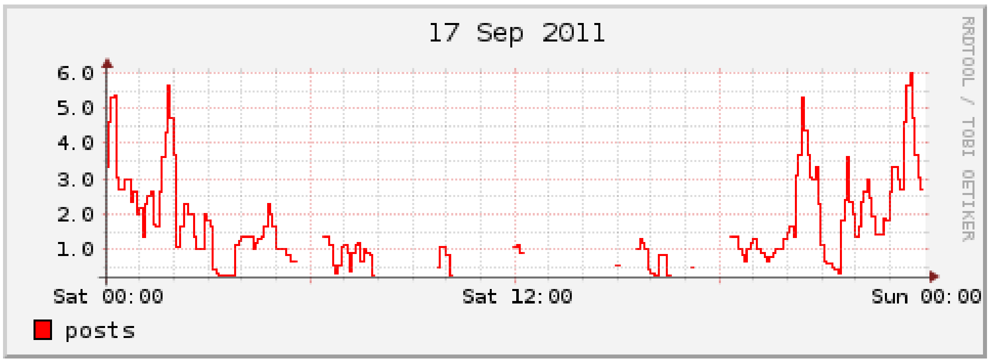


Figure D.1: NFL Posts Variation - September 17th

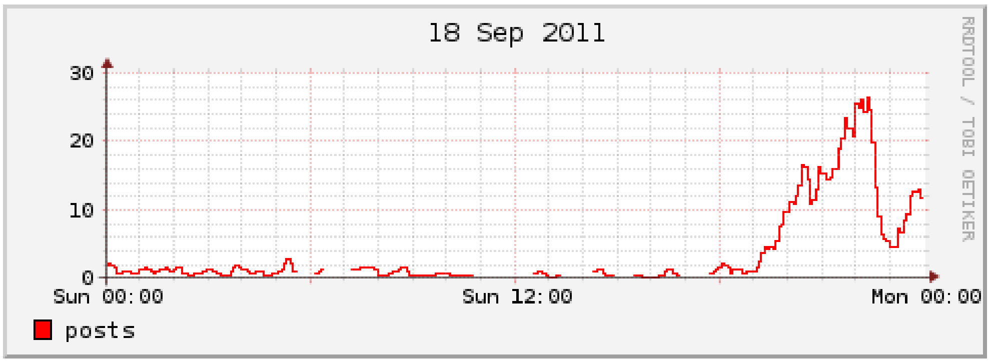


Figure D.2: NFL Posts Variation - September 18th

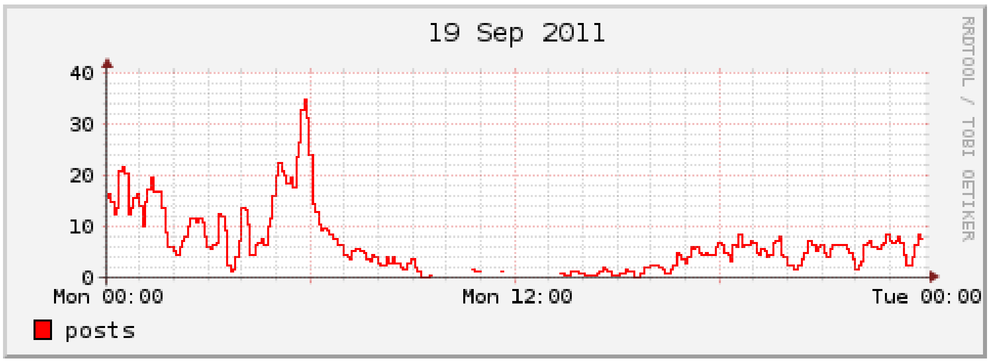


Figure D.3: NFL Posts Variation - September 19th

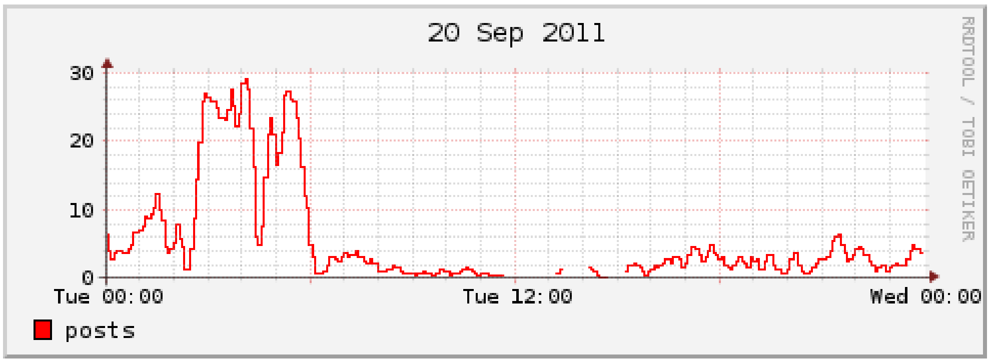


Figure D.4: NFL Posts Variation - September 20th

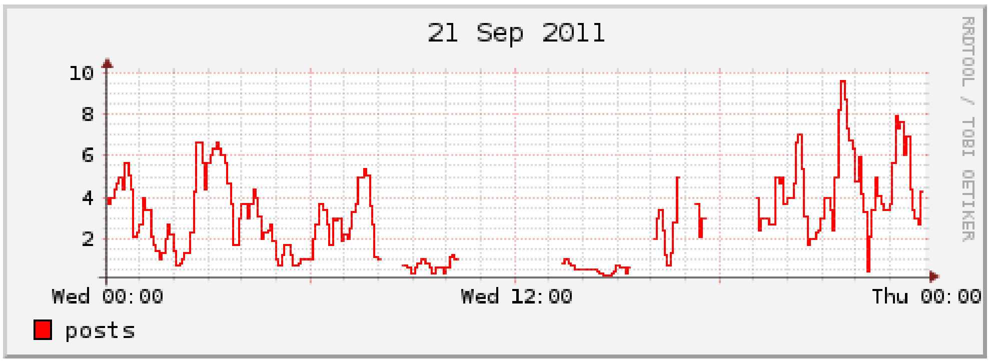


Figure D.5: NFL Posts Variation - September 21st

Appendix E

Forum events extraction tool: FrEventdetector.pl

```
#!/usr/bin/perl
###This script implements a detector for extracting Forum events###
# Needed packages
use Getopt::Std;
use strict ″vars″;
use Statistics::Descriptive;
use Math::NumberCruncher;
use Time::Local;

my $DEBUG = 0;
my $opt_string = ′dhf:′;
getopts(″$opt_string″,\my %opt) or usage() and exit 1;

if( $opt{′h′} ){
    usage();
    exit 0;
}

$DEBUG = 1 if $opt {′d′};
my $file = $opt {′f′};
die ″Error: −f is mandatory\n″ unless $file;

### Main script content###
debug(″Debug is enabled\n″);
debug(″Checking for the existance of $file\n″);
if(not −f $file){
    die ″Error:$file does not exist\n″;
}
debug (″File exists!\n″);
```

```
open(FILE,”$file”)or die ”Failed to open $file $!\n”;
my (@dates,@slopes,@posts,@contents,$date,$post,%LINE,$post1,$post2,$post3,$interval1,$interval2,$interval3);

while(my $line = <FILE> ){
    if($line =˜ /(\w+\s+\w+\s+\d+).*(\d\d\d\d)\s+\;\s+(.*)\;\s+(\d+)\;.*/){
        debug(”Read line from $file: $line\n”);
        my $currentpost = $3;
        my $slopein15;

        # Calculate the posts slope in last 15 minutes
        if (not $post1) {
            $post1 = $3;
        }elsif(not $post2){
            $post2 = $3;
            $interval1 = $4;
        }elsif(not $post3){
            $post3 = $3;
            $interval2 = $4;
        }else{
            $interval3 = $4;
            my $intervalsum = $interval1 + $interval2 + $interval3;
            $slopein15 = ($currentpost − $post1)/$intervalsum;
            $LINE{$line} = $slopein15;

            $post1 = $post2;
            $post2 = $post3;
            $post3 = $currentpost;

            $interval1 = $interval2;
            $interval2 = $interval3;
        }

        if (not $date){
            $date = $1 . ” $2”;
            push (@slopes,$slopein15);
            push(@dates,$date);
            push (@posts, $currentpost);
            push (@contents,$line);
        }else{
            my $newdate = $1 . ” $2”;
            if ($newdate eq $date){
                push(@slopes, $slopein15);
                push (@posts, $currentpost);
                push (@contents,$line);
            }else {
                my $threshold1 = avgslope(@slopes);
                my $threshold2 = avgposts(@posts);

                filter($threshold1,$threshold2);

                @slopes = ();
                @posts = ();
                %LINE = ();
                @contents = ();
                push(@slopes,$slopein15);
                push(@dates,$newdate);
                push (@posts,$currentpost);
                push (@contents,$line);
            }
            $date = $newdate;
        }
    }
}
close (FILE);

# Primary threshold: slope in 24 hours
sub avgslope {
    my @data = @_;
    my $stat = Statistics::Descriptive::Full->new();
