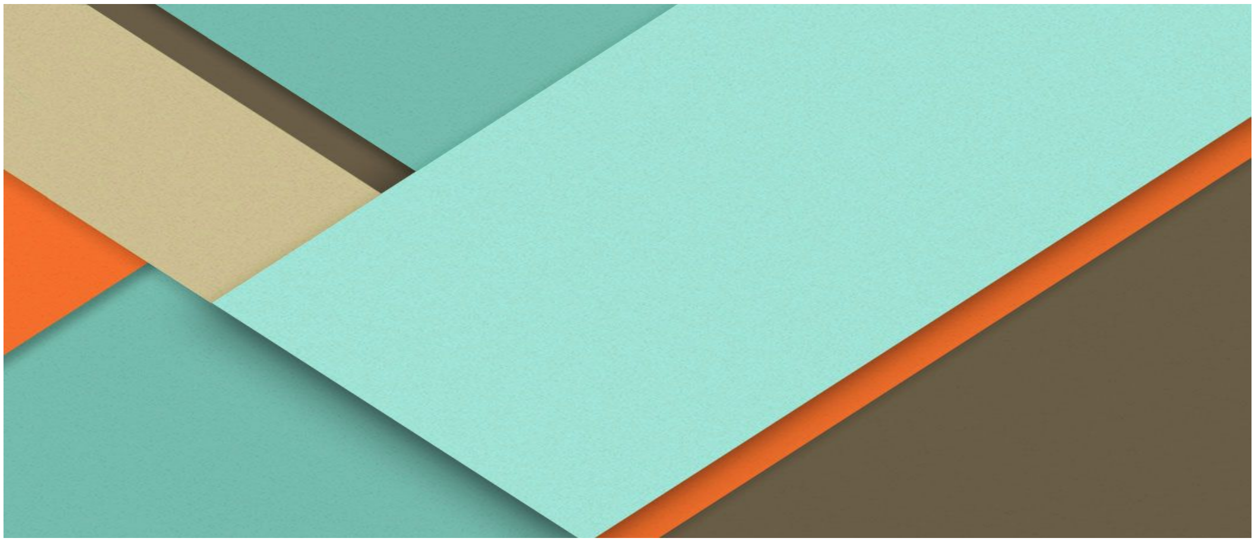


Fix the Fixing

Final Report on Tasks A26, A27, A28

01.10.2016

Team supervisor/coordination

Vakali Athena // avakali@csd.auth.gr

Research Assistants - Developers

Founta Antigoni-Maria // founanti@csd.auth.gr

Gogousis Pavlos // gogopavl@csd.auth.gr

Platis Konstantinos // platiskp@csd.auth.gr

Yfantidou Sofia // syfantid@csd.auth.gr

TABLE OF CONTENTS

INTRODUCTION

Overview

“Fix the Fixing” is an E.U. Erasmus+ project involving multiple countries and teams experts in this topic but also in information technologies and applications. It’s goal is to find, analyze and extract useful information about fixed or suspected matches of various sports and events.

Our team (Informatics Department, Aristotle University of Thessaloniki, Greece) had been assigned the task of supporting the digital side of the “fix the fixing” addressed problems, with a focus on understanding the social networks impact related to sports fixing. Therefore, an analysis on social networks data involved the processes of : collecting social media data, analyzing them, exporting useful results in readable and understandable graphs and presenting them in this report.

Idea and Motivation

Our motivation comes from the fact that, in today’s society people tend to share their opinion massively in social media, such as Facebook, Twitter and YouTube. When a scandal or important event is revealed, people share their opinion in public in order to express their anger or relief about it. Moreover, those media offer useful tools (APIs; see Appendix A) to developers and scientists who are willing to take advantage and extract valuable knowledge out of them. This process is rather useful and related with the term of “**Crowdsourcing**”.

> *“Crowdsource: to utilize information contributed by the general public (to a project), often via the Internet and without compensation”*

The idea behind the “Fix The Fixing Crowdsourcing task” was to collect data from multiple sources (Twitter and YouTube in our case) in order to extract valuable knowledge about the “Fixed Matches” issue. We have chosen two different Social Media platforms: Twitter[1], YouTube[2] and Google Plus[3]. These specific social media choices are justified by the following facts :

- Twitter users tend to share their opinion just when a scandal is announced and as a result we can analyze tweets over time; [1, 2]

[1] "Twitter." 2007. Retrieved 11 May, 2016 <https://twitter.com/>
[2] "YouTube." 2006. Retrieved 11 May, 2016 <https://www.youtube.com/>
[3] "Google+." 2011. Retrieved 11 May, 2016 <https://plus.google.com/>

- YouTube comments are attached and accompany a video’s subject (in our case scandals of fixed matches) so there is extra information offered. [3]
- Google Plus user profiles provide extra demographic information about the YouTube commenters.

As crowdsourcing techniques offer valuable results about the public opinion of people around the world, we are able to “conduct” large-scale opinion polls in order to mine people’s thoughts about fixed matches or matches being suspected to be fixed.

In our research we have collected raw data of Tweets and YouTube comments, relevant to the sports fixing problem, we have preprocessed them in order to reduce noise and finally extract useful knowledge. The derived knowledge is visualized by using multiple ways of presentation (such as tag clouds, diagrams etc.).

Challenges

During our research and results presentation we faced various challenges outlined next :

- *Linguistic issues*: Language is an important problem when it comes to crowdsourcing research, since English is the only well studied and formulated language in order to digitally process text and extract knowledge. As a result, we were limited in mining only Tweets and comments written in English.
- *Events Selection*: Events choice has also impacted our study because we were limited in selecting only events that attract English-speaking crowd (such as global events or matches in English-Speaking countries). Also, we limited our research to events that were proven or rumored to be fixed, in order to mine opinions on those matches.
- *Timespan Limitation*: Since Twitter and YouTube were created after 2007, our search about fixed matches or scandal announcements is bounded by that period.
- *YouTube Comments*: Most of the news media do not publish videos of the scandal on YouTube but on their own platforms. As a result, it was challenging to find and retrieve videos of high relevance to a particular scandal announcement and collect their comments.

CROWDSOURCING METHODOLOGY

The steps we followed to implement the crowdsourcing task as seen in Figure 1 are described below:

1. We collected data from various social media platforms. More specifically, we collected tweets from Twitter, video comments from YouTube and user information from Google+.
2. We processed the data, in order to achieve cleaner input for cleaner results. The processing included stop-word removal, punctuation removal, conversion to lowercase etc.
3. We analysed the clean data to produce the results, including statistical measurements, tag clouds and plots.

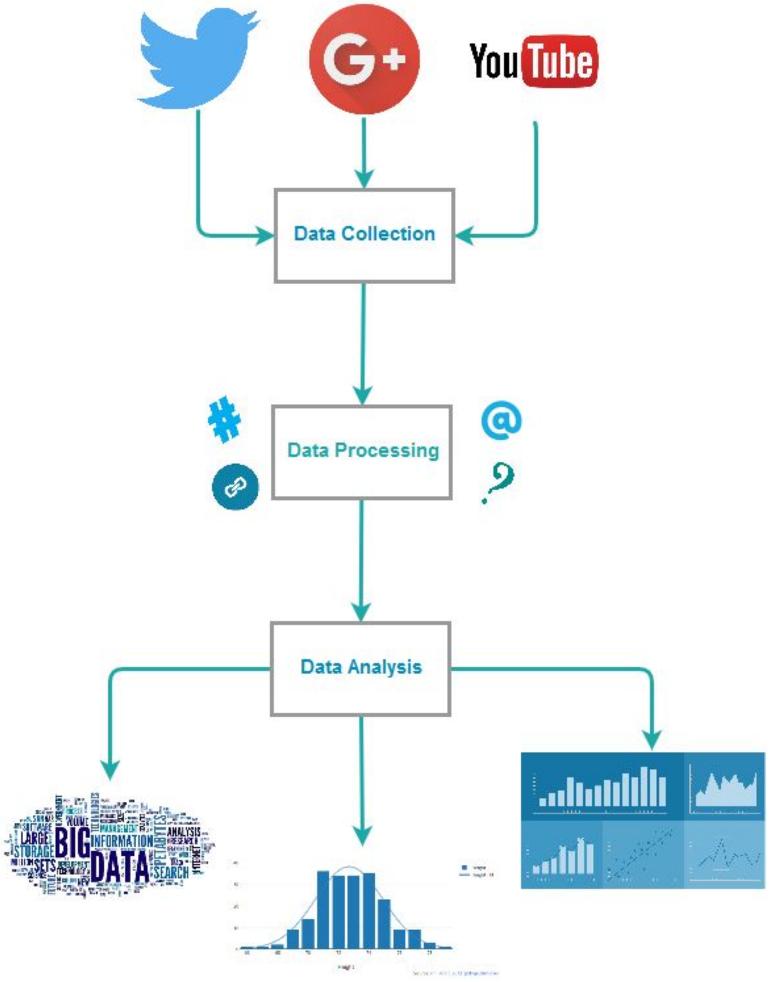


Figure 1: *The overall procedure*

Data Collection

Overview

Three social media platforms were used for the crowdsourcing related to match fixing, Twitter, YouTube and Google Plus. We tested multiple keywords, in order to fetch the most relevant results and concluded to a combination of case name, for instance Djokovic, and related keywords, such as betting, fixing, corruption, scam, gambling, etc. The data used for all of the results was obtained using the following APIs:

- Twitter REST API

  Twitter's REST API populates our Twitter database with all of the available tweets regarding a specific topic and supports tweet filtering based on various parameters, such as specific keywords, dates, hashtags and locations. We use this API, in order to collect user tweets relevant to fixed matches and events within a given timeframe.

  The following information is stored for every tweet: ID, screen name of the user, date it was posted, text, number of retweets, number of likes, mentions, hashtags, location (if available) and link to the tweet itself.

- YouTube API

  YouTube's API populates our YouTube database with comments from YouTube videos related to a specific topic. The selection of videos is done manually and the comments' collection is automated. We use this API, in order to collect the top 100 most characteristic user comments per video relevant to fixed matches and events.

- Google Plus API

  Google Plus' API populates our YouTube database along with the YouTube API. More specifically, for each comment retrieved we retrieve the authors ID as well. Then, we match this ID with the corresponding Google Plus profile and extract personal information about the author.
  The following information is stored for every comment: text, ID of the author, gender, birthday and location of the author (if available).

Data Storage

All the tweets and YouTube comments are stored in a MongoDB[4] database, which runs inside a Virtual Machine owned by the Aristotle University of Thessaloniki. For a short explanation of how MongoDB works see Appendix B.

MongoDB is a NoSQL Database Management System, which basically means that it is a document-oriented database. The main reasons we used Mongo are the following:

- Mongo is a cross-platform interoperable database;
- We were able to store all data in a JSON[5] format, so that it would be easy to use and post in the web; For a short explanation of JSON format see Appendix C.
- Mongo databases support connection with Java applications through the MongoDB-Drivers, which is quite convenient, since our data gathering mechanism was developed in Java.

Data Processing

For every tweet and comment we initiate a processing procedure to extract the "clean" parsed text. Specifically, we convert the text to lowercase, remove punctuation, single characters and stop-words, as well as mentions, numeric characters and URLs. This procedure is necessary, in order to produce qualitative results.

Data Analysis

For every collection of tweets or comments we extract multiple analytics, including:

- Total number of tweets and comments;
- Word frequency (which words are more prevalent among the tweets and comments);
- Hashtag frequency (which hashtags are more prevalent among the tweets);
- Location frequency (which cities were the comments and tweets posted from);
- Mention frequency (which users are mentioned more in the tweets).

The above are presented mostly in tag cloud formats.

Sentiment Analysis

[4] "MongoDB for GIANT Ideas | MongoDB." 2013. 9 May. 2016 <https://www.mongodb.com/>

[5] "JSON." 2003. 9 May. 2016 <http://www.json.org/>

For every collection of tweets or comments we perform a lexicon-based sentiment analysis technique, following the steps below:

1. For every one of the six basic emotions (primary emotions), which are anger, disgust, fear, joy, sadness and surprise [14], a list with their representative words is provided and used, as well as a list of related emoticons (pictorial representation of facial expressions that visualize a person's mood).
2. The list of the representative words is extended via considering their synonyms. In order to find these synonyms WordNet[6] lexical database is used. The words given in (1) along with their synonyms and the related emoticons constitute the representative words (secondary emotions) for each one of the six primary emotions.
3. Afterwards, SenticNet[7] dictionary is used, which includes a list of 30.000 words/phrases that express sentiment. In this dictionary each word/phrase is characterized based on different attributes, which contribute in better understanding the expressed sentiment. For this project only the information referred as polarity is utilized. The polarity ranges in [-1, 1], where words/phrases with negative values indicate negative sentiment, while with positive values positive sentiment.
4. Finally, the words/phrases that are included both in tweets/comments and in the SenticNet dictionary are spotted. The next step is to examine whether the spotted words are also included in the list of the representative words. Only these words are then used, in order to capture the expressed sentiment of a tweet/comment. The sentimental score for each emotion per tweet is calculated as follows:

$$\sum_{\forall emotion} \frac{\dfrac{score\ of\ an\ emotional\ word\ in\ SenticNet}{total\ number\ of\ emotional\ words\ that\ are\ related\ to\ the\ emotion\ under\ examination}}{total\ number\ of\ emotional\ words\ in\ the\ tweet\ under\ examination}$$

A sentimental word is a word that is included in the SenticNet dictionary and in the list of representative words, too, and refers to a specific emotion. In the above formula you should consider the absolute value of the value returned by the SenticNet dictionary. The above formula is calculated for each one of the six primary emotions. If the total score of a tweet for every primary emotion is equal to zero, then the tweet will be characterized as neutral.

Tools

[6] "Wordnet: A lexical database for English - Princeton University." 2015 <https://wordnet.princeton.edu/>

[7] "SenticNet - MIT Media Laboratory." 2009 <http://sentic.net/>

The tools that are used in order to implement the abovementioned approach are Java 1.8 using the IntelliJ IDE[8], the library Twitter4j[9] for downloading data from Twitter, and Github [10] in order to host the code of the whole project and facilitate the collaboration during the implementation. Also, for the data storage we used the Java MongoDB Driver[11]. Finally, for the data presentation we used Draw.io[12], an online diagram software, Plotly[13], an online analytics and data visualization tool and Tagul[14] and WordItOut[15], online tag cloud software. For the sentiment analysis part, the tools are described in the above section.

The code is available on github: https://github.com/OSWINDS/FixTheFixing. A complete Javadoc is available and can be exported from any IDE.

[8] "IntelliJ IDEA the Java IDE - JetBrains." 2006. 9 May. 2016 <https://www.jetbrains.com/idea/>
[9] "Twitter4J - A Java library for the Twitter API." 2009. 9 May. 2016 <http://twitter4j.org/>
[10] "How people build software · GitHub." 2008. 9 May. 2016 <https://github.com/>
[11] "Java MongoDB Driver." 2014. 9 May. 2016 <https://docs.mongodb.org/ecosystem/drivers/java/>
[12] "Draw.io." 2012. 9 May. 2016 <https://www.draw.io/>
[13] "Plotly | Make charts and dashboards online." 2013. 9 May. 2016 <https://plot.ly/>
[14] "Tagul - Word Cloud Art." 2014. 6 Jul. 2016 <https://tagul.com/>
[15] "WordItOut - Generate word clouds (and make custom gifts)." 2009. 9 May. 2016 <http://worditout.com/>

RESULTS-GENERAL

At first we apply the abovementioned procedure on a plain fixing dataset, in order to gather some generic information regarding match fixing, while next we analyze a series of scandals from all over the world.

Overview

All the information needed for the fixing dataset was gathered using both Twitter and YouTube. The timeframe in which data were collected from Twitter was on the past six months of 2016, more specifically from 01/01/2016 until 30/06/2016. The results include tag clouds, relating to the most frequent words and hashtags on the dataset as well as the most frequently mentioned users, some statistics, such as the amount of tweets per day, and a heat map of the locations tweeting or commenting about match fixing. Our total dataset is consisted of 62319 tweets and 1228 YouTube comments.

Analytics Summary

Below are presented the results of crowdsourcing for match fixing during 2016. We start by analysing three tables with the top 20 most frequent words, hashtags and mentions according to the dataset. All tables are also depicted in tag clouds, on Figures 2-3, 4 and 5 accordingly. On these first figures we notice that some of the words are quite predictable, such as match, fix, bet and twitter, but we also notice some words that are actually of a great value. For example, the world tennis is very high on the list, which could be interpreted as that many tennis-related fixed matches have been reported or discussed by users. An extended list of the top frequent words can be found on Appendix D. The same reasoning applies to the most frequent hashtags figure, in which we not only notice tennis being high on the list, but also other interesting words such as England and Euro 2016. This could be either due to the fact that both these terms are very popular in this period or because they are actually often reported for fixing.

The rest of the results about Fixing on 2016 are consisted of some plain statistics, such as a scatterplot with the amount of users tweeting about match fixing per day (Fig. 6), a bubble chart of the top 10 users that tweeted about the subject (Fig. 7) and a heat map showing which countries are interested the most and thus talk about fixing on Social Media (Fig. 8). We notice that the Tweets per Day plot has many minor spikes and two major ones. The first major spike appears to be on the 18th of January, the date when the Novak Djokovic scandal was raised. This is why the Djokovic scandal is the first one we will be analysing next. The top 10 users that tweeted the most about match fixing during the timeframe tested, appear to be mostly betting accounts while some of them offer tips for betting. That kind of analysis could be used to highlight possibly suspicious betting accounts that might take advantage of fixed matches. Last but not least, the heat map shows that users from all over the world are interested for this subject, with a peak at users from America, Australia and India. This is why, we also analyse two location-related case studies, one for the Australian Southern Stars and one for Pakistani Cricket.

