FIX
THEFIXING

TweetFix: Data Analytics on Match Fixing

OS WINDS
research group

Deliverables

WP5, Tasks: A26, A27, A28

Team

Team Supervisor / Coordination:
Athena Vakali, *avakali@csd.auth.gr*

Research Assistants - Developers:
Antigoni M. Founta, *founanti@csd.auth.gr*
Pavlos Gogousis, *gogopavl@csd.auth.gr*
Kostas Platis, *platiskp@csd.auth.gr*
Sofia Yfantidou, *syfantid@csd.auth.gr*

The Project

Our team created TweetFix, an online visualization platform, where users can explore the results of crowdsourced data analysis from Social Media on well-known Match Fixing cases.

Definitions

- Crowdsourcing
  The process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community.

- Data Analysis
  The collection and processing of data relative to one or more topics, in order to discover useful information and get a better understanding of these topics.

Contribution

- Capturing the social media crowd pulse regarding match fixing, per geographic location and per case;
- Analyse sentiment of the crowd per case;
- Creation of a thematic lexicon;
- Topic detection around a potential match fixing case;

Procedure

Two main tasks:

- Data Analysis
- Data Visualization

Data Analysis Task

Data Collection

- Social Media employed: Twitter, YouTube & Google+
  - Use of their APIs to collect the data
- Searching based on keywords relative to the cases
- Features
  - Twitter: ID, Username, Date, Tweet, Retweets, Favourites, Mentions, Hashtags, Location (if available)
  - YouTube: comments from videos of a specific topic
  - Google+: (for every YouTube comment) Text, Author's ID, Gender, Birthday, Location of the author (if available)

Data Preprocessing

- Text to lowercase
- Remove punctuation, numeric characters & single characters
- Remove stop words*
- Remove mentions & URLs

words with no important semantic significance such as the, to, and etc.

Cases & Datasets

- 1 general study & 4 case studies
- Cases
  - Novak Djokovic
  - Tim Donaghy
  - Southern Stars
  - Pakistani Cricket
- Final data: 191k tweets, 2674 comments, 51k users, 84k distinct words

… additional information can be found on the platform!

Data Analysis

- Total amount of Tweets, Comments, Users and Distinct Words
- Timeframe of data collection
- Total amount of Tweets per Month (*diagram*)
- Word, Hashtag and Mention frequencies (*tag clouds & tables of top-20*)
- Location frequency (*choropleth map*)
- Top active users (*bubble chart*)

62.319

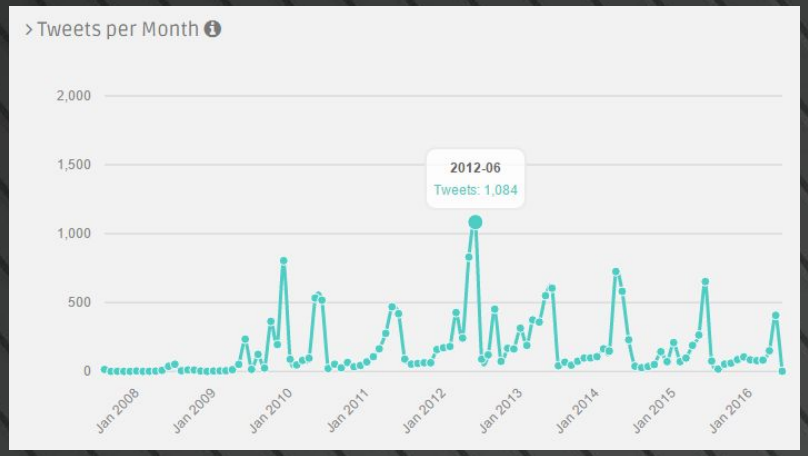
62.319 distinct tweets

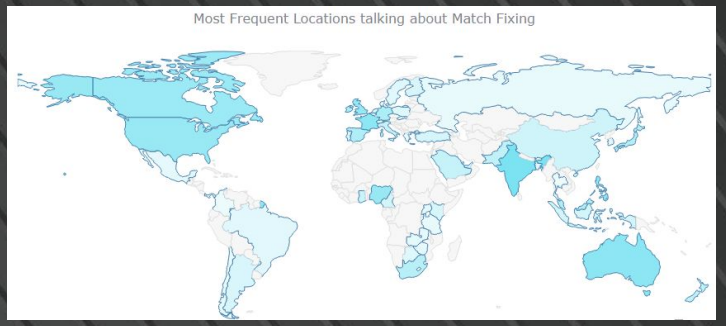
1228

34.227

53.526

01/01/2016 to 30/06/2016





› Most Frequent Words

#	Word	Frequency
1	match	52474
2	bet	36311
3	fix	32004
4	tips	6046
5	goal	5302
6	twitter	4616
7	win	3762
8	free	3629
9	today	3351
10	odds	3164