    $stat->add_data(@data);
    my $mean = $stat->mean();
    my $stdev = $stat->standard_deviation();
    my $threshold = $mean + 2.4*$stdev;
    return $threshold;
}
# Secondary threshold : posts amount in 24 hours
sub avgposts {
    my @data = @_;
    my $stat = Statistics::Descriptive::Full->new();
    $stat->add_data(@data);
    my $mean = $stat->mean();
    my $stdev = $stat ->standard_deviation();
    my $threshold = $mean + 1.5*$stdev;
    return $threshold;
}

# Detect the posts peak points in 24 hours
sub filter {

    my $threshold1 = $_[0];
    my $threshold2 = $_[1];

    my ($peakpoint,$peaktime);
    my $postspeak = 0;

    foreach my $line (@contents){
        $line =~ /(.*)\;\s+(.*)\;.*\;\s+(\d+)/;

        my $timestamp = $1;
        my $unixtimestamp = getUnix($timestamp);
        my $avgpost = $2;
        my $actualpost = $3;

        if ($LINE{$line} > $threshold1) {
            if ($avgpost > $threshold2){
                debug(”START::$line”);
                # spot upgoing trend in posts
                if ($postspeak eq 0){
                    $postspeak = $actualpost;
                    $peakpoint = $line;
                    $peaktime = $unixtimestamp;
                    debug(”Upgoing Point: $peakpoint”);
                }else{
                    if($actualpost > $postspeak){
                        $postspeak = $actualpost;
                        $peakpoint = $line;
                        $peaktime = $unixtimestamp;
                        debug(”Bigger Spikes Around: $peakpoint”);
                    }else{
                        # ignore smaller spikes around
                    }
                }
            }
        }

        # spot posts peak point
        if($postspeak){
            if($actualpost > $postspeak){
                $postspeak = $actualpost;
                $peakpoint = $line;
                $peaktime = $unixtimestamp;
                debug(”Increasing Value: $peakpoint”);
            }else{
                if ($unixtimestamp − $peaktime <= 950 ){
                    # ignore posts drift in about 15 minutes
                }else{
                    debug(”END::(Peak Point)$peakpoint”);
```

```perl
                    my $timetemp = scalar localtime($peaktime);
                    print "$peaktime;$postspeak;$timetemp;\n";
                    $postspeak = 0;

                }
            }
        }

    }
}

# Convert timestamp to Unix time
sub getUnix {

    my $date = $_[0];
    my %MONTHS = ("Jan" => 0, "Feb" => 1, "Mar" => 2, "Apr" => 3, "May" => 4, "Jun" => 5, "Jul" => 6, "Aug" => 7, "Sep" =>8,"Oct" => 9, "Nov" => 10, "Dec" =>11 );
    if ($date=~ /^\w{3}\s+(\w{3})\s+(\d+)\s+(\d\d):(\d\d):(\d\d)\s+(\d\d\d\d)/){
        my $month = $1;
        my $day = $2;
        my $hour = $3;
        my $minute = $4;
        my $second = $5;
        my $year = $6;
        my $unixtime = timelocal($second,$minute,$hour,$day,$MONTHS{$month},$year);

        return $unixtime;
    }
}

# prints the correct use of this script
sub usage {
    print "Usage:\n";
    print "-h Usage\n";
    print "-d Debug\n";
    print "-f Filename\n";
    print "./script [-d] [-h] -f filename\n";
}

sub debug {
    print "DEBUG: " . $_[0] if $DEBUG;
}
```