- Query Words: match, fixing, fix, betting, bet, corruption, scam, gambling, fraud, illegal, suspicious, manipulation, integrity

Top-20 Words/Hashtags/Mentions Frequencies

Word	Frequency	Word (cont.)	Frequency (cont.)
match	52474	tennis	3004
bet	36311	pre	2607
fix	32004	pic	2548
tips	6046	play	2408
goal	5302	finish	2239
twitter	4616	corruption	2145
win	3762	live	2095
free	3629	prediction	2045
today	3351	chance	1997
odds	3164	inplay	1846

Hashtag	Frequency	Hashtag (cont.)	Frequency (cont.)
#betting	3501	#Bet365	296
#prediction	1744	#football	383
#tip	1586	#gambling	288
#1X2	1335	#tennis	246

#livescore	899	#corners	191
#inplay_betting	897	#England	313
#FSTINPLAY	554	#bwin	179
#soccerbets	411	#freebets	162
#Euro2016	360	#WT20	156
#inplay	757	#soccer	149

Mention	Frequency	Mention (cont.)	Frequency (cont.)
@YouTube	159	@Rainbow6Game	54
@bet365	134	@IPL	50
@realDonaldTrump	115	@Prodige_Betting	49
@NaseemNsm1	101	@jtemplon	46
@DavidVonderhaar	89	@WWE	46
@Treyarch	75	@Bungie	44
@FootyAccums	73	@EASPORTSFIFA	44
@ManUtd	59	@paddypower	42
@SkyBet	58	@1	41
@FootySuperTips	54	@ATVIAssist	41

Tag Clouds

Word Frequency

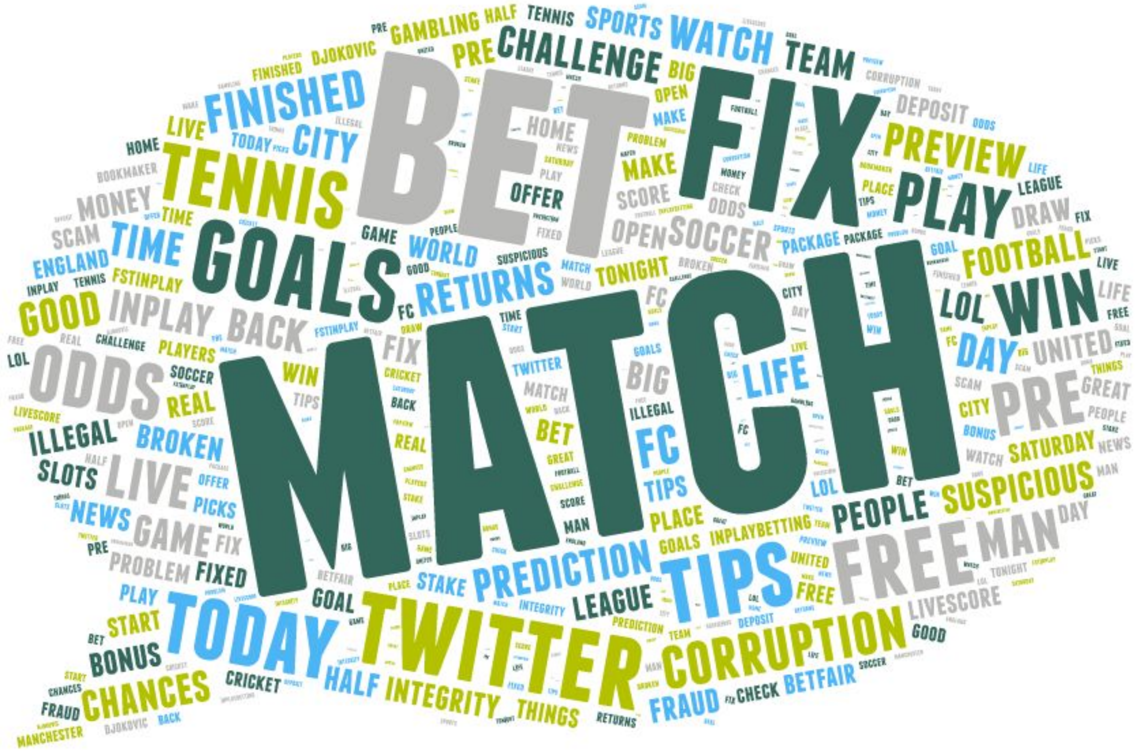


Figure 2: *Tag cloud of the most prevalent words of the dataset,* ***including*** *search terms*

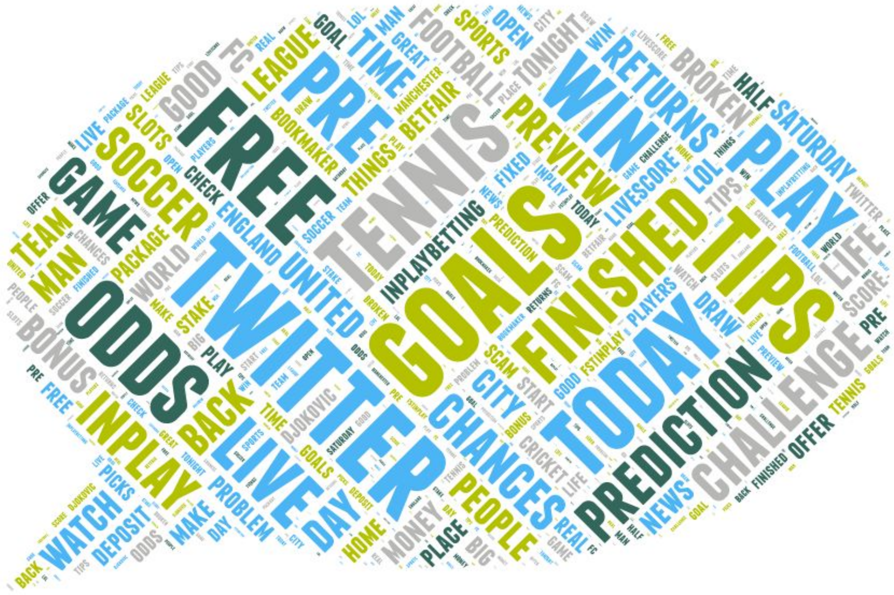


Figure 3: *Tag cloud of the most prevalent words of the dataset,* ***excluding*** *search terms*

Hashtags Frequency

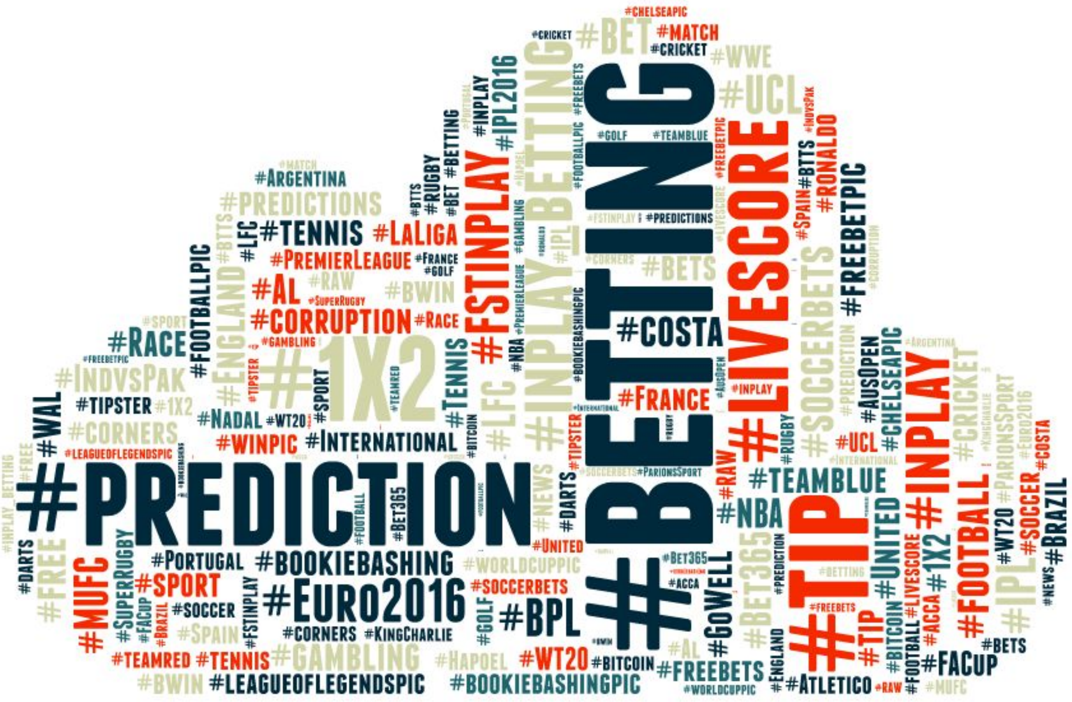


Figure 4: *Tag Cloud of most frequent hashtags in tweets*

Mention Frequency

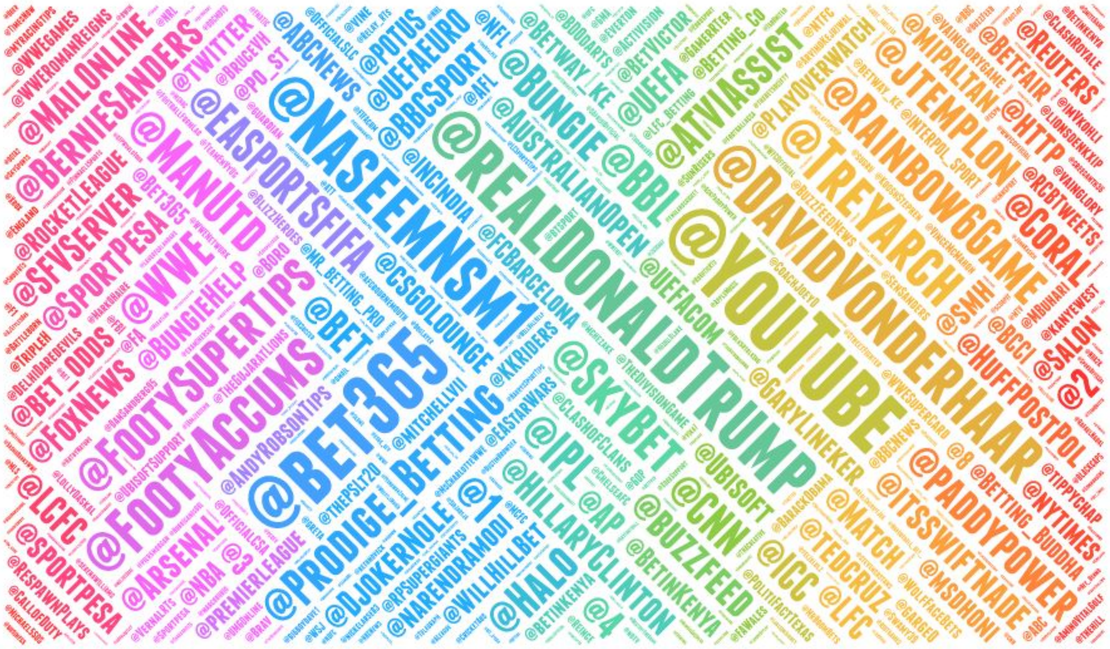


Figure 5: *Tag cloud of the most mentioned Twitter users in the tweets*

Statistics

Tweets per day

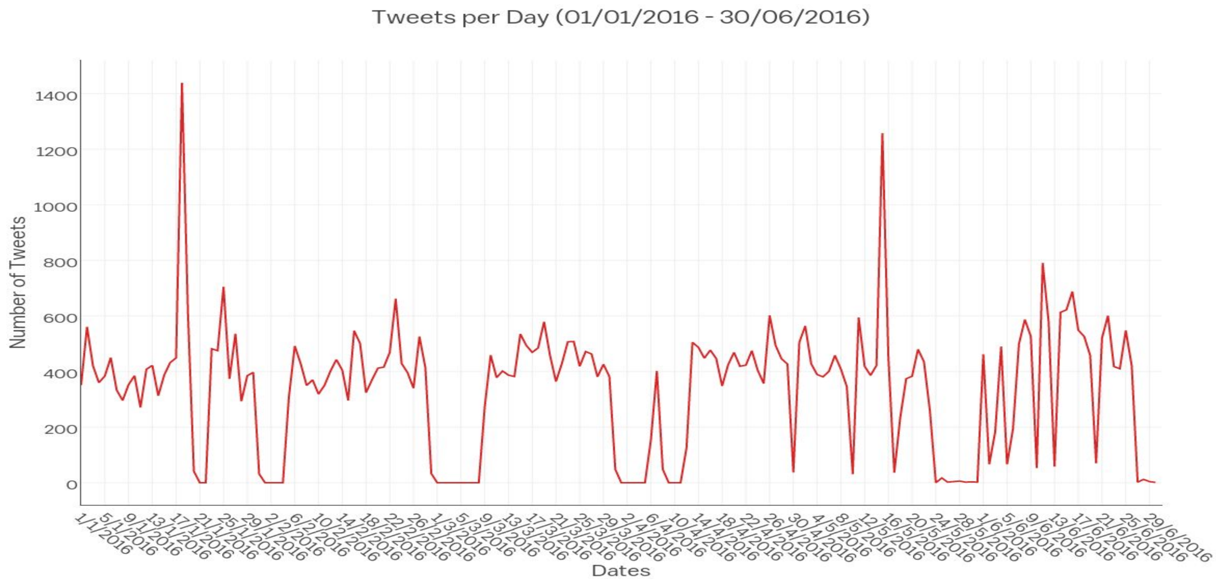


Figure 6: *Line plot representing tweets per day after Djokovic's announcement*

Most active users

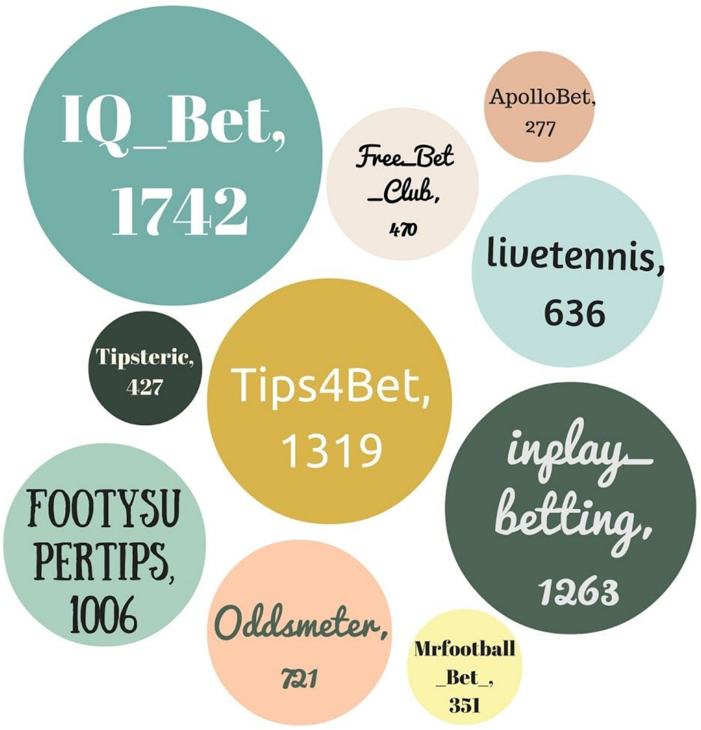


Figure 7: *Bubble chart of the top-10 Twitter users who talk about Match Fixing*

World Map of countries tweeted about the scandal

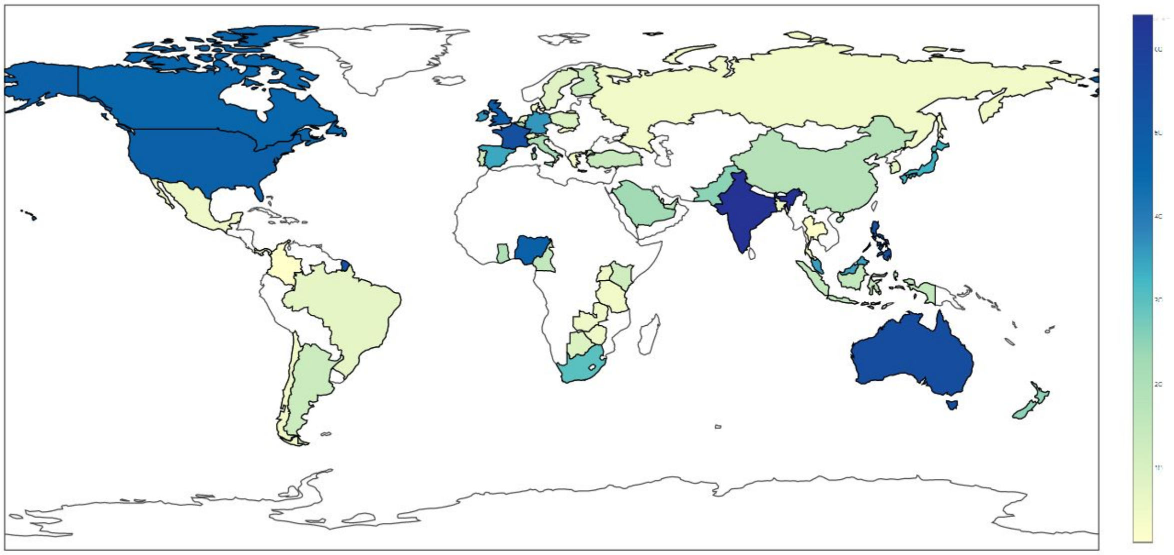


Figure 8: *Choropleth Map of countries that tweeted about match fixing, based on tweets count*

RESULTS - Case Studies

Indicative Scenario - The Djokovic Case

Overview

The procedure described above was tested on a case study of a real-world scandal including Novak Djokovic, a professional tennis player that, according to Wikipedia[16], is considered one of the greatest tennis players of all time. On January the 18th 2016, Djokovic revealed that some years ago he was approached indirectly with a £100,000 offer in order to lose a match. The revealed scandal provoked a storm of reactions from the other tennis players, news agencies as well as the social media, where there was an burst of users expressing their opinions regarding Djokovic, match fixing and other related topics.