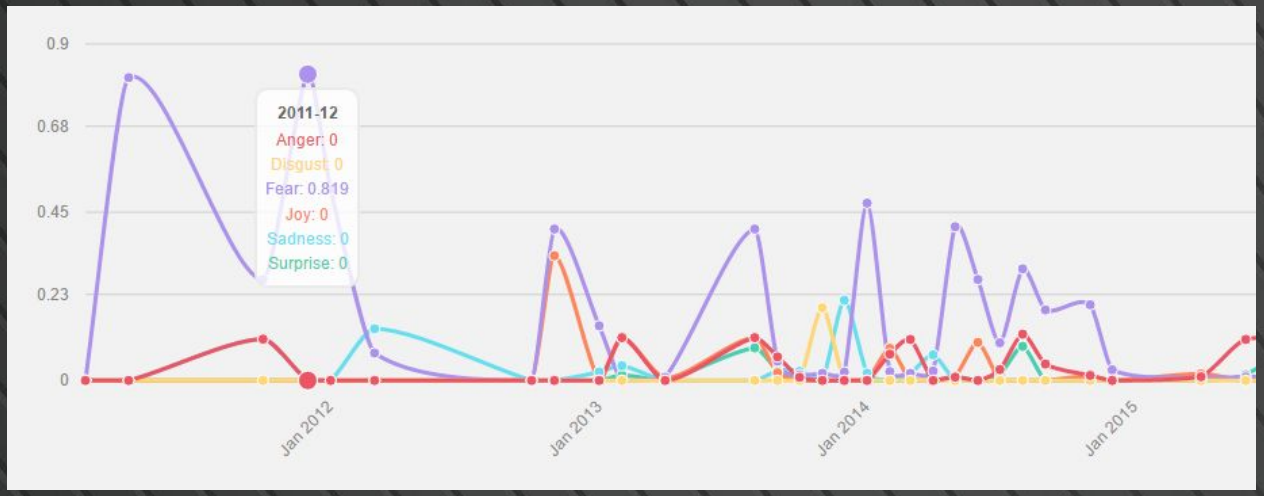
Sentiment Analysis

We perform a lexicon-based sentiment analysis in our dataset, of the six primary emotions (joy, sad, fear, anger, surprise and disgust).

"Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text. The opinions can be either binary (positive/negative) or multiple in terms of sentiments."

S.A. Results

- Sentiment per Month chart with results for all six sentiments
- Main sentiment for each case (extracted from the majority of the previous diagram)



Sentiment: Fear

The overall score of our sentiment analysis indicated the main sentiment of the users was Fear.

Data Visualization: TweetFix

TweetFix is our Data Visualization platform, where we gathered all the results from the previous analysis of the data.

- Home Page
- Cases Pages
- Lexicon Page
- About Page

Home
Cases
General
N. Djokovic
T. Donaghy
Southern Stars
Pakistani Cricket
Lexicon
About

> The Donaghy Case

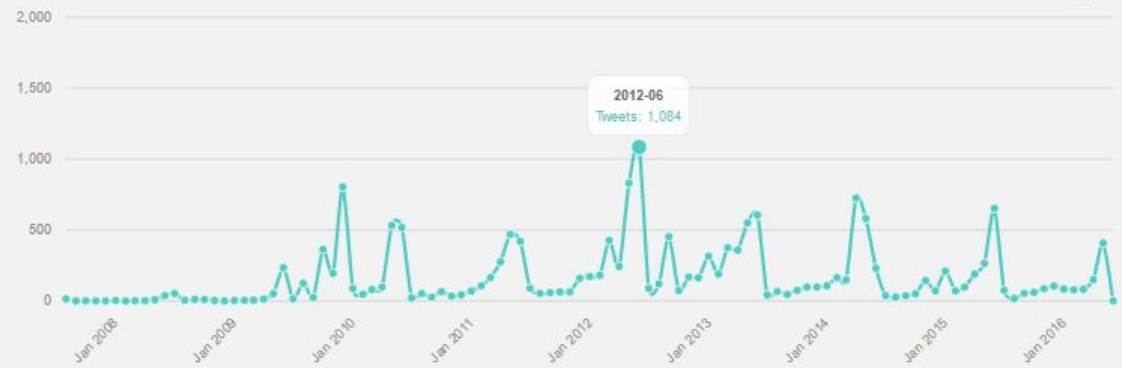
(1)