Appendix F

NFL Tweets and posts data comparison - November 2011

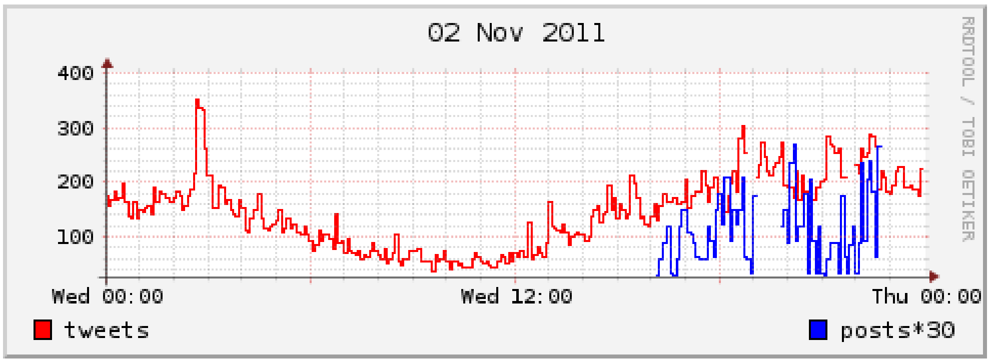


Figure F.1: NFL tweets and posts behavior - November 2nd

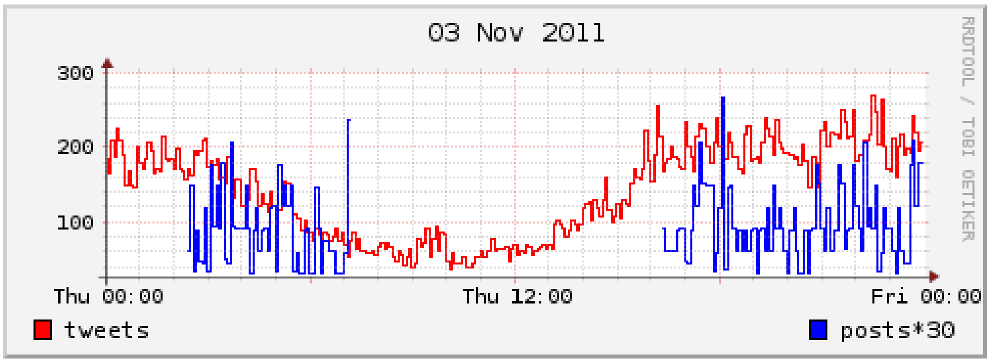


Figure F.2: NFL tweets and posts behavior - November 3rd

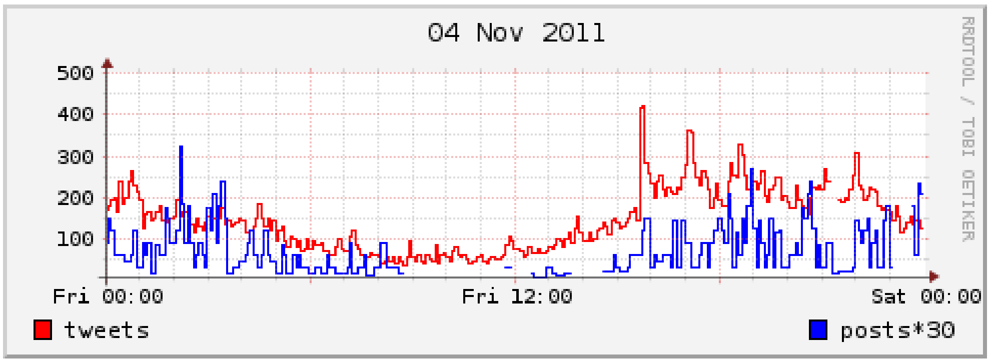


Figure F.3: NFL tweets and posts behavior - November 4th

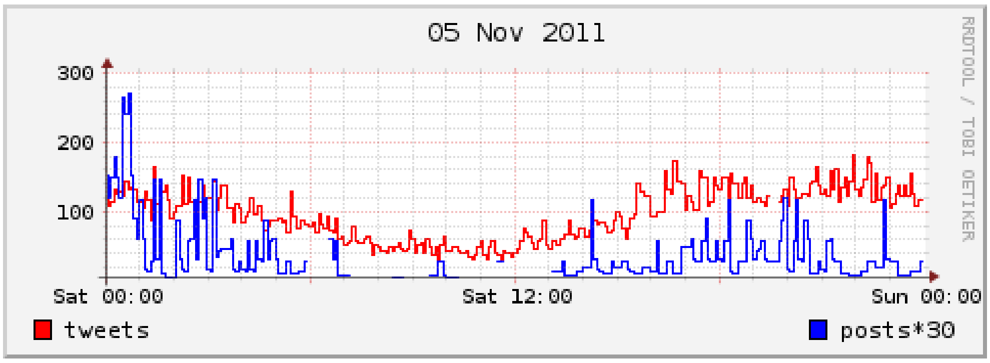


Figure F.4: NFL tweets and posts behavior - November 5th

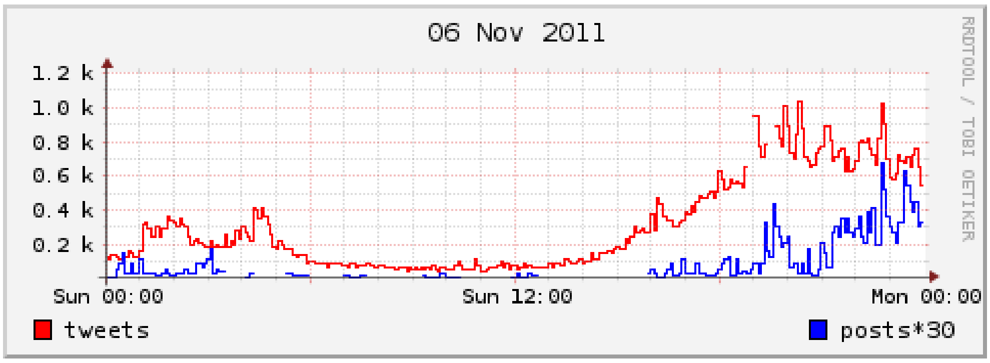