We gathered this information using Twitter and YouTube as analysed above and extracted the most frequent words of each text (e.g. tweet, comment), as well as the most frequent hashtags, and we present both these features in tag clouds. Also, we present some statistics regarding the collected dataset. We approached the Twitter Data collection on two different ways; first gathered plain data regarding Novak Djokovic, in order to have an objective view of what is discussed about the player, and then our second query included fixing-related terms such as corruption, fixing, suspicious etc. to find out what is discussed regarding the scandal. Our total Djokovic dataset is consisted of 105,188 tweets (102,973 general Djokovic tweets and 2,215 fixing-specific), 198 comments, 1,901 distinct users and 2,639 distinct words. The time frame of the data is three and a half months, from 18/01/2016 to 30/04/2016.

Analytics Summary

Below we present the results of crowdsourcing for Novak Djokovic's case. First, we present the outcome of the specific-query search, as described above, by displaying some frequency tables about the top 20 words, hashtags and mentions according to the frequencies on the dataset, as well as some figures that better showcase some interesting results. In Figures 9-15 there are some tag clouds about the distinct words (Fig. 9-10), the hashtags (Fig. 11-12) and the twitter mentions (Fig. 13) and some statistics about the amount of tweets per day (Fig. 14) and the top 15 users that tweeted about our subject (Fig. 15). Afterwards, in Figures 16 and 17 we show some results from the plain-query search in order to compare the two cases and make some assumptions in respect to the opinion of the crowd concerning Novak Djokovic in the time frame of the scandal and whether it has been altered due to the scandal.

[16] "Novak Djokovic - Wikipedia, the free encyclopedia." 2011. 9 May. 2016 <https://en.wikipedia.org/wiki/Novak_Djokovic>

The difference of the two search cases can be shown if we compare Figure 9 and Figure 16. In Figure 9 most of the words, if not all, are related to the scandal while in Figure 16 the words are mostly related to tennis and Novak Djokovic himself. On one hand this is expected because of the query's content, and on the other hand that shows that the scandal did not affect the player's image and he did not get stigmatised from that, which can be partially explained by the nature of the scandal, that Djokovic himself brought it to light while he also stated he condemns such behaviour.

The same thing can be collaborated by the fact that although Twitter users do not mention the Djokovic fixing case long after the announcement (Fig. 14), they do mention Djokovic himself in a steady basis (Fig. 17). The peaks in Figure 17 can be explained by the various tennis matches Djokovic played at these dates during the Australian Open and are not only due to the scandal.

Both the tag clouds and the plots can be used in various ways. For instance, in Figures 9 and 10 we can mine public opinion by observing the most prevalent words. For example, we can notice the crowd's anger as words like “absurd” and “corruption” are pretty frequent in both figures. Furthermore, as we notice people tend to associate match fixing with betting as Djokovic never mentioned “betting” on his statement, but at the same time the word “betting” is frequent. As a result, we conclude that “betting” is a conclusion made by the crowd. Moreover, in Figures 11 and 12 the most prevalent hashtags used in tweets relevant to the Djokovic case can be found. These could be used to draw attention to the project's Twitter account (if available) by using them in posted tweets about match-fixing. Also, in Figure 13 we showcase the most mentioned Twitter users in the collected tweets. Among them are many news agencies e.g. Newsweek Europe, BBC Sport, France24 etc., which apparently show interest in match-fixing cases, and could be potentially used to promote the project's results. In Figure 14, we notice that the striking majority of the tweets about the Djokovic's case were posted the first 4 days after the athlete's announcement. This points out the fact that due to today's digital world's information overload, the people's interest in a topic, such as match-fixing, can be only caught for short spans of time. The project could take advantage of such time spans to promote itself possibly simultaneously with a similar match-fixing announcement. Finally, in Figure 15 we can see that most of the top users that tweeted about our case are news agencies or accounts that reproduce news, while many of the users are either tennis-related or betting-related, or both.

- **CASE I - Query Words: Djokovic, Novak, fixing, betting, corruption, scam, gambling, fraud, illegal, suspicious, manipulation, integrity**

Top-20 Words/Hashtags/Mentions Frequencies

Word	Frequency	**Word (cont.)**	Frequency (cont.)
djokovic	2214	world	216
match	1679	open	211
fixing	1315	allegations	204
novak	1230	approached	187
tennis	873	questions	166
fix	577	australian	149
offered	417	number	143
betting	278	sport	142
reveals	274	plays	138
approach	254	admits	116

Hashtag	Frequency	**Hashtag (cont.)**	Frequency (cont.)
#Djokovic	92	#prediction	12
#tennis	75	#sports	11
#AusOpen	55	#News	11
#Tennis	53	#djokovic	10
#betting	44	#fixing	9

#news	32	#tennisracket	9
#AustralianOpen	23	#novakdjokovic	9
#tipster	17	#tennisfixing	9
#Novak	16	#TennisRacket	9
#AusOpenpic	12	#economy	8

Mention	Frequency	**Mention (cont.)**	Frequency (cont.)
@TOISportsNews	7	@FRANCE24	4
@DjokerNole	7	@monachris	4
@knovak832_novak	6	@AJENews	4
@YouTube	6	@Yolitatennis	4
@ABCNews	6	@unibethttp	4
@smiley2410	6	@beastieaw	4
@NewsweekEurope	5	@LeeRock	4
@BBCSport	5	@timesofindia	3
@AustralianOpen	5	@Reuters	3
@ABC	4	@Annepappas22	3

Analytics visualization with Tag clouds

Word frequency

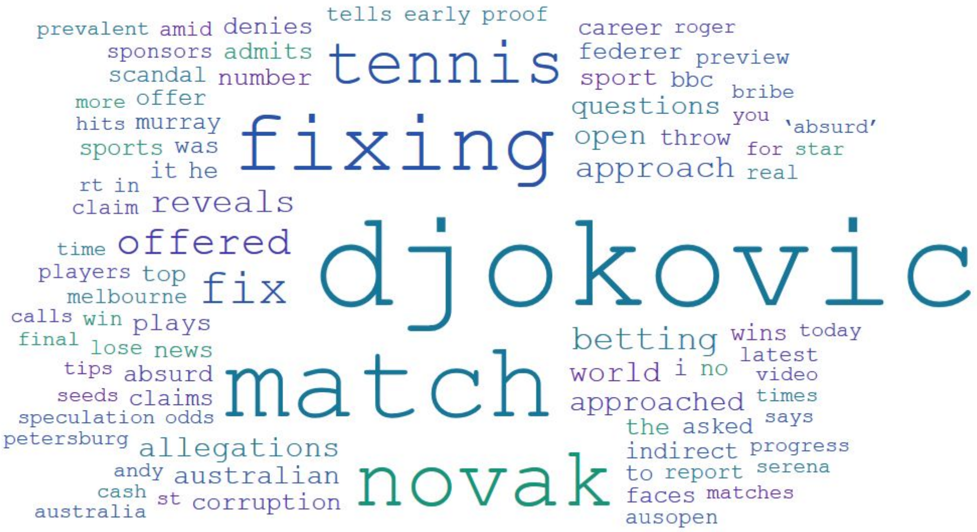


Figure 9: *Tag cloud of the most prevalent words of the tweets and YouTube comments,* ***including*** *search terms*

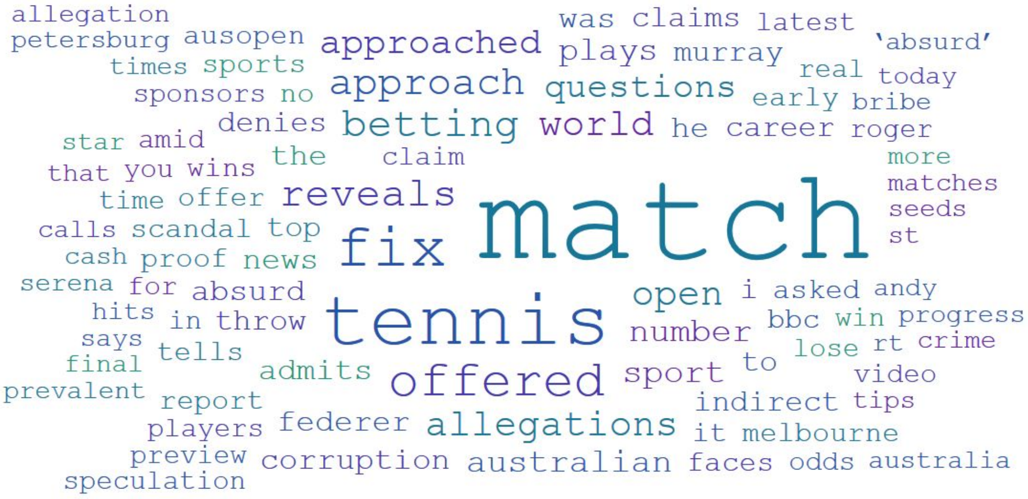


Figure 10: *Tag cloud of the most prevalent words of the tweets and YouTube comments,* ***excluding*** *search terms*

Hashtag frequency

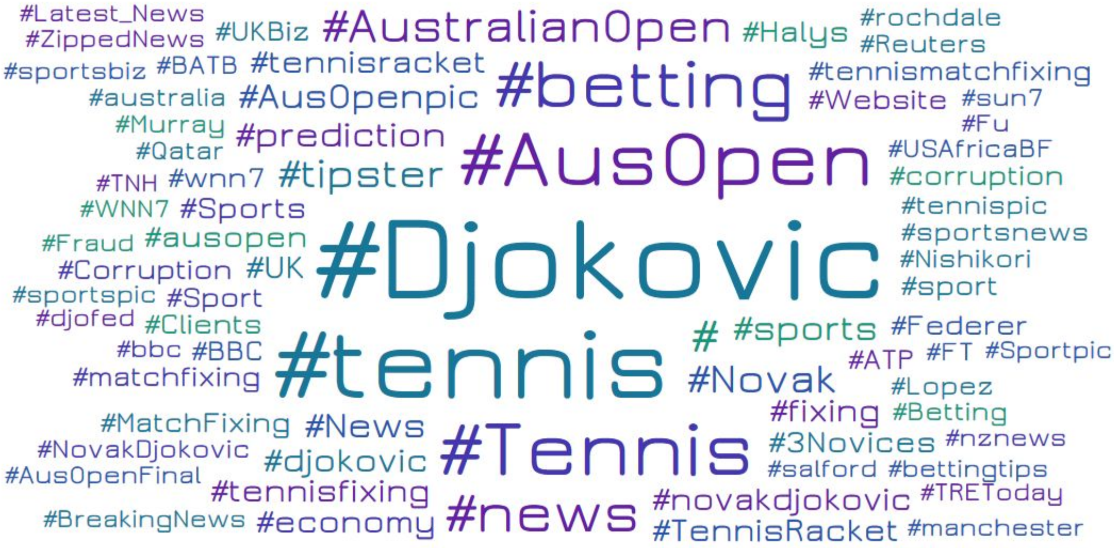


Figure 11: *Tag cloud of the most prevalent hashtags in the tweets, including search terms*

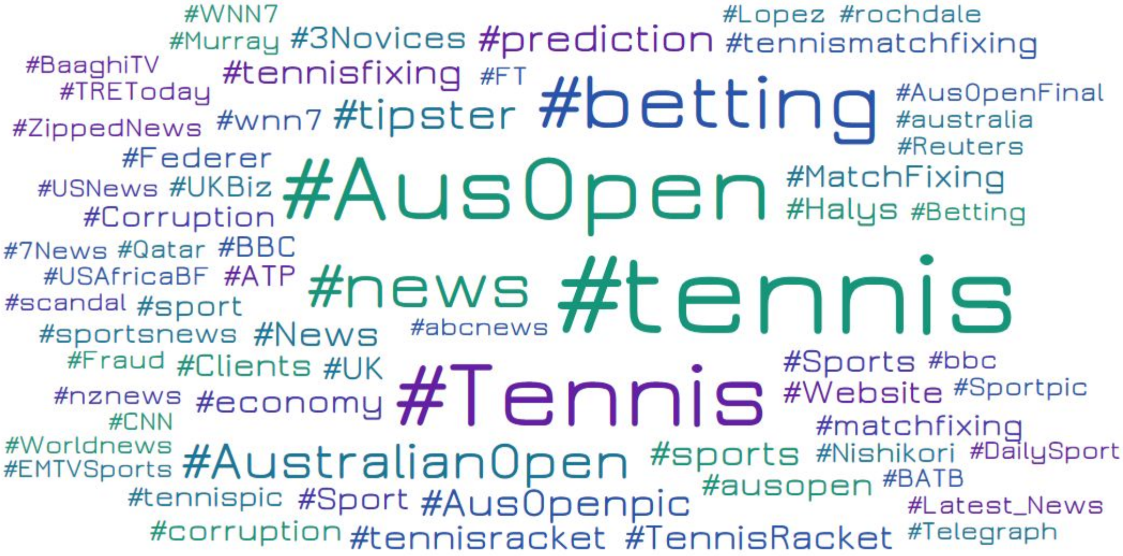


Figure 12: *Tag cloud of the most prevalent hashtags in the tweets, excluding search terms*