A second case study of a real-world scandal is this of Tim Donaghy, a former professional basketball referee, who worked for the National Basketball Association (NBA) for 13 seasons, from 1994 to 2007. Donaghy filed his resignation on the 9th of July, 2007. Later this year, the Federal Bureau of Investigations (FBI) published a report of an investigation on Donaghy for allegedly betting on matches that he officiated during the seasons 2005-06 and 2006-07 and making calls that affected the point spread of those games. On August 15th, 2007, Donaghy pleaded guilty to two federal charges related to the investigation, and a year later he was sentenced to 15 months in prison and three years of supervised release. The scandal provoked a storm of reactions in social media, where there was a burst of users expressing their opinions regarding Donaghy, match fixing and other related topics.

(2) Query Words: tim donaghy

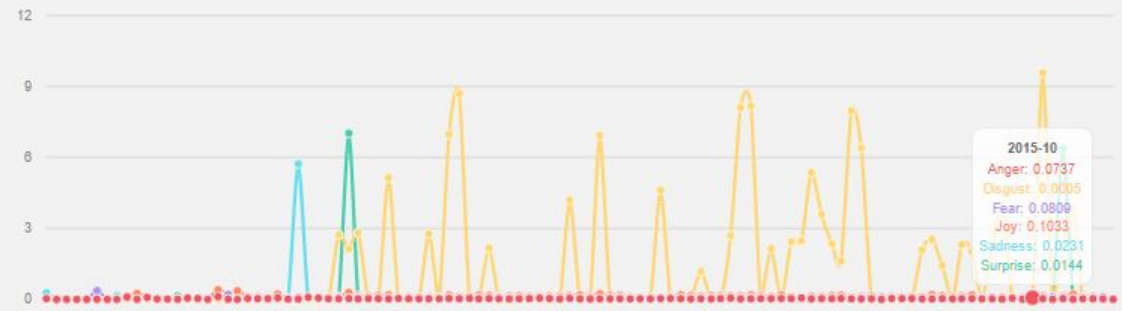
> Tweets per Month

(4)



> Sentiment per Month

(5)



> Infobox

(3)

17.344

446

13.023

21.577

15/08/2007 to 01/06/2016

Sentiment: Disgust

Useful Links

- TweetFix Home Page: http://oswinds.csd.auth.gr/tweetfix/
- Data Analysis Repository on GitHub: https://github.com/OSWINDS/FixTheFixing
- TweetFix Repository on GitHub: https://github.com/OSWINDS/TweetFix

Results

Common Patterns

1. The results of each case study are proportional to the social impact the case subject has.
2. The tweet frequency peaks are usually formed during the days of the scandal announcement and afterwards there is little or no discussion regarding the incident.
3. Most of the accounts that appear to be mainly involved in the user activity are news agencies and betting accounts.
4. There are many common terms in almost all cases. The 20 most frequent common terms between all five cases are:

match	betting	game	win	play
fixing	twitter	today	pic	corruption

Use on Educational Activities I

1. **Assessing social media users influence**: Some of the top Twitter users in most diagrams are mass-media and betting-relevant accounts, indicating that news of such scandals spread rapidly as well as that betting agencies try to take advantage of scandals.

   These results can be highlighted at an athletes' or trainers' educational task as they indicate that no scandal can be covered up (hidden) for long. Since athletes' and trainers' careers are closely related to mass media and news agencies, such an indication will increase their awareness and will impact their future choices.

Use on Educational Activities II

2. **Impact on athlete's or trainer's popularity and fame**: the Djokovic scandal doesn't seem to have affected his image, and he proceeded to an honest public declaration, while Tim Donaghy's career was stigmatised by the scandal. In the same way, in the cricketer's case, public opinion was strongly influenced about their career and their names were stigmatised with words like "corruption" and "betting". Thus, considering to accept such offers and be a part of a corrupted match could cost an athlete's career.

   These results could be used to educate athletes and trainers about the risks of taking into consideration such offers and how their image is strongly connected with their attitude, and especially their willingness to reveal any non ethical fixing case. Athletes should be aware of the consequences caused by match fixing and this research presents them in a clear way.

Conclusion

Our work can be useful to FtF, in order to:

- Support surveying based on the detected topics or in the context of the revealed thematic;
- Monitor awareness of use cases and survey participants with respect to TweetFix results;
- Prepare questionnaires with social media analytics results cross check points or remarks and validate them with ftf users;
- Utilize them in the planned educational material or in the project's potential road mapping

However, there is still lack of ground truth to evaluate our results. Could you help us?

Thank you! Questions?