Figure F.5: NFL tweets and posts behavior - November 6th

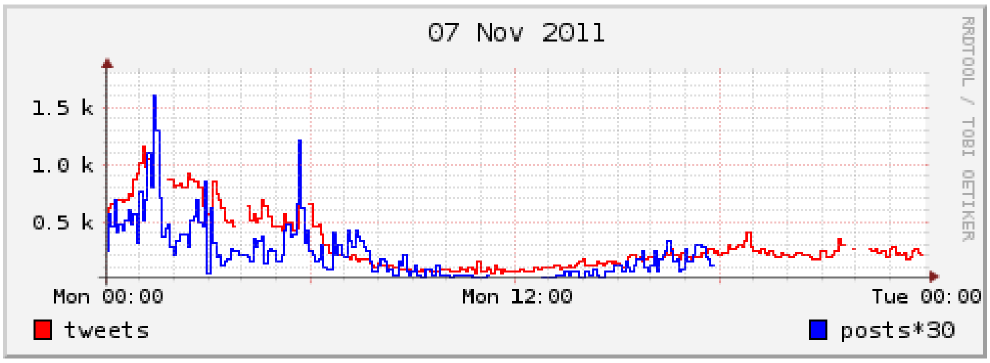


Figure F.6: NFL tweets and posts behavior - November 7th

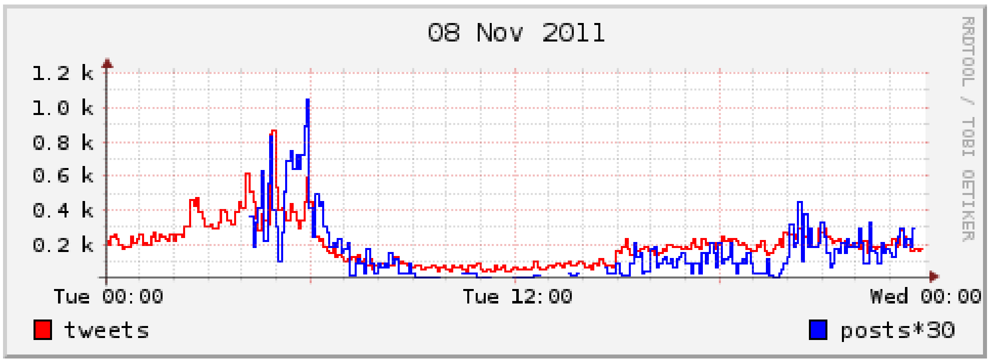


Figure F.7: NFL tweets and posts behavior - November 8th

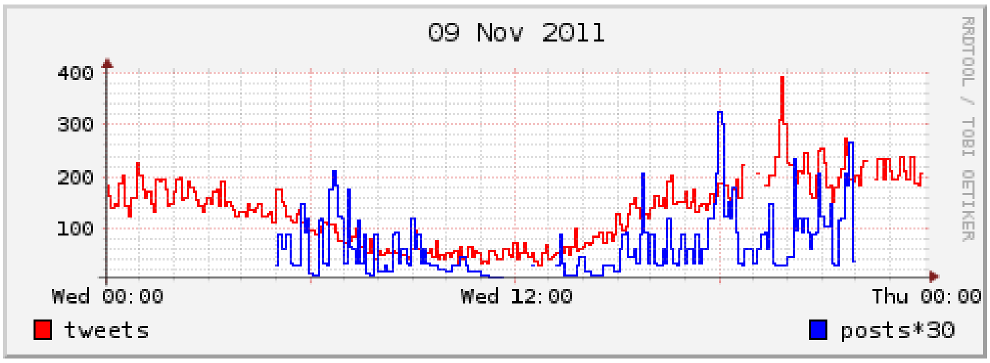


Figure F.8: NFL tweets and posts behavior - November 9th

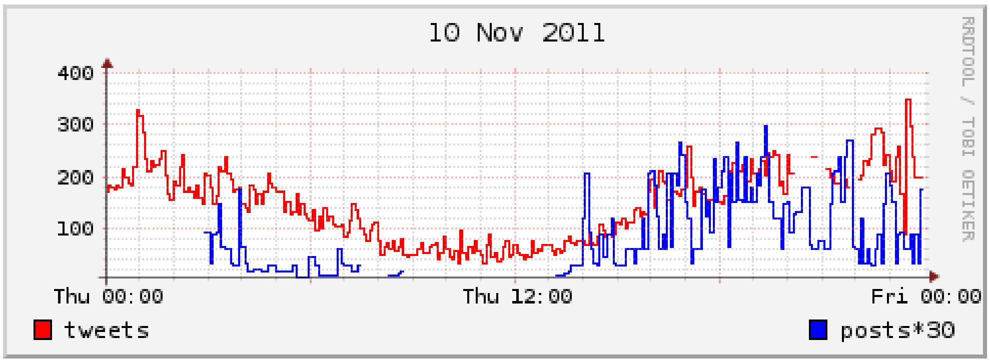


Figure F.9: NFL tweets and posts behavior - November 10th

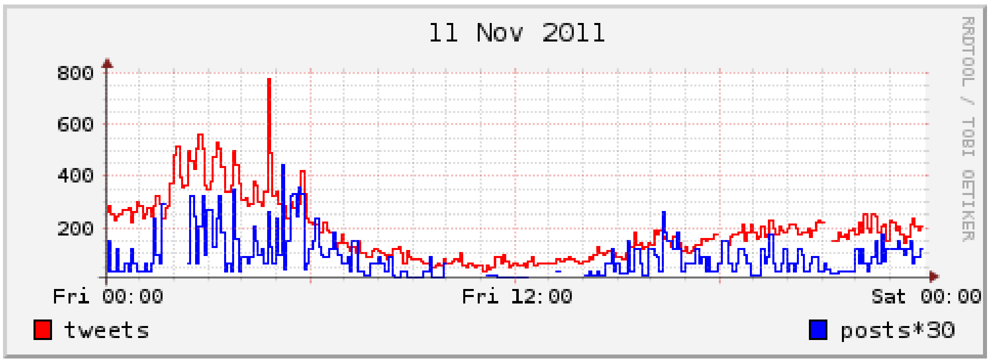


