Analyzing Big Data with Twitter: Mid-Project Report

Pigskin: Visualizing Football Tweets

Team Members & Roles

Andrew Chao (@acchao / andrew.c.chao@ischool)
Project Manager, primary researcher
Gilbert Hernandez (@thegilby / gahernandez@ischool)
Coder – visualization (d3.js), website, frontend
Jacob Portnoff (@jacobportnoff / jacob.portnoff@ischool)
Writer – data gathering/processing, backend

Mentor

Gilad Mishne (@gilad / gilad@twitter.com)

Goals

The hope of this project is to build an exploratory visualization application from data gathered from Twitter. We will gather data during football games from Twitter to determine the number of tweets that relate to professional football games. We hope to determine the most popular teams every week as well as a number of different metrics by which to compare games, teams, and the sport as a whole.

Strategy

In order to achieve our goal we developed a strategy of heavy tweet collection followed by a significant analysis. While we originally explored the angle of pursuing data collection and analysis simultaneously using Twitter's search filter, we discovered that our term/filter list was longer than what Twitter would allow (over 500 terms versus a maximum of 400 search terms). Further we wanted to explore not only the interest in particular football games, but the general interest in football versus total twitter 'chatter.'

As a result, we established a data collection that centered on gaining as much data as possible during the times of interest. Specifically, we did data collection from 10:00am to 9:00pm (Pacific Time) on Sunday. This time frame allowed us to see what was happening on twitter during football games on Sunday.

Our analytics strategy was designed to minimize our work load and the number of false positives and false negatives. As a result we narrowed down quite a lot on what it was we were interested in seeing. Tweet rates and number of tweets were particularly interesting to us. We plan on utilizing such information to determine popularity of teams and games. In addition to the items of specific interest, our strategy accounts for the variety of timeframes during which our data collection takes place. The level of a team's twitter popularity across multiple game days is particularly interesting, especially when considering that other events are influencing interest.

Project Timeline (Done by NFL Weeks)

Week 7 Sunday (October 22, 2012) – collect data on all games
ID boundaries of the data
Assumptions: time, user, tweet, location, hashtags
Gather research material
Begin working on exploratory visualization application
Build expectation model
Build corpus of tags across teams, general football stuff
ID how to identify location without geolocation tag on (if possible)
Sentiment analysis tool testing

Week 8 Sunday (October 28, 2012) – collect data on all games
Visualization model for football interest versus other trends

Week 9 Sunday (November 4, 2012) – collect data on all games
Visualization model for mapped information

Week 10 Sunday (November 11, 2012) – collect data on all games
Visualization mockups for all expected outcomes + sample code
Sentiment analysis testing completed

30%-40% done point – Nov 13 – 1st report due

Week 11 Sunday (November 18, 2012) – collect data on all games
Testing hypotheses
Visualization application work

Week 12 Sunday (November 25, 2012) – collect data on all games
Continued analysis and programming

Week 13 Sunday (December 2, 2012) – collect data on all games
Visualization application completed
Finishing touches

100% done – Dec 10

Literature Review

1. Understanding the Demographics of Twitter Users

by Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, J Niels Rosenquist

This paper tries to analyze the demographic of twitter users. Researchers have begun using twitter as a means of measuring and predicting real world phenomena; such predictions would be greatly enhanced with better demographic data and provide more insight to possible biases.

Takeaways

75.3% of publicly visible users listed a location
Gender can be predicted by the user of the first name. With a US-centric corpus, there was a match for names of 64.2% of users via a list of 5836 names.
Last name is not a good indicator of ethnicity.

2. Using Twitter to Detect and Tag Important Events in Live Sports

by James Lanagan and Alan F. Smeaton

Summary

This paper compares the effectiveness of Twitter vs audio-visual content analysis on detecting important events during live sports. The primarily method for analysis for twitter is through the use of volume of tweets or the delta through time.

Takeaways

Twitter is a reactionary response
It’s effective at capturing goals scored, but not events like bookings (violations). It brings up an interesting point on what is worth tweeting in sports.

Goals scored closer to the end of the game will generate more conversation. The base line of conversation rises which could make threshold detection more difficult.

3. Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games

by Siqi Zhao, Lin Zhong, Jehan Wickramasuriya, Venu Vasudevan

Summary
Using Twitter to discover in real-time, events occurring in a NFL football game within 40 seconds and with 90% accuracy.

Takeaways

Good at predicting events such as touchdowns but poor at fumbles (64%) and other less important events.

Streaming API is more useful for real time data.

Team names appear in 60% of game-related tweets.

top 10 most frequent words are either game terminology or team names.

Processing was done by first removing urls, @username, emoticons, punctuation, and stop words.

Post rate was used to determine the important event with an adaptive window, but streaming api has a post rate limit of 50 tweets per second; failed during superbowl.

check out www.sportsense.us

only really works with predetermined keywords.