Mention frequency

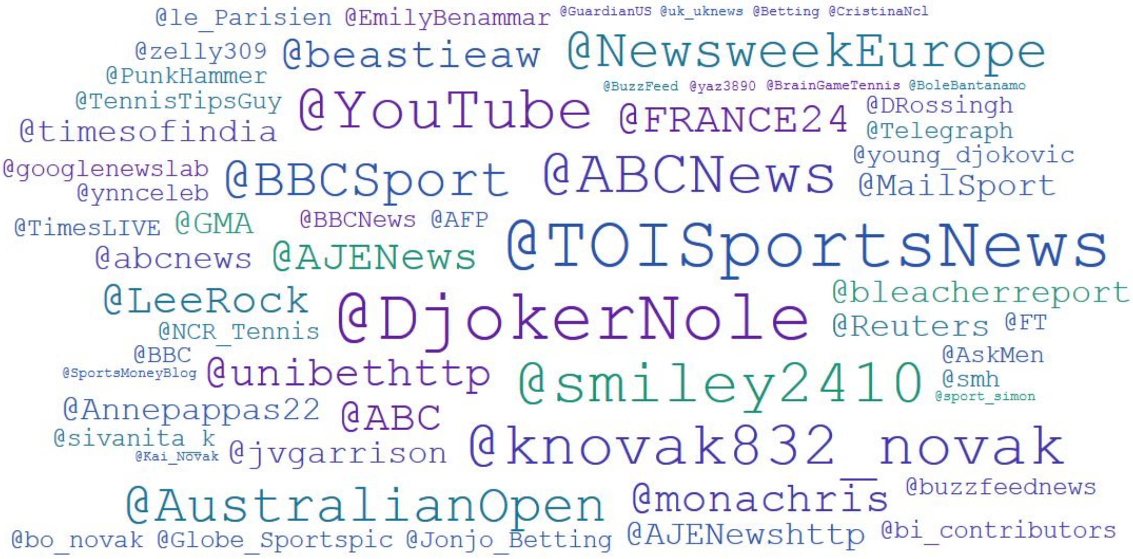


Figure 13: *Tag cloud of the most mentioned Twitter users in the tweets*

<u>Statistics</u>

Tweets per day

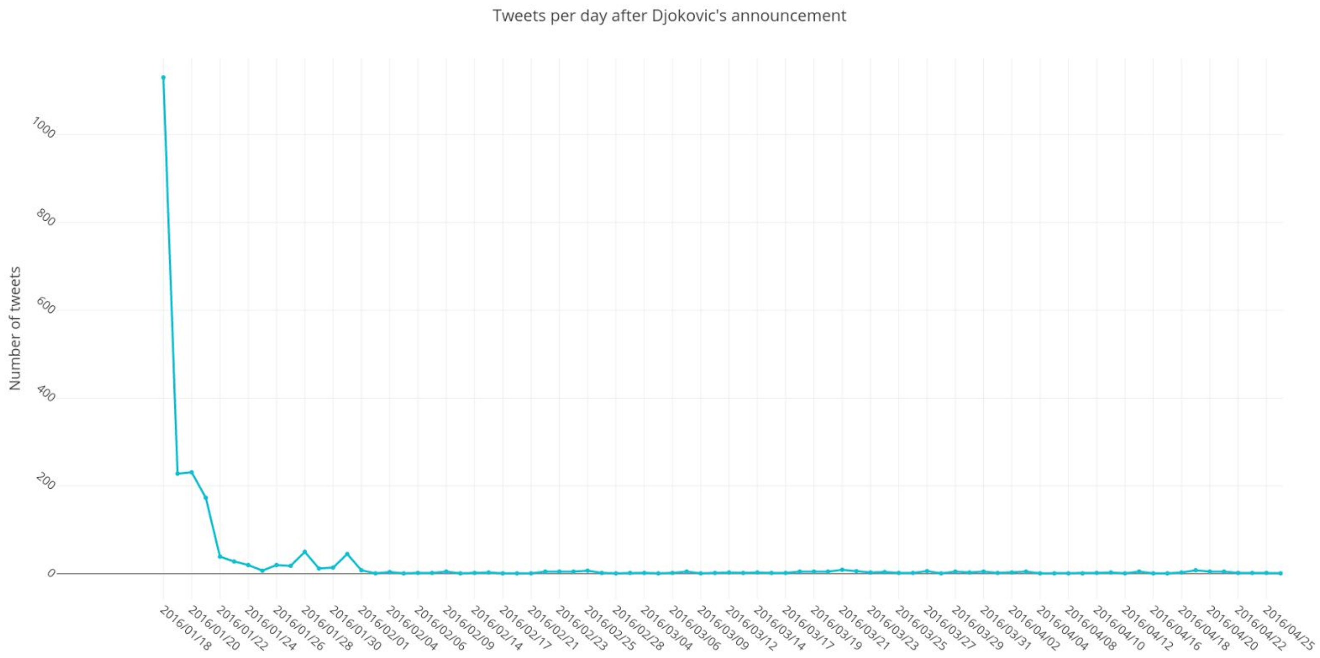


Figure 14: *Line plot representing tweets per day after Djokovic's announcement*

Most active users

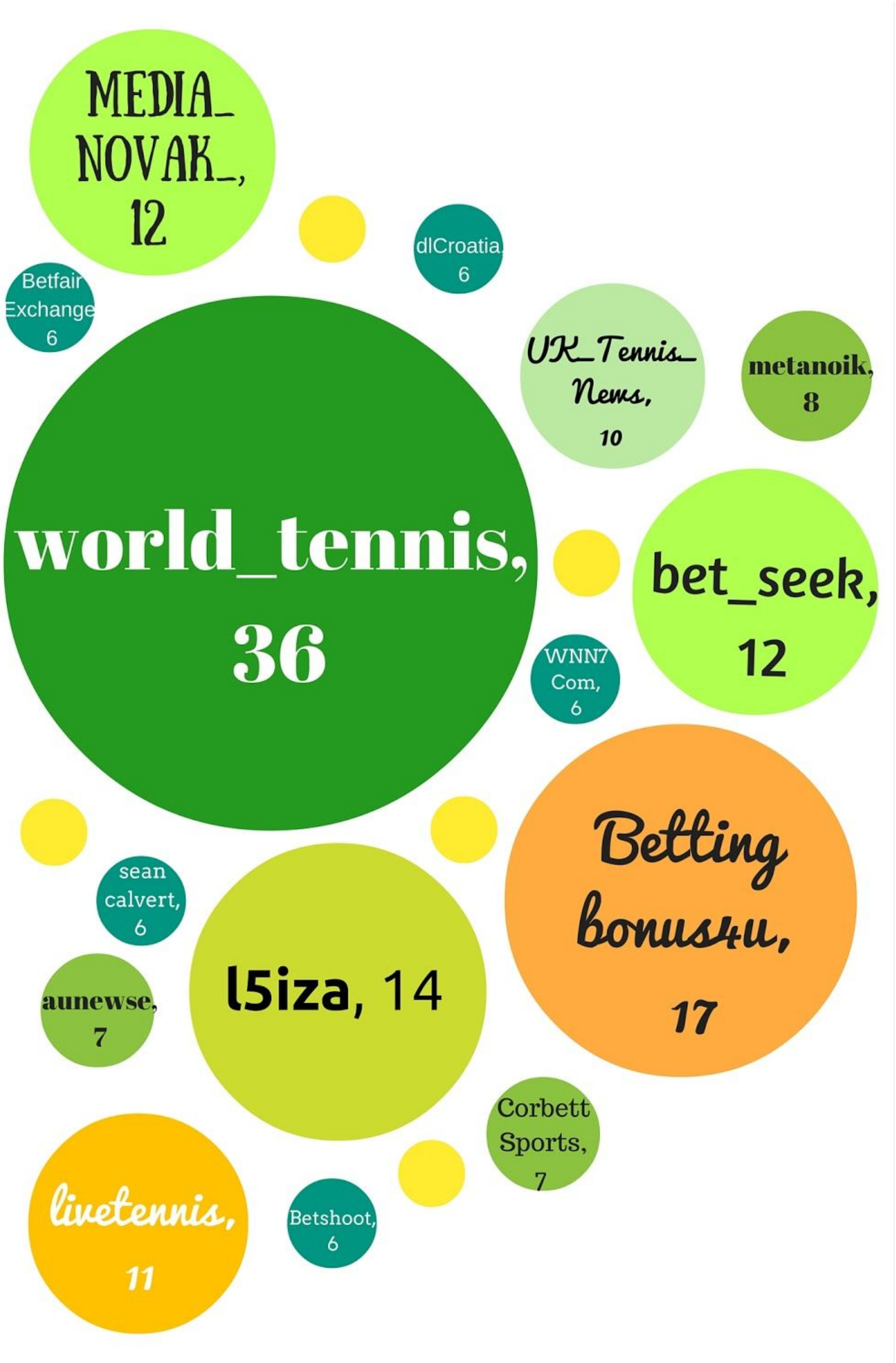


Figure 15: *Bubble chart of the top-15 Twitter users who talk about Djokovic*

● CASE II - Query Words: Djokovic, Novak

Tag clouds

Word frequency

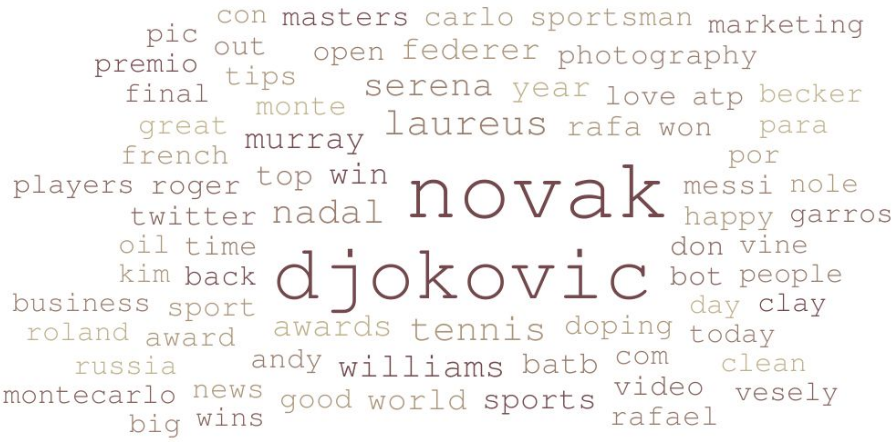


Figure 16: *Tag cloud of the most prevalent words of the tweets and YouTube comments, **including** search terms*

Statistics

Tweets per day

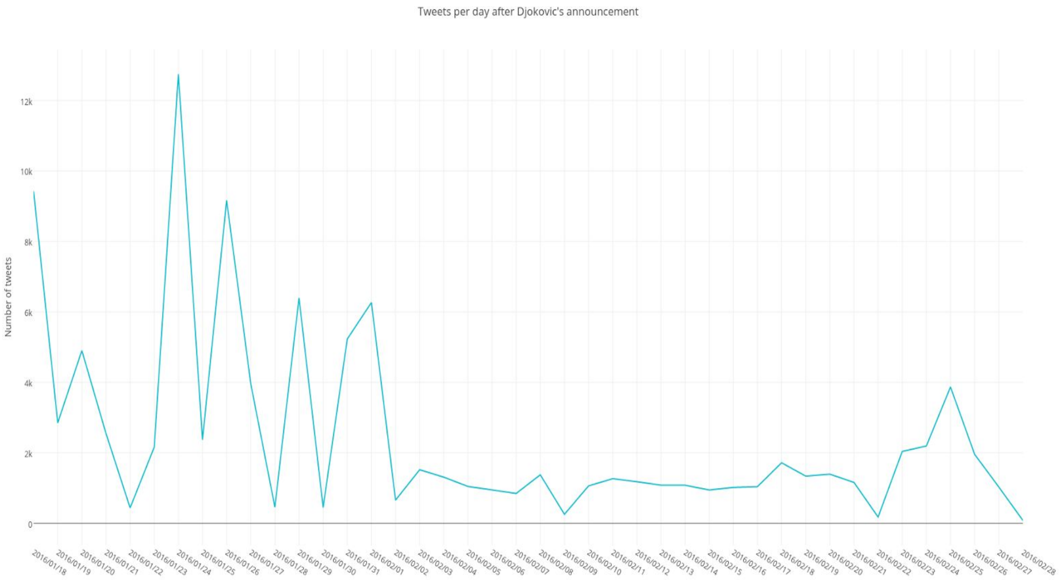


Figure 17: *Line plot representing tweets per day after Djokovic's announcement*

Basic Sentiments Scores

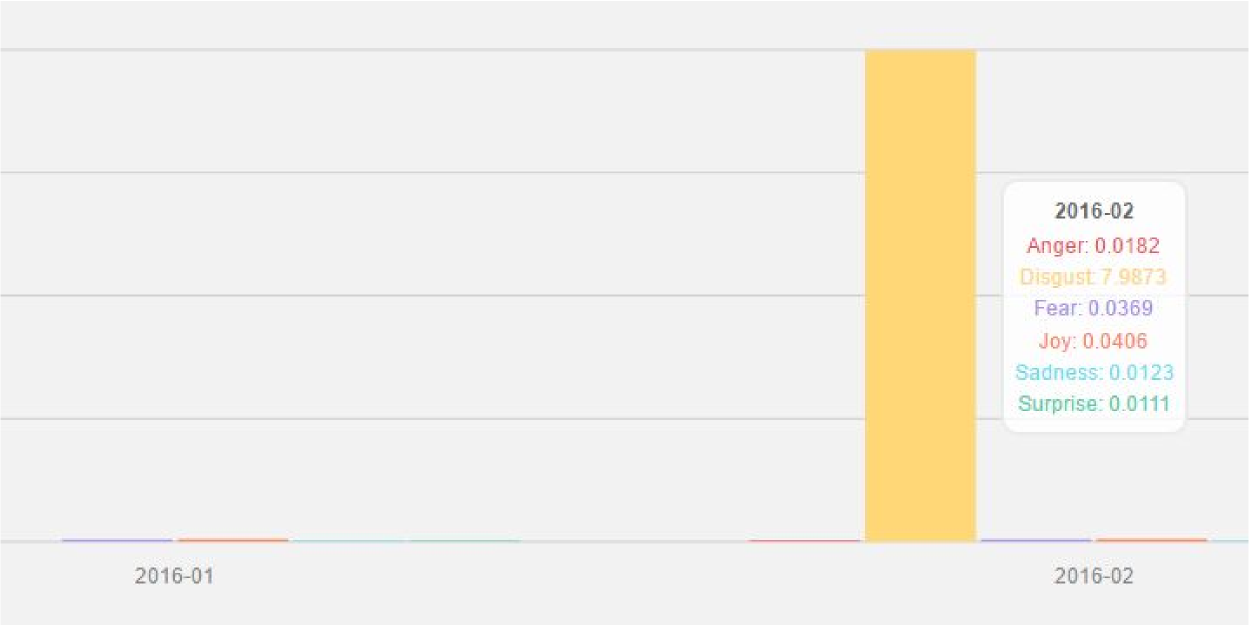


Figure 18: *Bar diagram representing levels of the six basic sentiments (Anger, Disgust, Fear, Joy, Sadness & Surprise), on each month.*

Indicative Scenario - The Donaghy Case

Overview

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.

We gathered this information using Twitter and YouTube as analysed above and performed the same analysis as with Djokovic's scandal. We approached the Twitter data collection by gathering plain data regarding Tim Donaghy after the day he pleaded guilty, in order to have an objective view of what is discussed about the former referee. Our total Donaghy dataset is consisted of 17,435 tweets, 447 comments, 13,023 (12,625 from

Twitter and 398 from YouTube) distinct users and 21,577 distinct words. The time frame of the data is 8 years and 9 and a half months, from 15/08/2007 to 01/06/2016.

Analytics Summary

First of all, it is important to mention here that, contrary to the Djokovic case, Tim Donaghy did perform match fixing and was sentenced to 18 months in prison for his crimes, which he did not reveal himself like Djokovic, but was "forced" to admit after FBI's investigation on him. Thus, differences in public opinion between Djokovic and Donaghy cases were expected and were confirmed by the crowdsourcing task. In this section we will mainly pinpoint these differences.

The graphs, plots and other visual facilitation means used to present the results of crowdsourcing for Tim Donaghy's case follow the same pattern as the ones in Novak Djokovic's case.

Starting from the tag clouds in Figures 19-23, it is already obvious that Tim Donaghy's scandal did not hurt only his reputation as a referee, but tamed NBA's reputation and decreased its integrity as a whole, as "NBA" is one of the most frequent words used in users' tweets and comments, and NBA teams like "Raptors", "Lakers", etc. are used quite frequently as well, doubting the integrity of their matches' officiating. In addition, words like "gang", "mob", "gambling", etc. are commonly used, indicating a connection between match fixing and more shady dealings.

In Figure 23, we can notice that amongst the most mentioned twitter users are again some news agencies, including, but not limited to, @CBCBoston, @NBATV, @SLAMonline and @NYMag, which are different from the ones elicited for the Djokovic case. This information indicates that different media channels and agencies should be used for different sports and/or countries, in order to achieve maximum publicity.

However, the most interesting results arise from the plot in Figure 24, representing the number of tweets per day after Donaghy pleading guilty to two federal charges on August 15th, 2007. In Djokovic case there was a spike in tweets that lasted only for a few days after the announcement, thus we can conclude that the match fixing announcement has not hurt the tennis player's reputation in the long run. This is not true though for Tim Donaghy. In Donaghy's case plot we notice that there are multiple spikes throughout the course of the 8 years following the scandal, proving the irreparable damage the latter caused to his career. For instance, on December 7th, 2009, the spike in tweets is caused by a Tim Donaghy interview with "60 minutes", talking about match fixing in NBA and his recently released book. Also, the spike on September 25th, 2012, is caused by Tim Donaghy questioning the integrity of NFL's replacement referees during an interview. These two cases prove how interwoven Donaghy's name is with match fixing. From the tweets' text on the specific dates one can understand that users either condemn the trust placed on Donaghy for commenting on match fixing issues or agree with his points, presenting him as a "match fixing guru", but at the same time making him infamous in the field on basketball officiating. These results can be used to emphasize that this is not the kind of fame an athlete, coach, referee, etc. wants surrounding one's name when it comes to one's career. In addition, other spikes occurred during NBA suspicious matches that were officiated by other NBA referees (29/05/2012: Heats vs Celtics suspicious officiating, 29/05/2013: Heats vs Pacers suspicious officiating, 14/05/2014: Thunder vs Clippers suspicious officiating), where Tim Donaghy's name was used ironically, humorously at times or as a synonym of match fixing in the tweets' text. All the above can be used to showcase that match fixing accusations can permanently harm one's career or even end it (Tim Donaghy has never officiated a big match since the scandal broke) and tame one's name.

- **Query Words: Tim, Donaghy**

Top-20 Words/Hashtags/Mentions Frequencies

Word	Frequency	Word (cont.)	Frequency (cont.)
donaghy	17472	officiating	907
tim	17298	book	788
nba	6272	timdonaghy	759
game	3213	call	757
ref	2314	crawford	620
refs	2171	stern	605
bit	1470	reffing	549
referee	1360	sports	543
reveals	274	twitter	537
former	1150	foul	534

Hashtag	Frequency	Hashtag (cont.)	Frequency (cont.)