Figure F.10: NFL tweets and posts behavior - November 11th

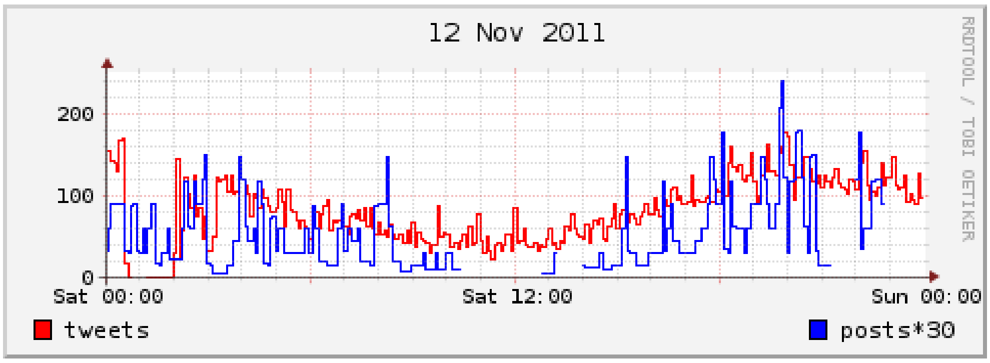


Figure F.11: NFL tweets and posts behavior - November 12th

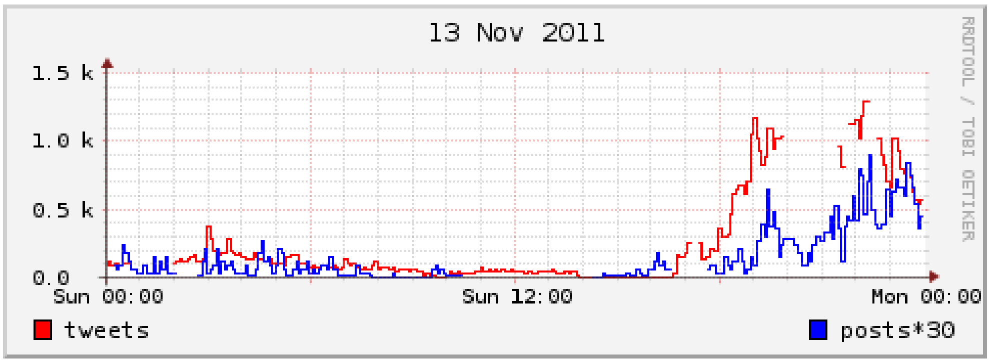


Figure F.12: NFL tweets and posts behavior - November 13th

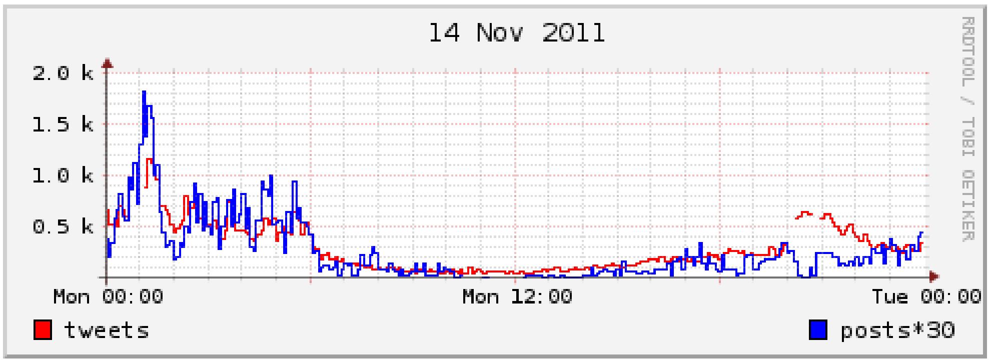


Figure F.13: NFL tweets and posts behavior - November 14th

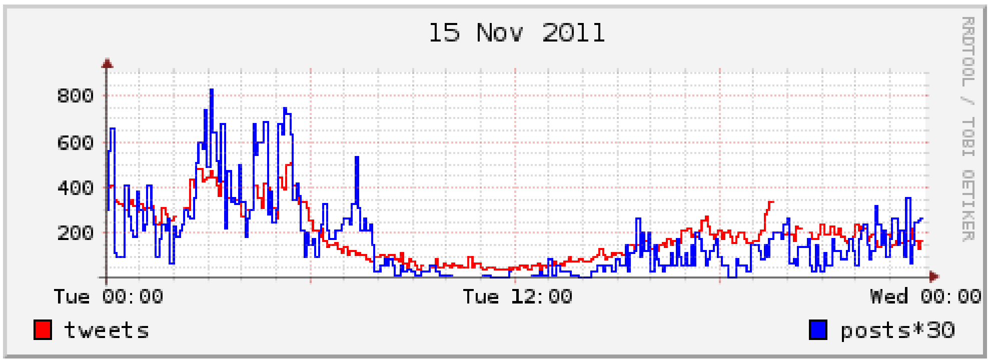


Figure F.14: NFL tweets and posts behavior - November 15th

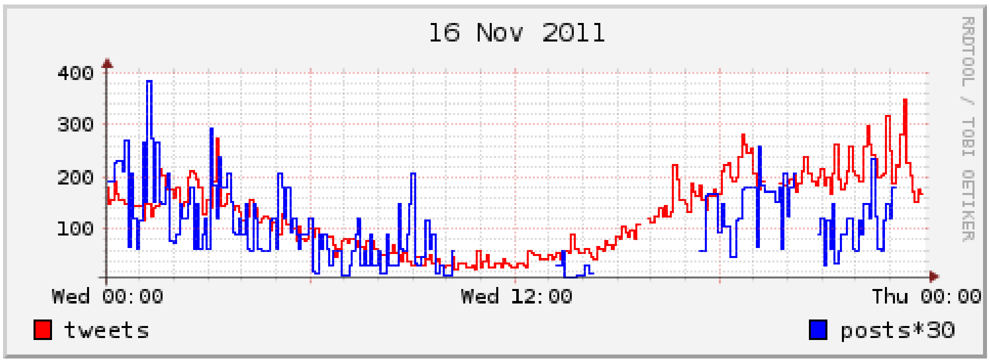