4. Lexical Normalisation of Short TExt Messages: Makn Sens a #twitter

by Bo Han, Timothy Baldwin

Summary
Due to the 140 character limit, this paper proposes a method for identifying and normalising ill-formed words. Some of the important issues that need to be addressed include the deabbreviation of words such as “b4” to their canonical versions, “before”. The proposed method is a cascaded one that builds upon sms text normalization, using only single token words; It also excludes hashes, mentions, and urls.

Takeaways

Uses word similarity, dictionary look ups,

Not all ill-formed words provide useful context.

5. Analyzing Twitter for Social TV: Sentiment Extraction for Sports

by SiQi Zhao, Lin Zhong, Jehan Wickramasuriya and Venu Vasudevan

Summary

Building upon their past research, they use their real-time twitter sports event detection website to assist in analyzing real-time sentiment of NFL games. Their goal is assist with advertisers and possible product placement that makes more sense and coincides with the sentiment of tv viewers.

Takeaways

Emoticons are a strong signal and useful source for determining sentiment in the tweet.

Positive Sentiment is the dominant sentiment, comprising of upwards 90% of tweets.

Sportsense recognized 92% of touchdowns, 75% of interceptions, 74% of fumbles, 67 % of field goals in 33 games.

6. The Wisdom of Bookies? Sentiment Analysis vs. the NFL Point Spread

by Yancheng Hong and Steven Skiena

Summary

Using sentiment analysis, they were able to identify the winner roughly 60% of the time; the prediction of the winner was better in the second half of the season compared to the first.

Takeaways

local media is less reliable than national media due to local bias.

Social media is just as informative as professional newspaper media.

7. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena

by Johan Bollen, Huina Mao, Alberto Pepe

Summary

By adapting a traditional psychometric instrument, they were able to map out a six dimensional mood vector for each day to twitter users. The six mood states consisted of tension, depression, anger, vigor, fatigue, and confusion.

The following steps were used to prepare data for POMS scoring:

1. Separation of individual terms on white-space boundaries
2. Removal of all non-alphanumeric charaters from terms
3. Conversion to lower-case of all remaining characters
4. Removal of 214 standard stop words, including highly common verb-forms
5. Porter stemming of all remaining terms in tweet.

Takeaways

General mood matched well to large events even if delayed.

8. Event Summarization Using Tweets

by Deepayan Chakrabarti and Kunal Punera

Summary

Using Hidden Markov Models, they explored the ability to summarize event-based tweets. They assess highly structured and recurring events such as football games. They apply a summarization method on each tweet, and use time intervals to separate events. Their algorithm also has difficulty with proper names.

Takeaways

Events are bursty.

subevents can also occur within close temporal proximity.

This paper focuses on generalizing learning a language model to identify the event, but within the case of our project, we have a set of well defined key terms.

Event detection on scoring plays are easier to find.

9. TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration

by Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, Robert C. Miller

Summary

They utilize a novel streaming algorithm to identify peaks of high tweet activity and label them; This also allows for the drill down to subevents. . Twitinfo utilizes signal processing literature. The algorithm requires that users first define the event by providing a keyword query.

Takeaways

- Also utilizes the tweet post rate for identifying the peak.
- Served mostly as an exploratory tool.
- Users that evaluated the visualization did not trust the sentiment analysis.

Accomplishments to Date

- Further refinement of project goals
  - Narrowed down project focus and our definition of what a “football” tweet vs. “non-football” tweet is.
- Data gathering
  - 4 Sundays worth of tweets (roughly 8 million tweets, 2 million per day) gathered so far
  - Data structure: date time, screen name, tweet, hashtags, and geolocation
  - Began transforming data to JSON or .csv format for visualization use
- Algorithm development
  - Tweets weighted for expected football relevance
    - Tweets, hashtags, and user names were evaluated
    - Tweets checked for football relevant terms and teamnames
    - Hashes and usernames evaluated for teamnames
  - Initial individual Sunday tweet evaluation completed
- Software architecting/Coding
  - Java (twitter4j) implemented data gathering

- Python (tweepy) implemented data gathering
- Set up cronjob to automatically gather data on Sundays
- Pig implementation of algorithms
- d3.js implementation of sample graphs

- Interface design
    - Basic website framework established
    - Website interaction flow established
    - Team logos gathered

Next steps

- Data gathering
    - Implement data gathering from verified team accounts
- Algorithm Development
    - Improve term list by adding commonly used hashtags, and other nuanced terms used by football fans
    - Integrate player names into term list
    - Improve weighting mechanism
        - Weight multiple words within a single tweet versus a binary option of ‘football relevant terms within’ or not
    - Develop analysis for total tweet counts versus individual Sunday tweet counts
    - Explore sentiment analysis of tweets further
- Software architecting/Coding
    - Automate processing/exporting of JSON or .csv data
    - Integrate backend data with frontend visualizations

Work Percentages

	Research	Backend Coding	Frontend Coding	Presentation & Report
Andrew	75%	15%	10%	30%
Gilbert	15%	10%	75%	40%
Jacob	10%	75%	15%	30%