#NBA	596	#celtics	52
#TimDonaghy	411	#Raptors	44
#timdonaghy	287	#RTZ	44
#nba	209	#fixed	40
#NBAFinals	105	#Lakers	39
#NBAPlayoffs	92	#Basketball	37
#Celtics	79	#heat	34
#NFL	74	#ebook	34
#rigged	63	#Knicks	34
#Heat	63	#Bulls	31

Mention	Frequency	Mention (cont.)	Frequency (cont.)
@NBA	419	@jimrome	27
@nba	106	@deadspin	27
@Deadspin	75	@TimDonaghy	26
@sportsguy33	63	@youtube	26
@Tim_Donaghy	60	@SportsCenter	25
@espn	41	@TheCWW	24
@NBAOfficial	38	@mcuban	23
@Youtube	32	@Raptors	21
@nfl	30	@dpshow	21

@BillSimmons	27	@NBAonTNT	20

Tag Clouds

Word Frequency

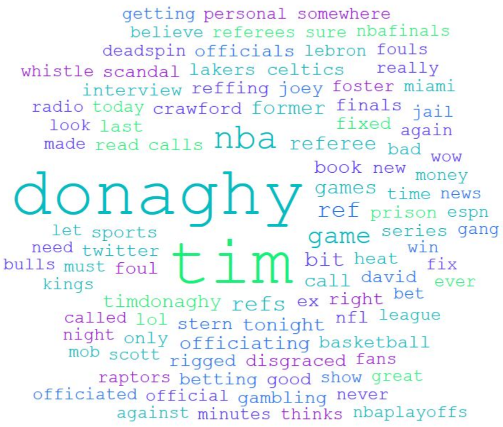


Figure 19: *Tag cloud of the most prevalent words of the tweets and YouTube comments, **including** search terms*

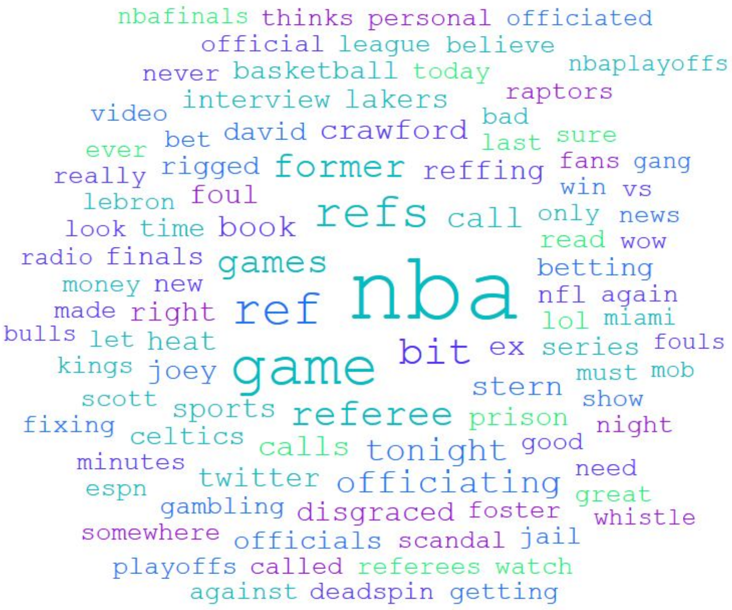


Figure 20: *Tag cloud of the most prevalent words of the tweets and YouTube comments, **excluding** search terms*

Hashtag frequency

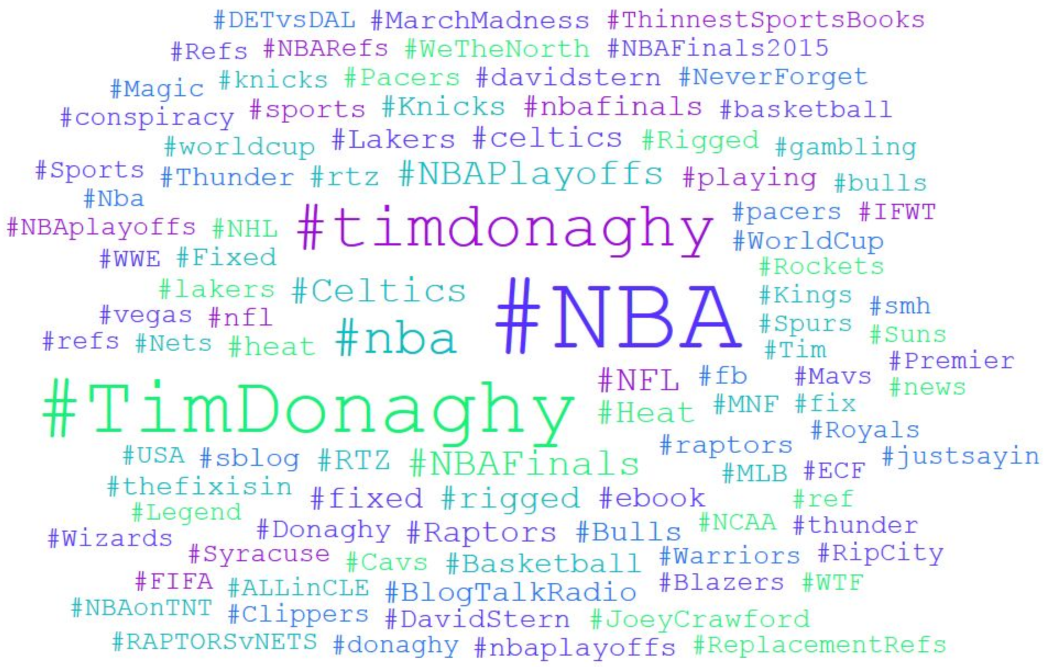


Figure 21: *Tag cloud of the most prevalent hashtags of the tweets, including search terms*

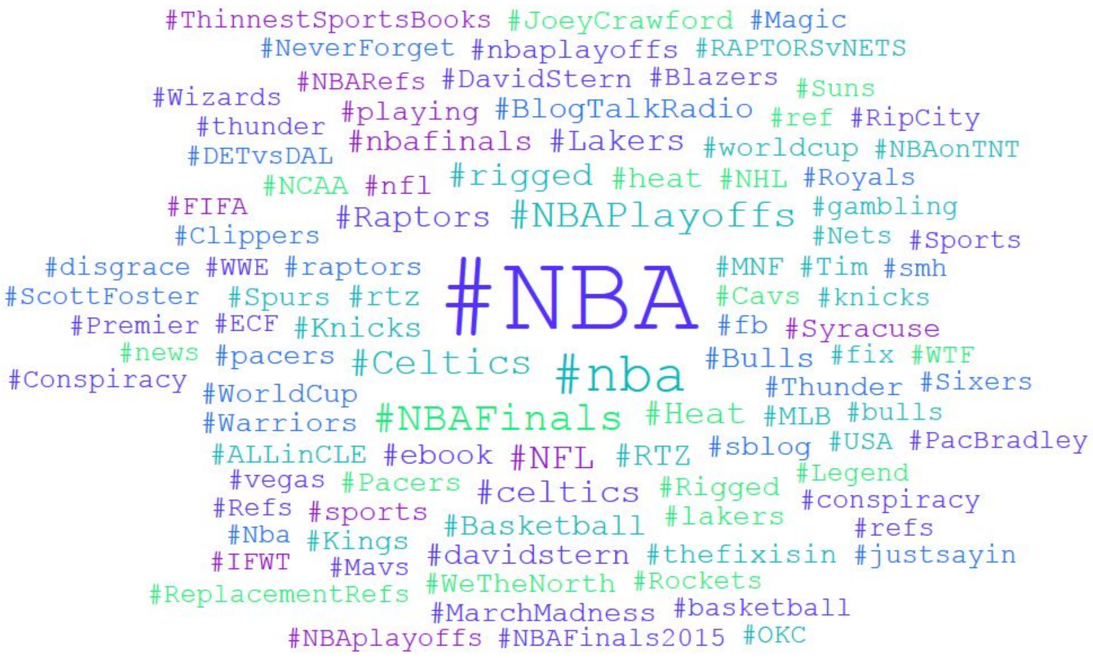


Figure 22: *Tag cloud of the most prevalent hashtags of the tweets, excluding search terms*

Mention frequency

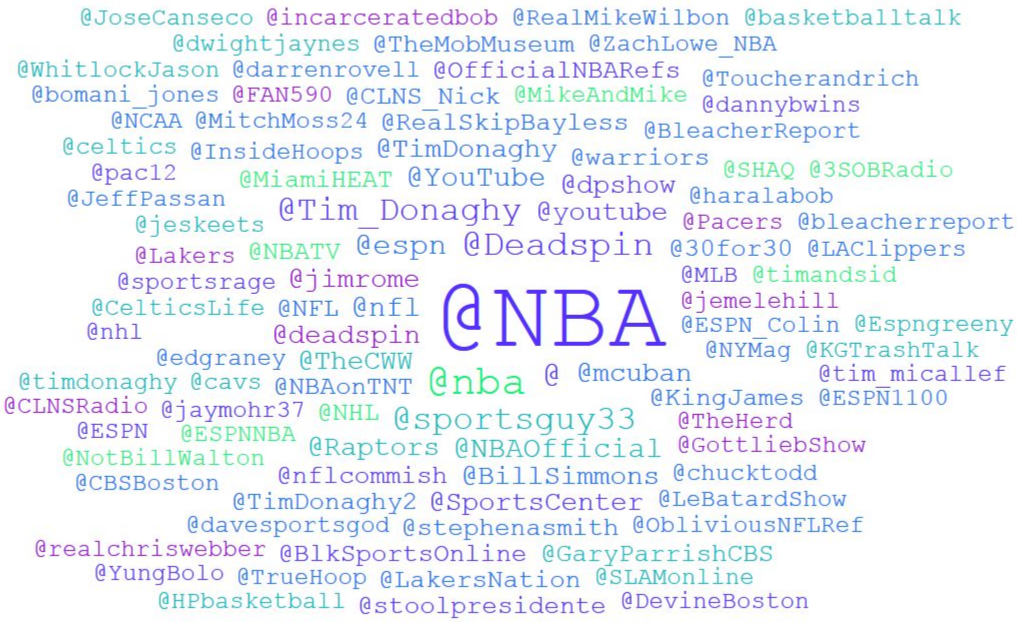


Figure 23: *Tag cloud of the most prevalent mentions of the tweets, including search terms*

Statistics

Tweets per day

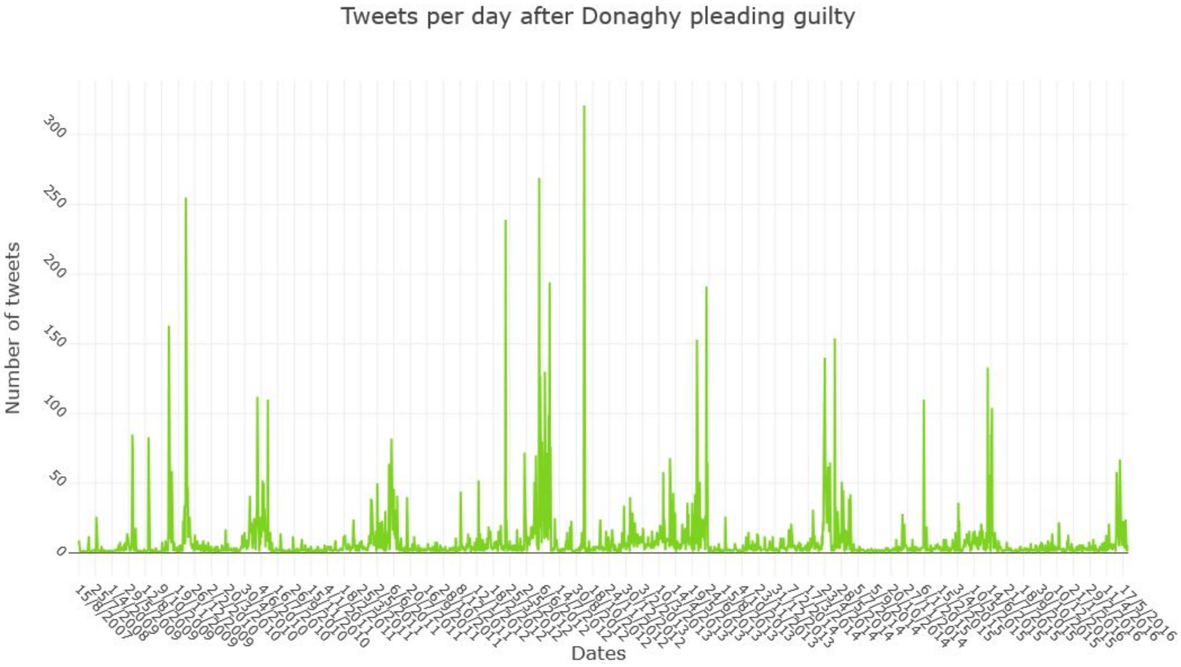


Figure 24: *Line plot representing tweets per day after Donaghy's announcement*

Most active users

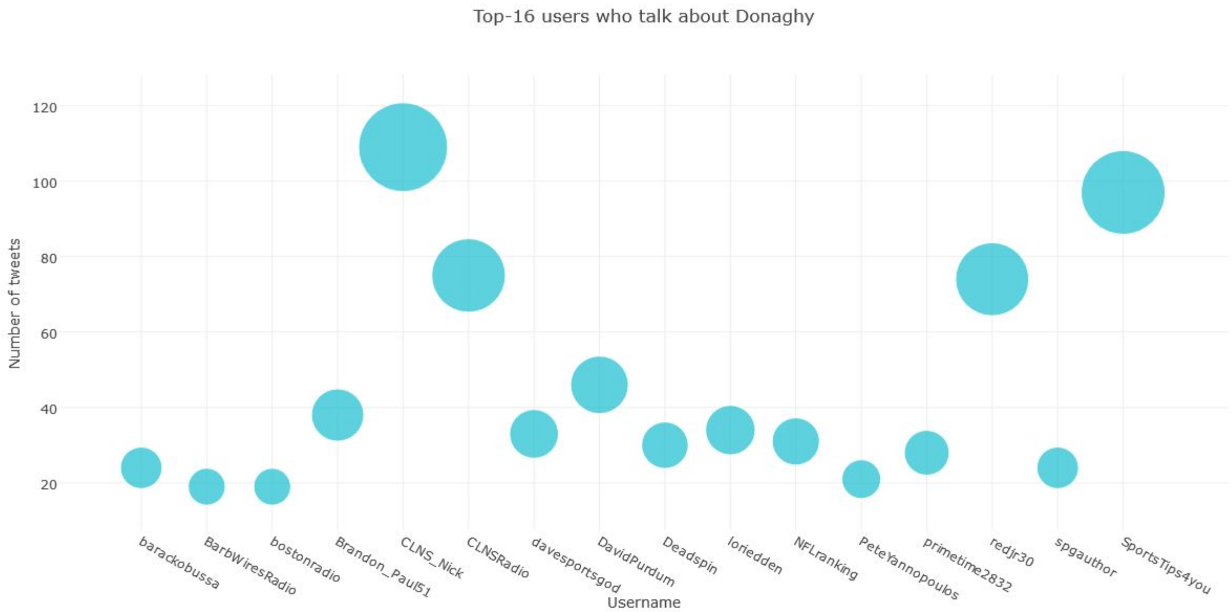


Figure 25: *Bubble chart of the top-16 Twitter users who talk about Donaghy*

Basic Sentiments Scores

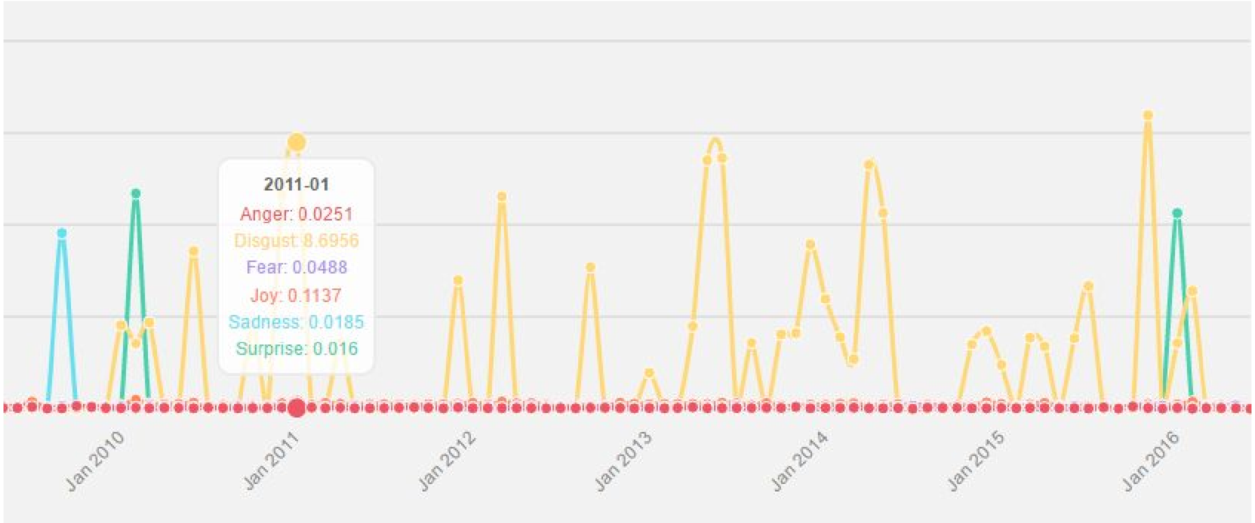


Figure 26: *Line plot representing levels of the six basic sentiments (Anger, Disgust, Fear, Joy, Sadness & Surprise), on each day and month.*

Indicative Scenario - Australia's Southern Stars Case

Overview

Another study case we examined regarded Australia's biggest match-fixing scandal. The scandal was hatched between Hastings, Sussex and Singapore in late 2012 and was blown apart in September 2013, when 10 people were arrested for alleged involvement in throwing games.

We retrieved data only from Twitter, due to the lack of Southern Star videos on YouTube. The tweets range from July 2, 2011 to May 7, 2016. We visualized the retrieved information in tag clouds by all tweets' word frequencies, user frequencies, mention frequencies, hashtag frequencies, and we plotted the number of tweets posted per day (from the first day a relevant tweet was posted, up to the most recent one). Our collection consists of 204 distinct tweets, 197 users and 72 mentions.

Analytics Summary

We start by presenting the top twenty words that appeared, based on the frequency of their appearances. It is quite obvious that among the first words, one would expect "southern" and "stars" to show up at the top of the list.

Besides those two, however, we find "match", "fixing", "betting" and "scandal" to frequently occur after the team name, which indicates that fixing allegations were a very popular topic during the search dates.

Similar terms show up further down in the list, such as "fix", "alleged" and "arrested". Moving on to the top twenty hashtag frequencies table, we can tell that hashtags like "#matchfixing", "#betting" and "#integrity" are of higher interest to us. Regarding the top twenty mentions table, most tweets are sent directly to, or at least mention, the "@SouthernStars" official account. More words and terms can be spotted at the terms', hashtags' and mentions' tag clouds.