Figure F.15: NFL tweets and posts behavior - November 16th

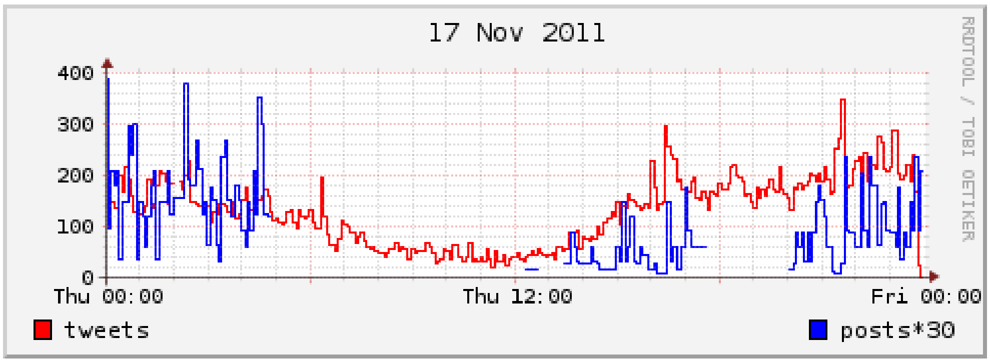


Figure F.16: NFL tweets and posts behavior - November 17th

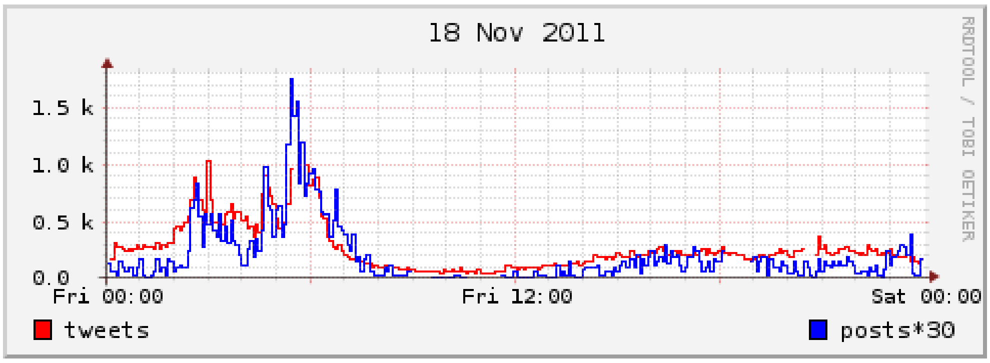


Figure F.17: NFL tweets and posts behavior - November 18th

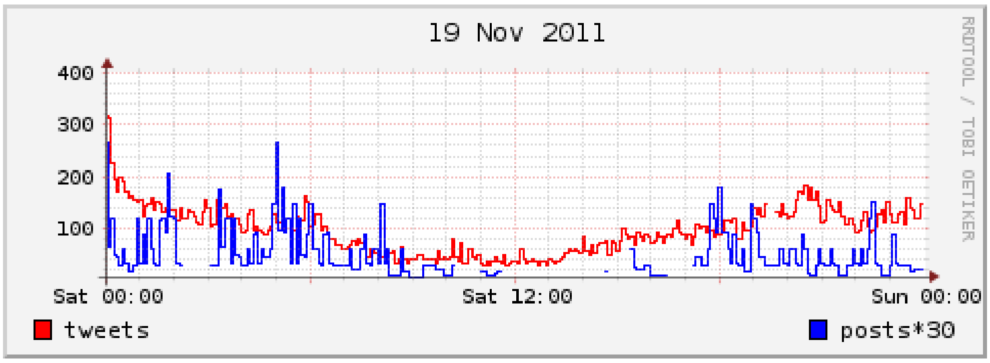


Figure F.18: NFL tweets and posts behavior - November 19th

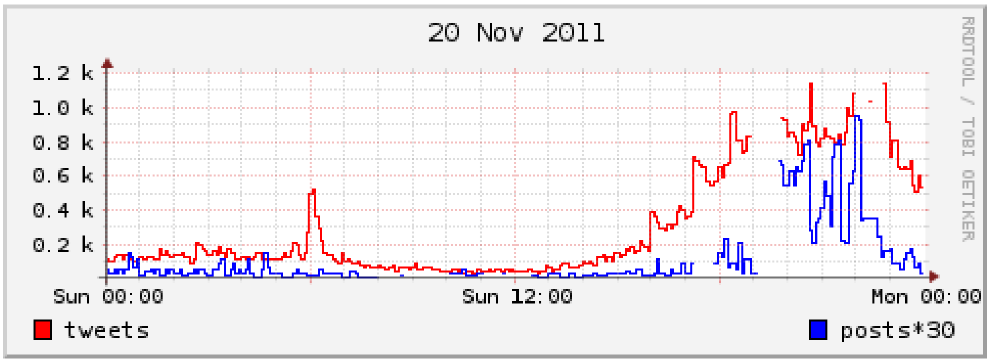


Figure F.19: NFL tweets and posts behavior - November 20th