Lastly, the tweets per day line plot clearly demonstrates that the highest activity took place between the dates 6/8/2013 and 22/9/2013, during the time that they admitted to conspiring to fix games against the Oakleigh Cannons.

- **Query Words: Southern, Stars, fix, fixing, betting, corruption, scam, gambling, fraud, illegal, suspicious, manipulation, integrity**

Top-20 Words/Hashtags/Mentions Frequencies

Word	Frequency	**Word (cont.)**	Frequency (cont.)
stars	177	soccer	19
southern	175	alleged	17
fixing	137	club	15
match	125	team	14
betting	39	arrested	13
scandal	34	australia	13
players	33	southernstars	13
vpl	29	fc	12
football	22	coach	11
fix	21	league	11

Hashtag	Frequency	**Hashtag (cont.)**	Frequency (cont.)
#VPL	15	#australia	2
#SouthernStars	7	#Victoria	2
#southernstars	5	#betting	2
#vpl	5	#VictoriaCrimesAct	2
#matchfixing	4	#integrity	2

#HorseRacing	4	#ALeague	2
#Australia	4	#FFA	2
#football	4	#SportsLaw	2
#Football	3	#BETTING	2
#VictorianPremierLeague	2	#News	1

Mention	Frequency	**Mention (cont.)**	Frequency (cont.)
@SouthernStars	10	@bad_boy_six	1
@_SouthernStars	6	@BCCI	1
@abcnews	4	@WormsleyCricket	1
@tennewsmelb	3	@foxfootball	1
@theage	3	@FIFAcom	1
@westindies	3	@dannyahh	1
@2ser	2	@NickMetallinos	1
@2	2	@FinancialReview	1
@FFV_VPL	2	@smh	1
@7NewsMelbourne	1	@joshcalle13	1

Tag clouds

Word frequency

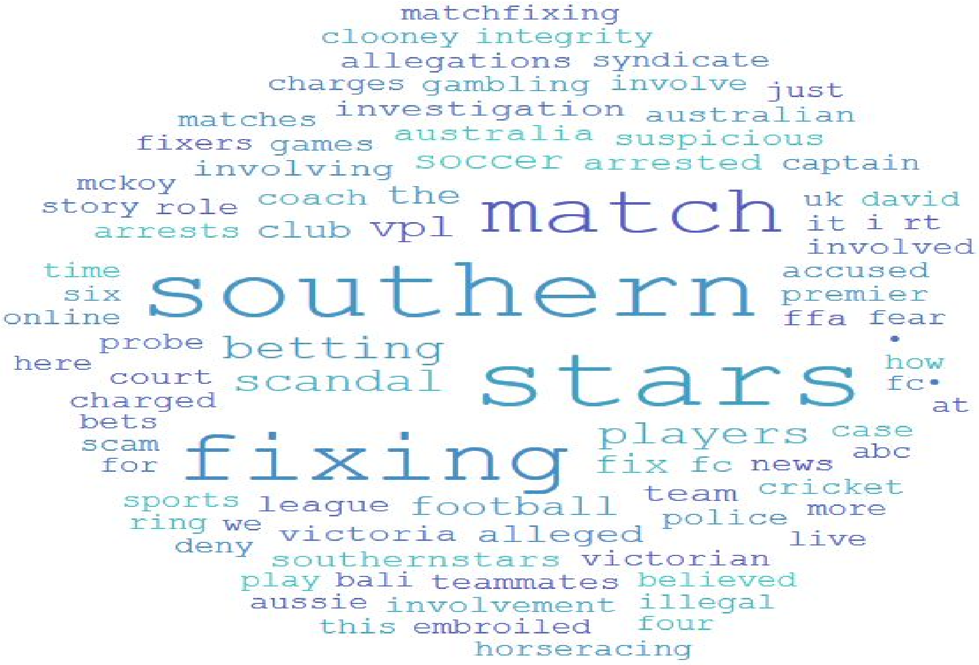


Figure 27: *Tag cloud of the most prevalent words of the tweets,* ***including*** *search terms*

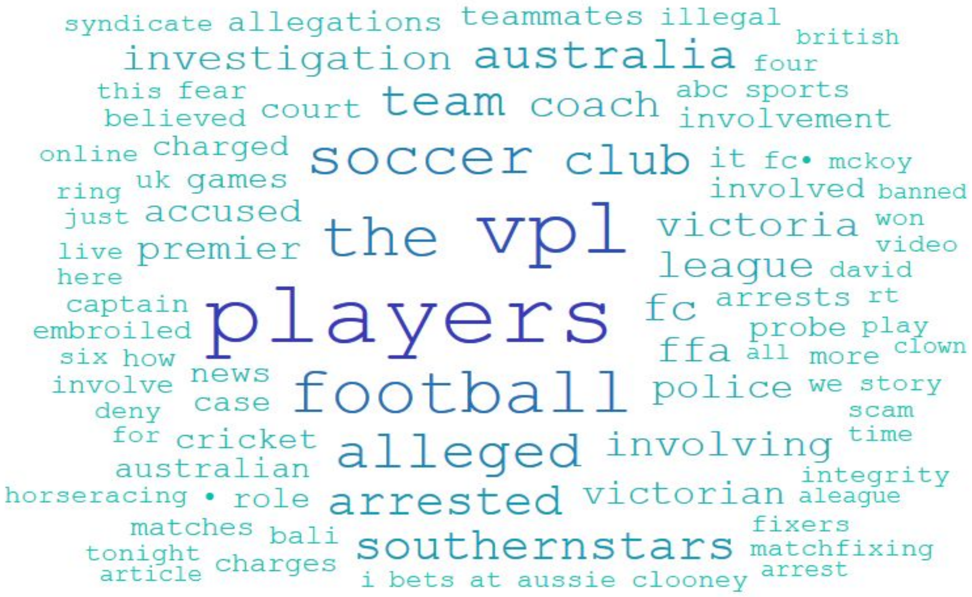


Figure 28: *Tag cloud of the most prevalent words of the tweets,* ***excluding*** *search terms*

Hashtag frequency

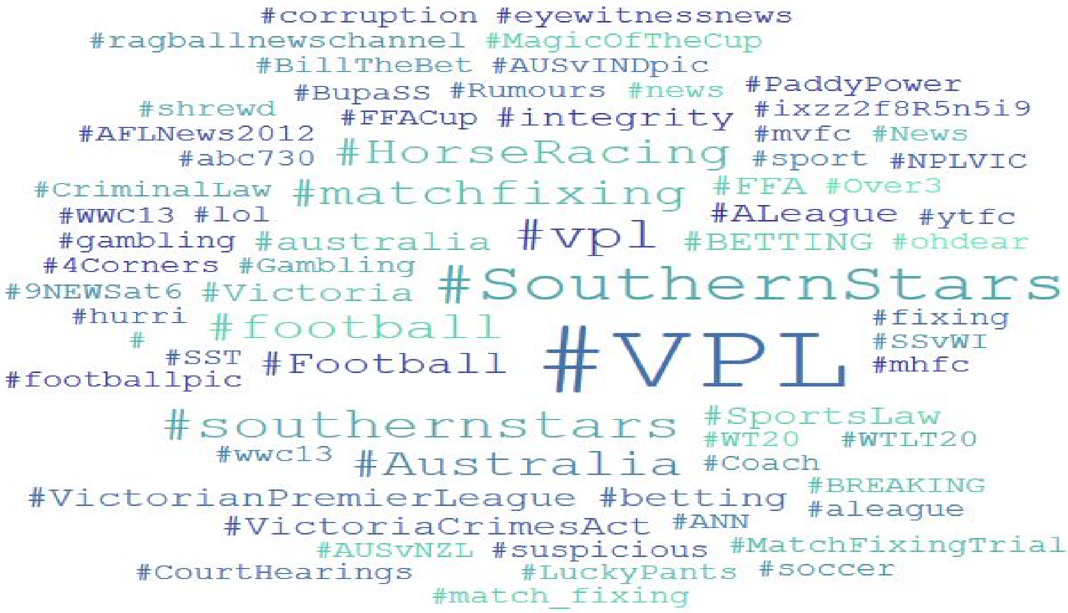


Figure 29: *Tag cloud of the most prevalent hashtags of the tweets, including search terms*

Mention frequency

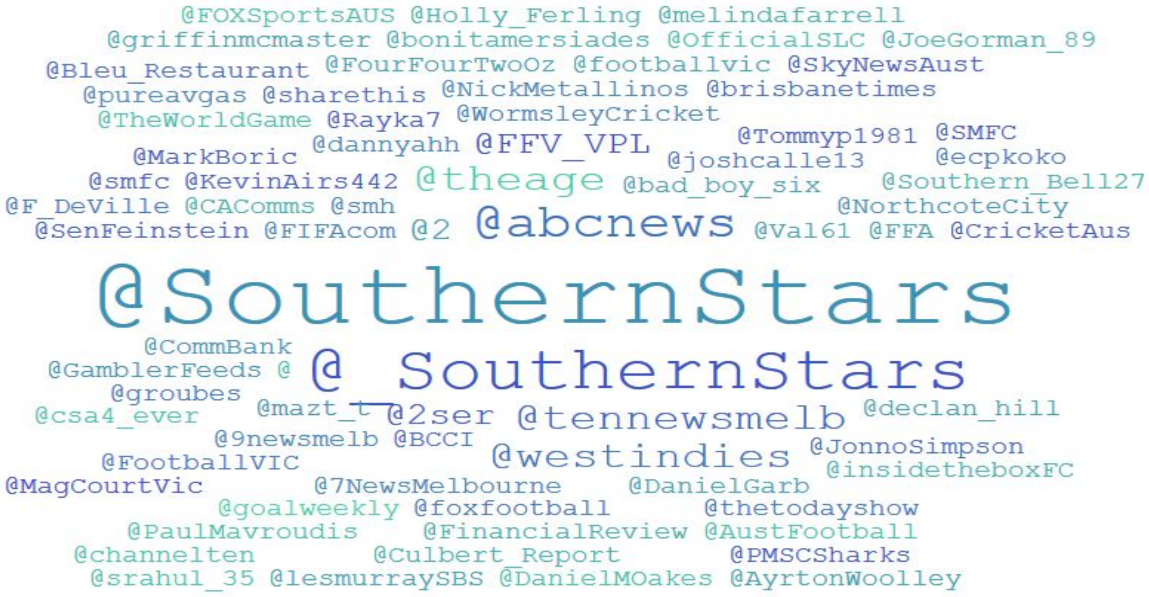


Figure 30: *Tag cloud of the most prevalent mentions of the tweets, including search terms*

Statistics

Tweets per day

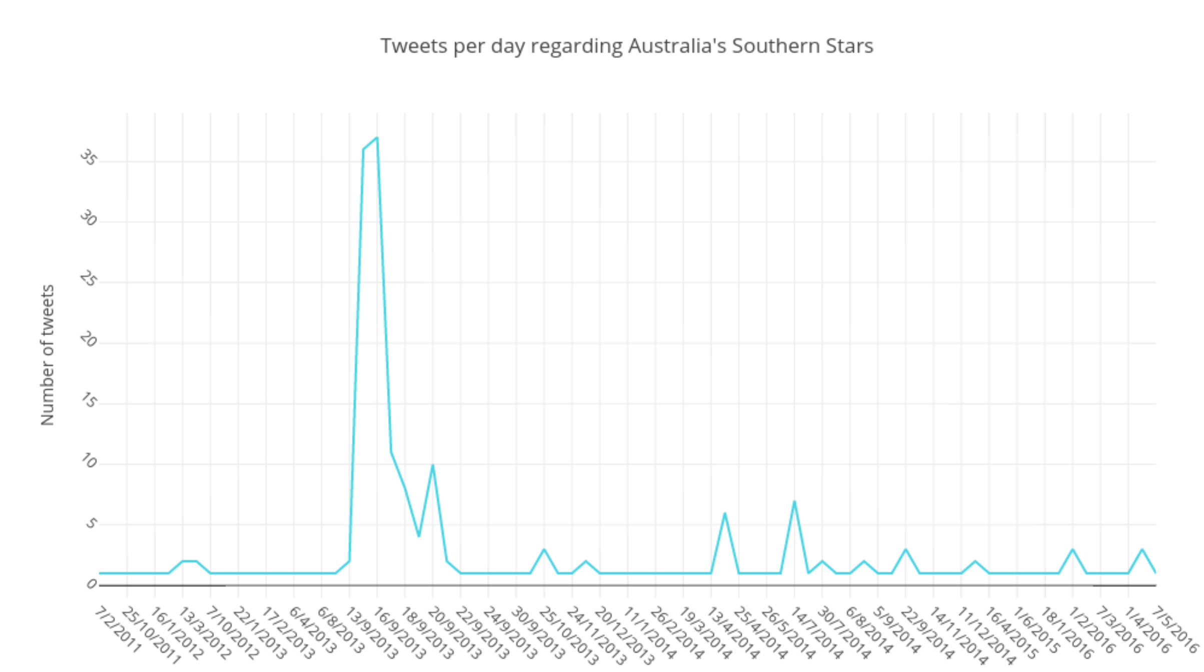


Figure 31: *Line plot representing tweets per day in the context of the Southern Stars*

Most active users

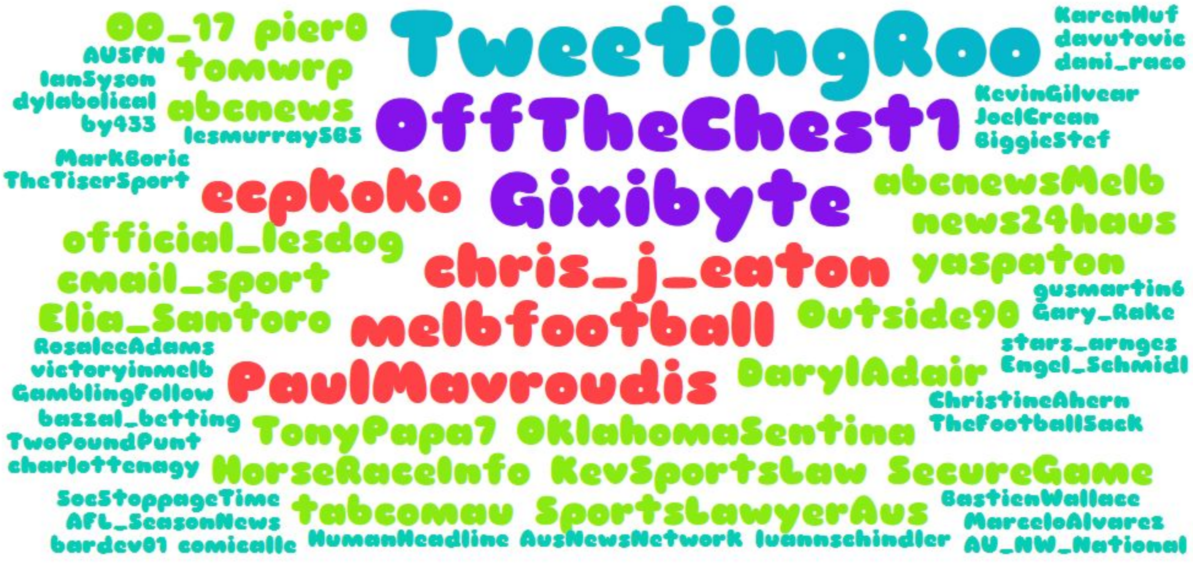


Figure 32: *Tag cloud of the top 90 users posting tweets relevant to the Southern Stars*

Basic Sentiments Scores

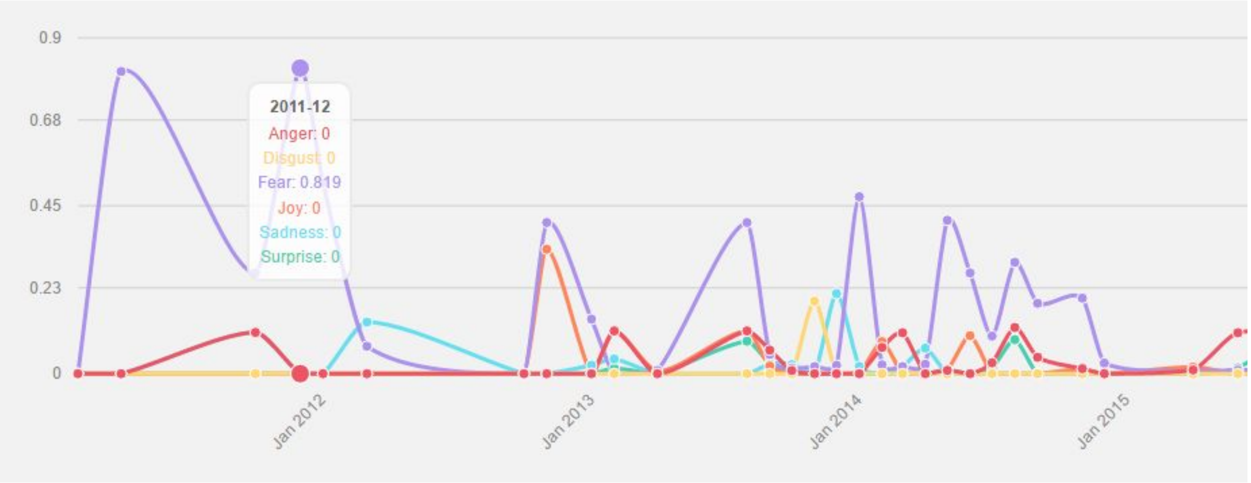


Figure 33: *LIne plot representing levels of the six basic sentiments (Anger, Disgust, Fear, Joy, Sadness & Surprise), on each day and month.*

Indicative Scenario - Pakistan cricket spot-fixing scandal

Overview

Finally, the last case study we analyzed was the “Pakistan cricket scandal”. This scandal of 2010 centres on certain members of Pakistan’s national cricket team being convicted of taking bribes from a bookmaker, Mazhar Majeed, to under-perform deliberately at certain times in a Test match at Lord's Cricket Ground, London, in 2010. More specifically, some reporters videotaped the bookmaker accepting money and informing the reporters that some players would deliberately bowl no balls at specific points in an over. Three cricket players were banned and convicted for this case; Salman Butt, Mohammad Asif and Mohammad Amir.

In this case, we analysed data both from Twitter and YouTube. The tweets range from 29/8/2010, when the scandal was revealed, until 01/06/2016. We visualised the collected data in tag clouds of frequent words, hashtags, users and plotted the number of tweets per day. Also, we created a world map visualizing the countries that tweeted about the scandal. Our total dataset is consisted of 2926 tweets and 401 YouTube comments.

Analytics Summary

Below we present the results of our research using multiple tag clouds, plots and maps.

First of all, we present tag clouds of frequent words and mentions used in tweets and comments (Fig. 34-35). In this tag cloud, we notice that the most frequent words are related to corruption and fixed matches (e.g. betting, corruption, fixing, bet, match). Also, "cricket" and "pakistan" are the most frequent words and hashtags, indicating that most of the tweets were linked to this specific cricket scandal. In the meantime, we observe that most of the users that tweeted about the scandal (Fig. 36) were betting and news media accounts. Furthermore, in Figure 37 we notice that, as in most cases, users tweeted mostly during the days when the scandal was announced (Fig. 37). Finally, during our research about the location of the users, we noticed that most of the users tweeted (Fig. 38) from countries where cricket is famous like India, Pakistan, Great Britain and Australia.

- ### Query Words: cricket, Asif, Amir, fixing, betting, corruption, bet, fix, scandal

Top-20 Words/Hashtags/Mentions Frequencies

Word	Frequency	Word (cont.)	Frequency (cont.)
cricket	4519	twitter	555
corruption	2085	amir	520
bet	1791	scandal	347
betting	1335	india	324
pakistani	1299	tips	314
fixing	960	asif	301
match	768	team	299
free	676	news	280
fix	670	khan	277
pakistan	663	win	269

Hashtag	Frequency	Hashtag (cont.)	Frequency (cont.)
#cricket	349	#bet	88
#pakistan	237	#TENNIS	81
#IPL	197	#BASKETBALL	81
#WT20	133	#GPL2016	80
#CRICKET	132	#SNOOKER	79
#news	126	#FOOTBALL	79
#IPL2016	123	#CRiCKET	79
#T20	113	#Pakistani	78
#inplaymagic	105	#Inplay	68
#bet	88	#betting	68

Mention	Frequency	Mention (cont.)	Frequency (cont.)
@YouTube	78	@herefordrich	15
@GoPaisaCom	70	@ECB_cricket	14
@ti_asif	53	@TheRealPCB	14
@Cricket_Tipster	42	@KlasraRauf	13
@BCCI	34	@IPL	11
@ICC	32	@HarperCollinsIN	11
@ImranKhanPTI	27	@dwnews	11

@eastbridge	21	@msdhoni	11
@Ihab_Amir	16	@emraanhashmi	10
@imVkohli	16	@ianuragthakur	10

Tag Clouds

Word Frequency

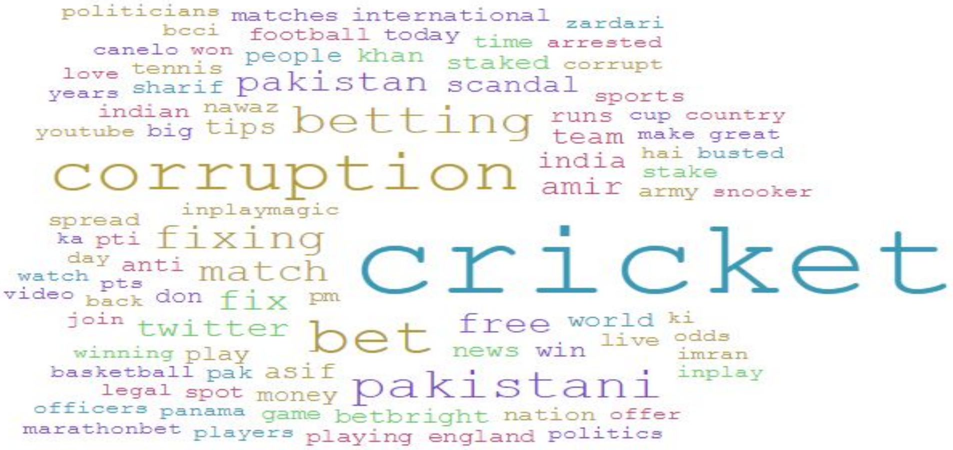


Figure 34: *Tag Cloud of most frequent used words in tweets and YouTube comments*

Hashtags Frequency

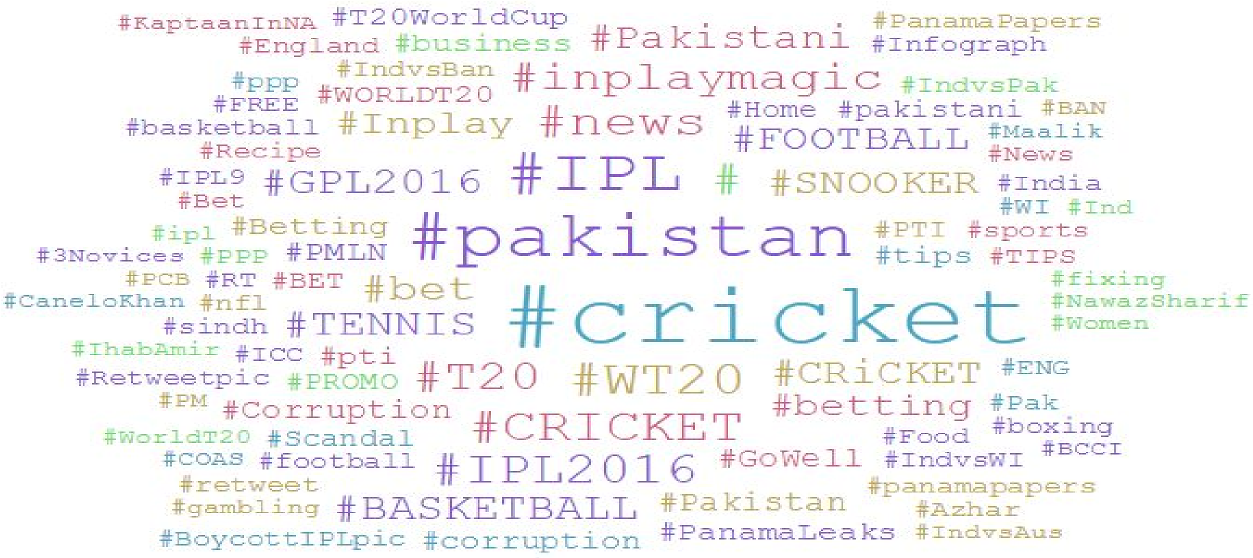


Figure 35: *Tag Cloud of most frequent hashtags in tweets*

User Frequency

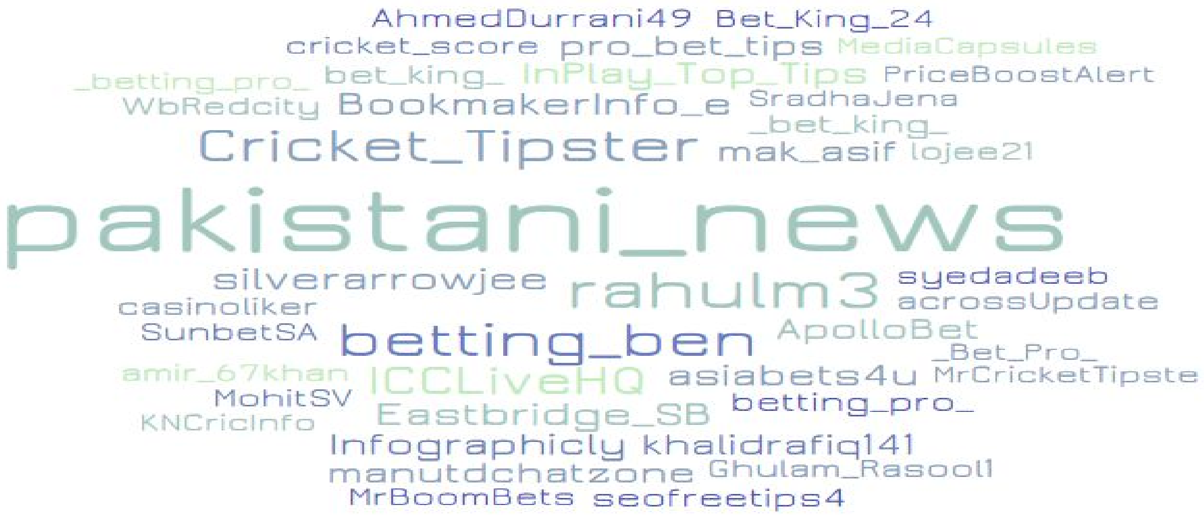


Figure 36: *Tag Cloud of most tweets per user about the scandal*

Statistics

Tweets Per Day

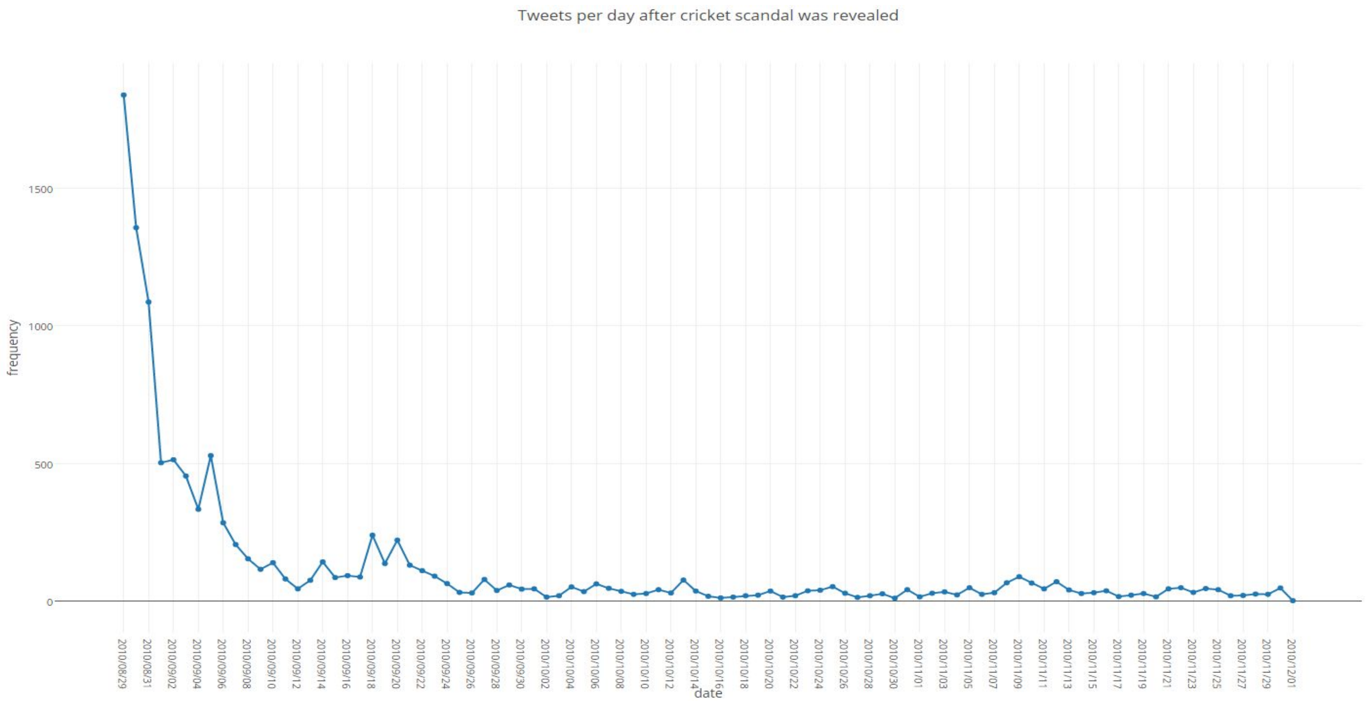


Figure 37: *Line plot of tweets per day after the scandal was revealed*

World Map of countries tweeted about the scandal

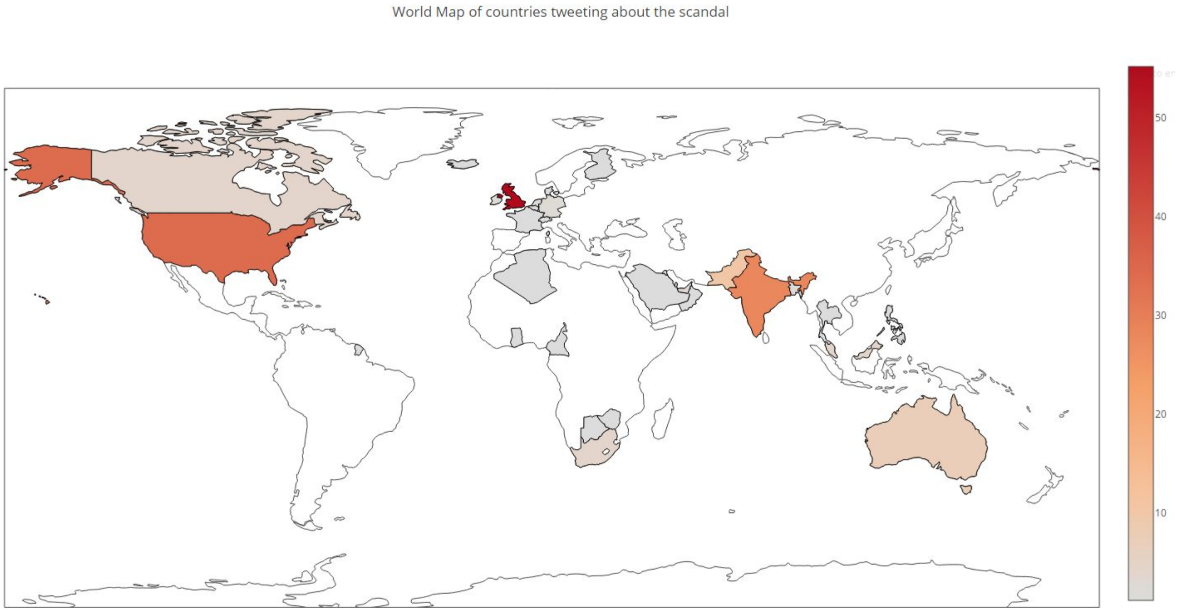


Figure 38: *Heat Map of countries tweeted about the scandal based on tweets count*

Basic Sentiments Scores

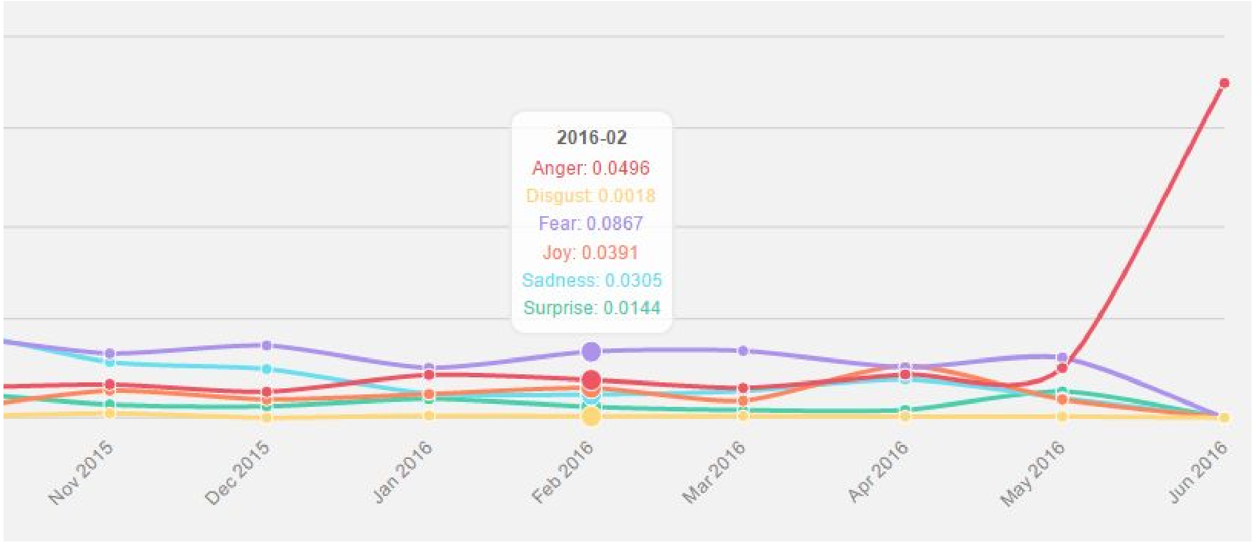


Figure 39: *Line plot representing levels of the six basic sentiments (Anger, Disgust, Fear, Joy, Sadness & Surprise), on each month.*

CASE CONCLUSIONS

There are many differences among the case studies, mainly due to the number of users discussing the subject, which varies according to the prestige of the professional as well as the location of the scandal. However, a reader of this study can easily notice some **common patterns** between the results of each case. Some of these patterns are the following:

- The results of each case study are proportional to the social impact the case subject has. For instance, the Djokovic case had many more tweets than the Australian Southern Stars case, due to the fact that Novak Djokovic is ranked as the number one tennis player in the world - therefore is world widely known - and the Southern Stars compete in the Victorian League, which is followed mostly by Australian viewers.
- In most cases, 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. The Donaghy scandal constitutes an exception to this rule, because of the extend of the scandal (FBI took part in the investigations).
- Most of the accounts that appear to be mainly involved in the user activity are news agencies and betting accounts, which shows that the many of the accounts involved are interested on taking advantage of the scandals either for popularity or directly for betting.
- There are many common terms in almost all cases, such as fixing, betting, odds etc. In the table below there is a cross-use cases vocabulary with the 20 most common terms between the four cases and the general dictionary.