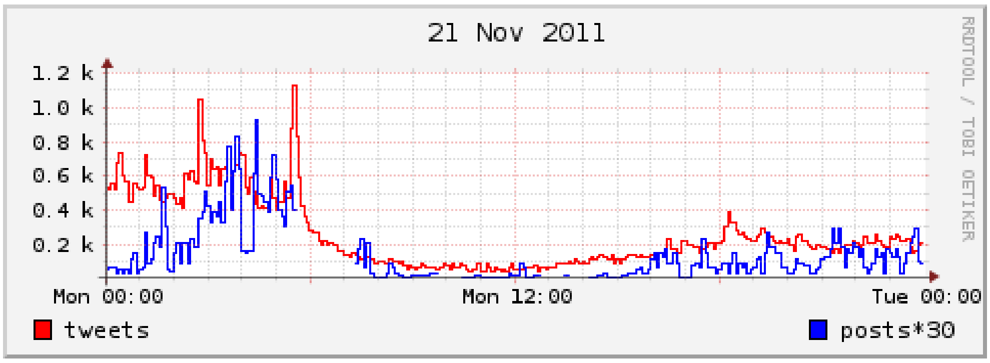


Figure F.20: NFL tweets and posts behavior - November 21st

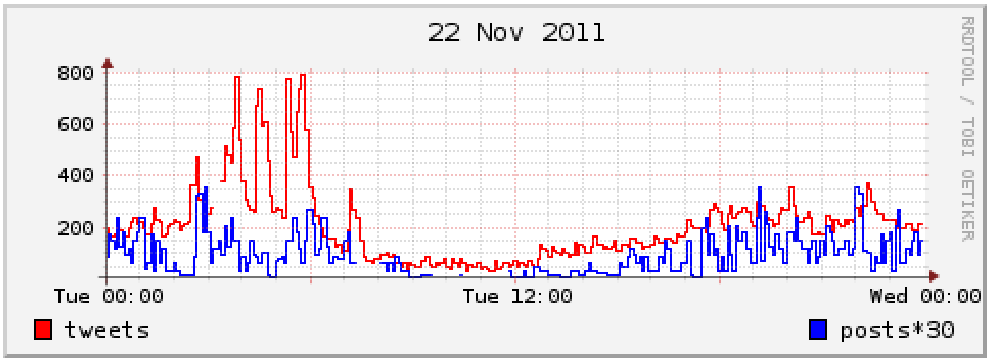


Figure F.21: NFL tweets and posts behavior - November 22nd

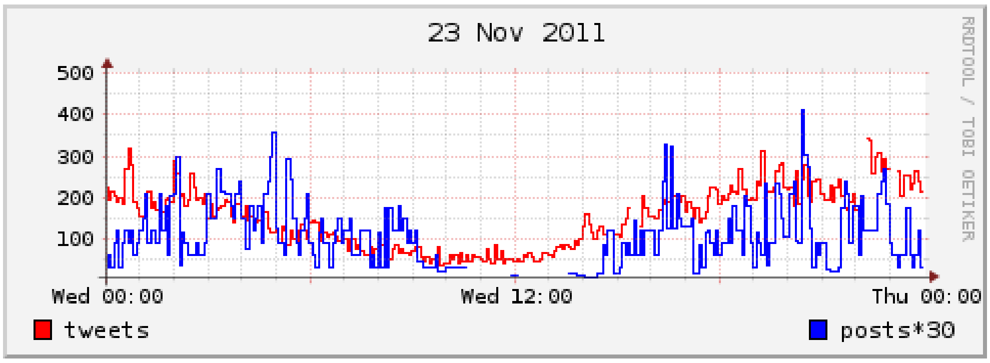


Figure F.22: NFL tweets and posts behavior - November 23rd

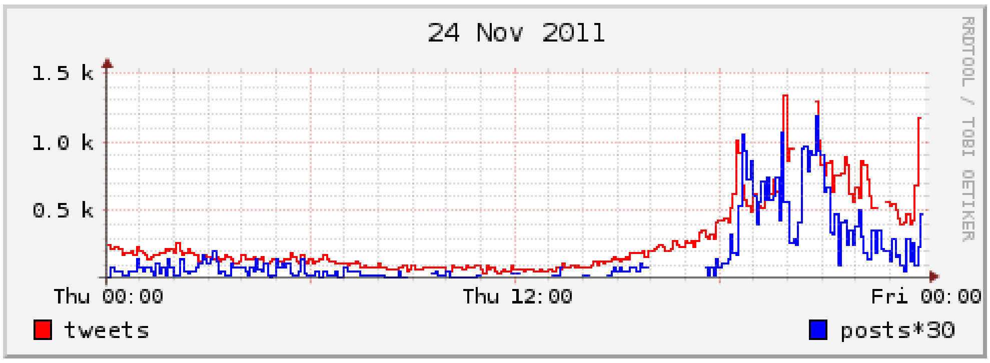


Figure F.23: NFL tweets and posts behavior - November 24th

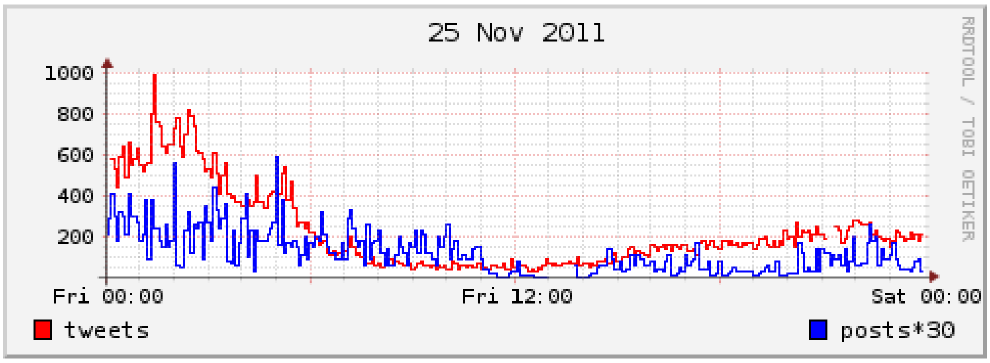


Figure F.24: NFL tweets and posts behavior - November 25th

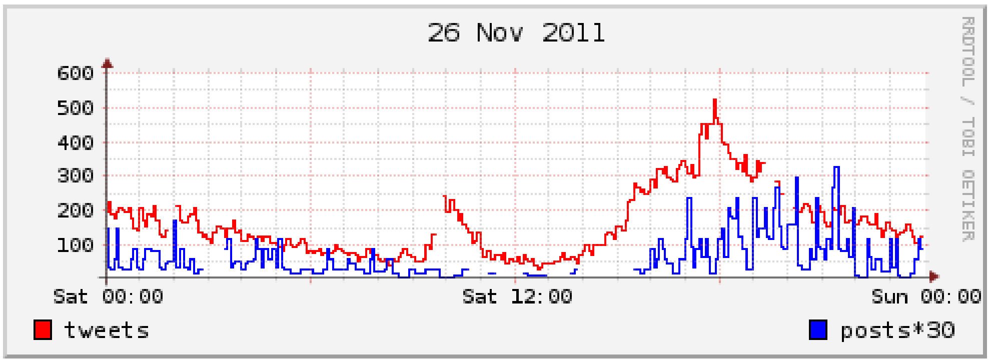


Figure F.25: NFL tweets and posts behavior - November 26th

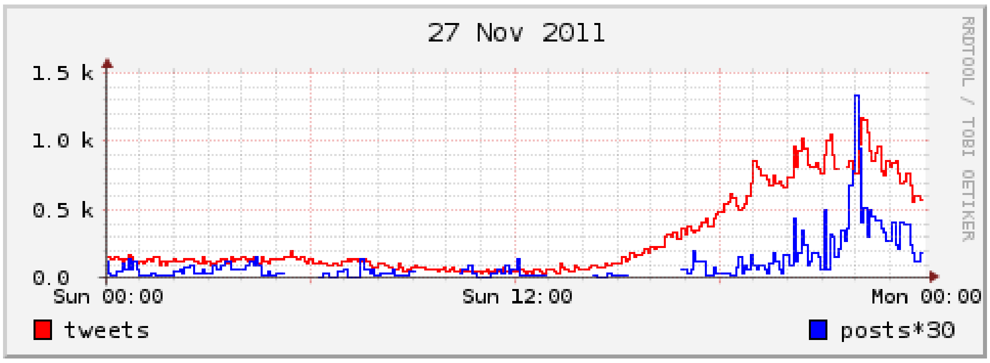


Figure F.26: NFL tweets and posts behavior - November 27th

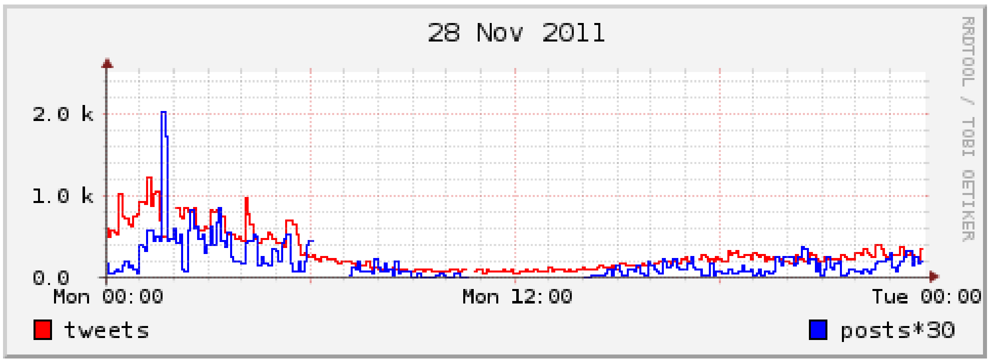


Figure F.27: NFL tweets and posts behavior - November 28th

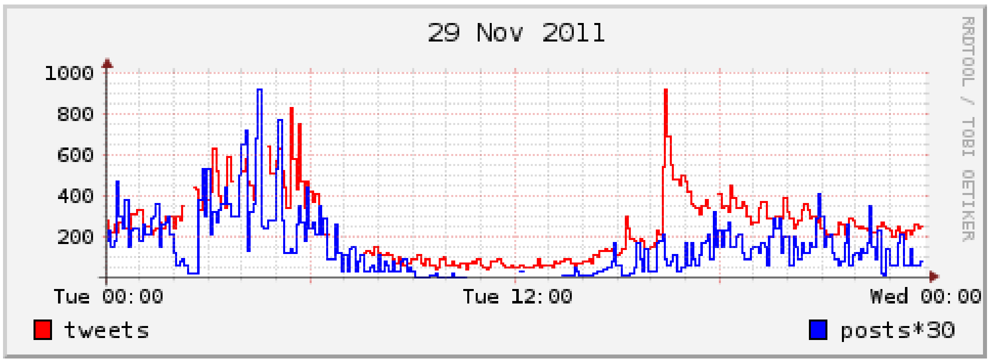


Figure F.28: NFL tweets and posts behavior - November 29th

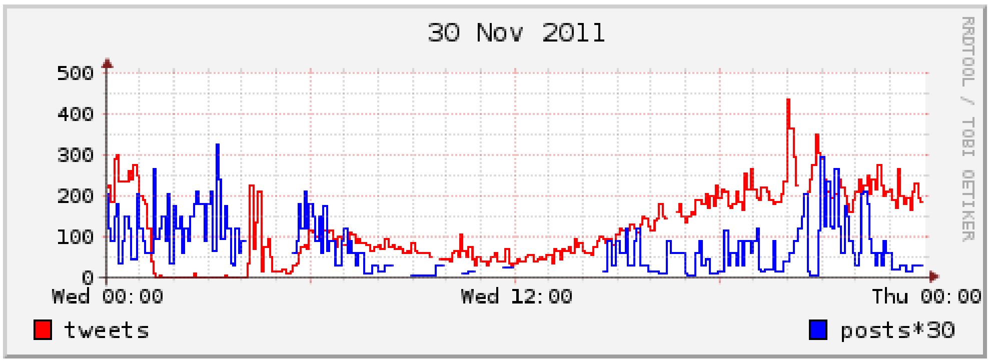


Figure F.29: NFL tweets and posts behavior - November 30th