20 Most Frequent Common Words in all Cases			
match	game	corruption	rt
fixing	today	live	games
fix	win	tonight	ll
betting	pic	time	people
twitter	play	bit	suspicious

Additionally, social media harvesting results can be quite **beneficial** for **educational activities**. More specifically, such results can be exploited to focus on the following aspects:

- *Assessing social media users influence*: in Figure 15 we notice that most of the Twitter Accounts who tweeted about the scandal are mass media (news agencies accounts), i.e. it seems that the news about the Djokovic Scandal were spread massively, as media accounts have the biggest influence in spreading news (due to their large social network scale). The same applies for Figures 25, 32 and 36. At the same time, betting relevant twitter accounts of great influence are included in the top Twitter users in most of the diagrams, including Figure 7, who interacted heavily about the scandal, indicating that betting agencies try to take advantage of a scandal. In the Djokovic case for example @BetfairExchange (101K followers) and @Betshoot (3K followers) are very high among the top-15 users. Along with users profile, the timeline of influence is also indicated in Figures 6, 14, 17, 24, 31 and 37. So it seems that the tweets about the Djokovic scandal were spread rapidly, as most of the tweets were submitted in the first two days of the announcement, whilst the diagram of tweets about Donaghy is spread with many spikes throughout the years.
    - **Educational activity**: these results can be highlighted at an athletes' or trainers' educational task since they indicate that no scandal can be covered up (hidden) for long since user accounts of large impact (media accounts) tweet massively when it comes to scandals and big names. 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.
- *Impacting on athlete's or trainer's popularity and fame*: the Djokovic scandal doesn't seem to have affected the image of the athlete who proceeded to an honest public declaration. This is indicated by the second case and the results of Figure 16 which show that the majority of users didn't make negative posts about N. Djokovic. This is possibly due to the fact that Djokovic refused the offer and condemned fixing publicly, so he was not accused of any allegations. On the other hand, Tim Donaghy's career was stigmatised by the scandal and as we notice on Figure 24, users are still discussing about him even after all these years. In the same way, we observe that in the cricketer's case, public opinion was strongly influenced about their career and even though some of the players returned to action, their names were stigmatised with words like "corruption" and "betting". Thus, scandals like this could cost an athlete's career, if they would consider to accept an offer like that and be a part of a corrupted match.
    - **Educational activity**: these results could be used to educate athletes and trainers about the risks of taking into consideration such offers and how

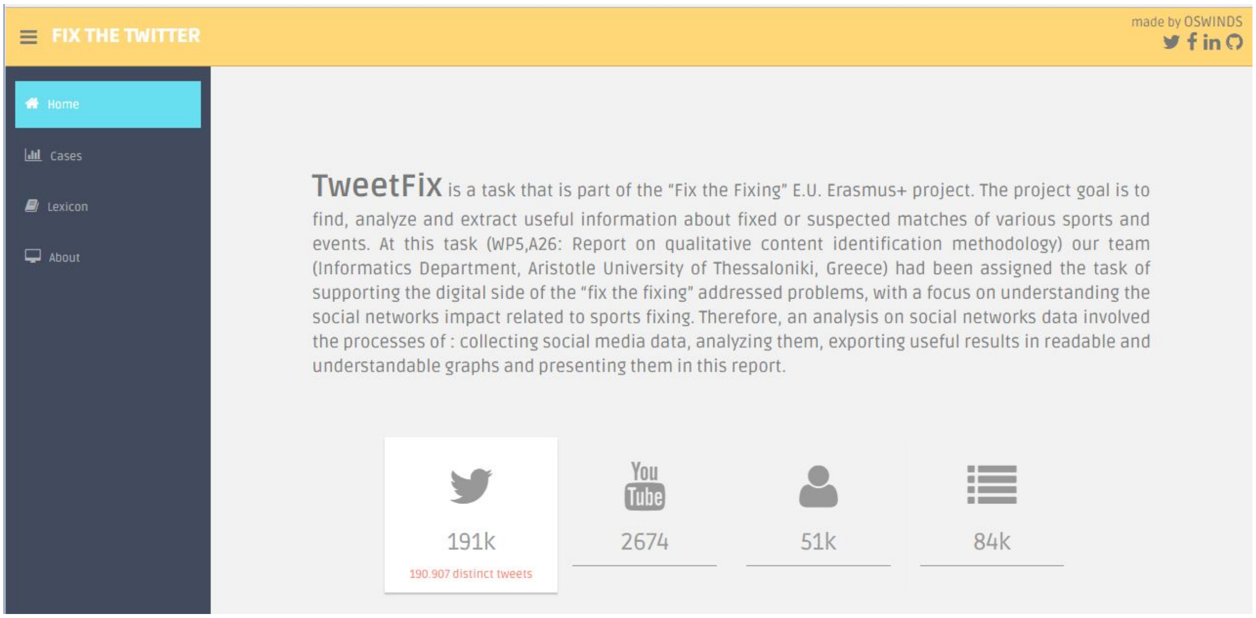
their image is strongly connected with their attitude, and especially their willingness to reveal any non ethical fixing case. For most sports-related professionals, their careers is something they build over time and it takes years to build a career like Djokovic or Donaghy did. The cases we researched are real life cases and the results come from real life opinions· they are cases of real athletes and some of them are cases of world famous athletes. Athletes and trainers should be advised to follow Djokovic's example in order to save their careers and in the same time to avoid Donaghy's example. Because, as we have concluded in our research, world famous careers can be stigmatised and even destroyed in a couple of days (2-3 most of the times). Athletes should be aware of the consequences caused by match fixing and this research presents them in a clear way.

WEB TOOL

We implemented a web tool to visualize the results of all the cases using interactive libraries, for better understanding. There is a Home page, where we describe our task TweetFix, some pages where we present the results for each case separately, a page with the lexicon including the most frequent words, as presented in Appendix D, and the About page, where we present the project and our team.
The tool can be found here: http://oswinds.csd.auth.gr/tweetfix/ and the repository with our code can be found here.

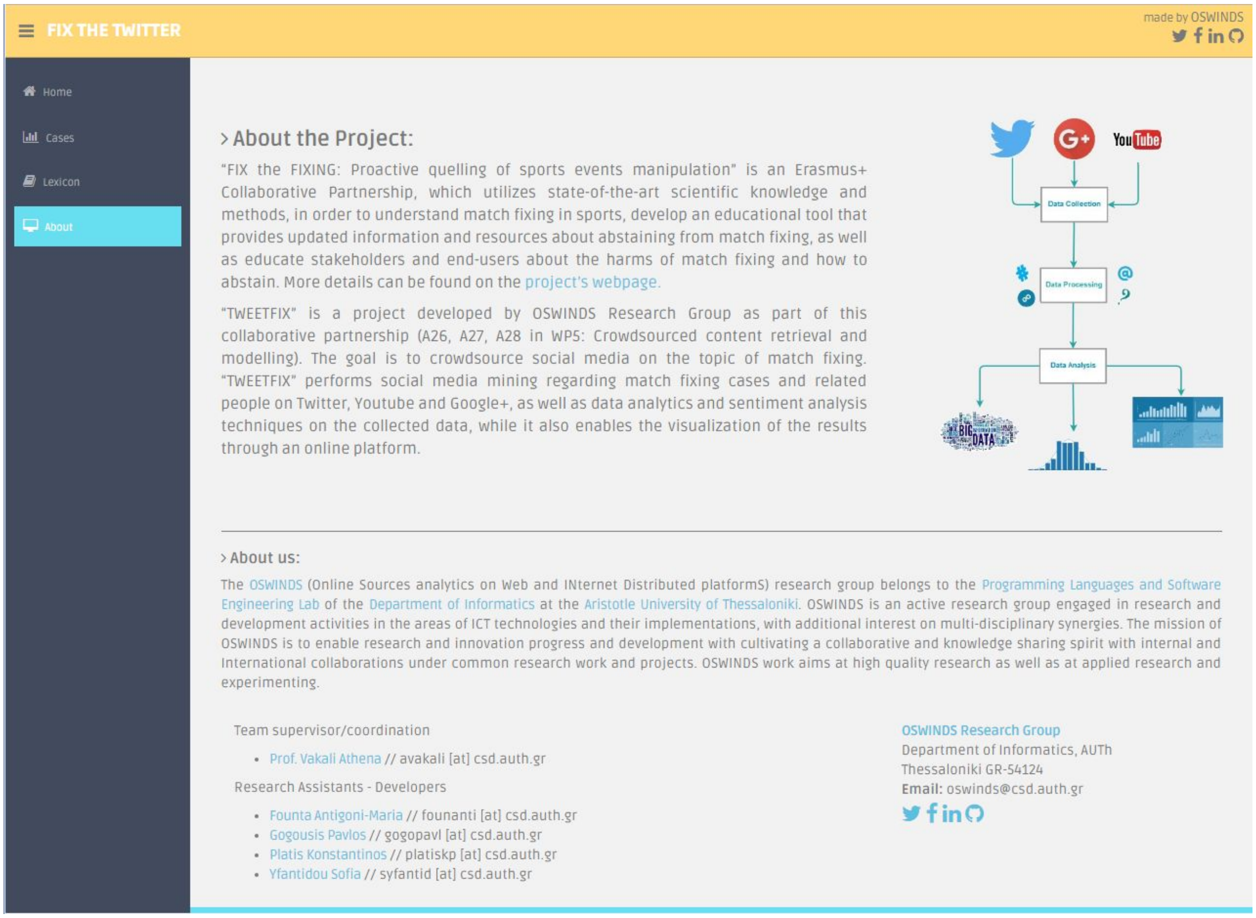
Home & About pages

In Home page the user can find useful information regarding our task as well as information about the summary of the data we collected; the summary of tweets collected for all five cases, the youtube comments, the number of users that participated and the total amount of distinct words in all the cases.



TweetFix is a task that is part of the "Fix the Fixing" E.U. Erasmus+ project. The project goal is to find, analyze and extract useful information about fixed or suspected matches of various sports and events. At this task (WP5,A26: Report on qualitative content identification methodology) our team (Informatics Department, Aristotle University of Thessaloniki, Greece) had been assigned the task of supporting the digital side of the "fix the fixing" addressed problems, with a focus on understanding the social networks impact related to sports fixing. Therefore, an analysis on social networks data involved the processes of : collecting social media data, analyzing them, exporting useful results in readable and understandable graphs and presenting them in this report.

The About page, on the other hand, describes Fix the Fixing and our task in a little more detail, while afterwards there is a description and contact information of the OSWINDS research group and the developing team of TweetFix.



> About the Project:

"FIX the FIXING: Proactive quelling of sports events manipulation" is an Erasmus+ Collaborative Partnership, which utilizes state-of-the-art scientific knowledge and methods, in order to understand match fixing in sports, develop an educational tool that provides updated information and resources about abstaining from match fixing, as well as educate stakeholders and end-users about the harms of match fixing and how to abstain. More details can be found on the project's webpage.

"TWEETFIX" is a project developed by OSWINDS Research Group as part of this collaborative partnership (A26, A27, A28 in WP5: Crowdsourced content retrieval and modelling). The goal is to crowdsource social media on the topic of match fixing. "TWEETFIX" performs social media mining regarding match fixing cases and related people on Twitter, Youtube and Google+, as well as data analytics and sentiment analysis techniques on the collected data, while it also enables the visualization of the results through an online platform.

> About us:

The OSWINDS (Online Sources analytics on Web and INternet Distributed platformS) research group belongs to the Programming Languages and Software Engineering Lab of the Department of Informatics at the Aristotle University of Thessaloniki. OSWINDS is an active research group engaged in research and development activities in the areas of ICT technologies and their implementations, with additional interest on multi-disciplinary synergies. The mission of OSWINDS is to enable research and innovation progress and development with cultivating a collaborative and knowledge sharing spirit with internal and International collaborations under common research work and projects. OSWINDS work aims at high quality research as well as at applied research and experimenting.

Team supervisor/coordination

- Prof. Vakali Athena // avakali [at] csd.auth.gr

Research Assistants - Developers

- Founta Antigoni-Maria // founanti [at] csd.auth.gr
- Gogousis Pavlos // gogopavl [at] csd.auth.gr
- Platis Konstantinos // platiskp [at] csd.auth.gr
- Yfantidou Sofia // syfantid [at] csd.auth.gr

OSWINDS Research Group
Department of Informatics, AUTh
Thessaloniki GR-54124
Email: oswinds@csd.auth.gr

Case pages

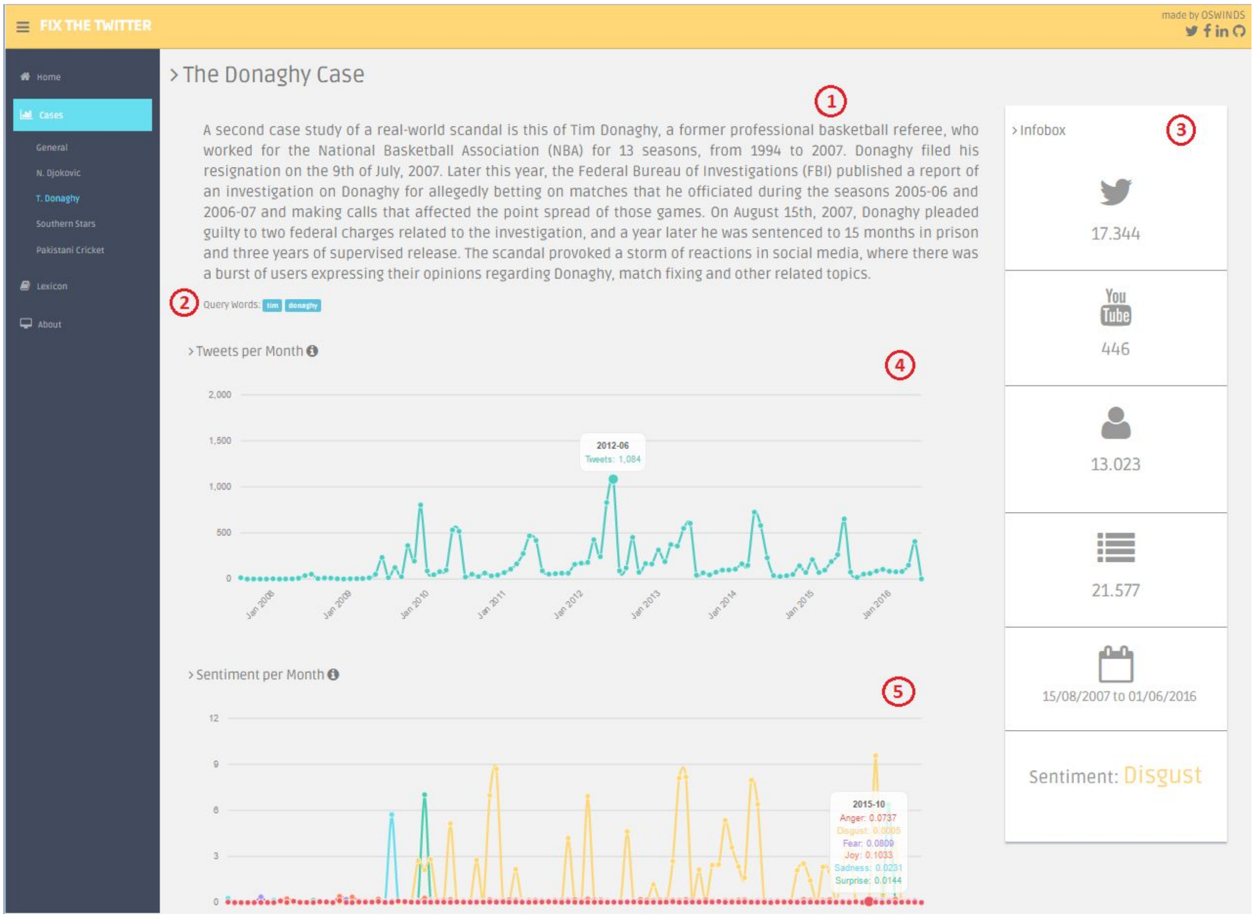
There are five cases presented in this report that were related to match fixing and each one of these cases has its own page in the web tool. The pages are:

- General Fixing case
- Novak Djokovic case
- Tim Donaghy case
- Southern Stars of Australia case
- Pakistani Cricket case

For each one of the cases we present:

1. A description;
2. The query words that we used to search for this case on social media;
3. An information box in which we mention some information regarding the dataset, such as the amount of tweets and youtube comments we collected, the time frame of the data, the amount of distinct words and users as well as the results of our Sentiment Analysis summarized in the dominating sentiment;

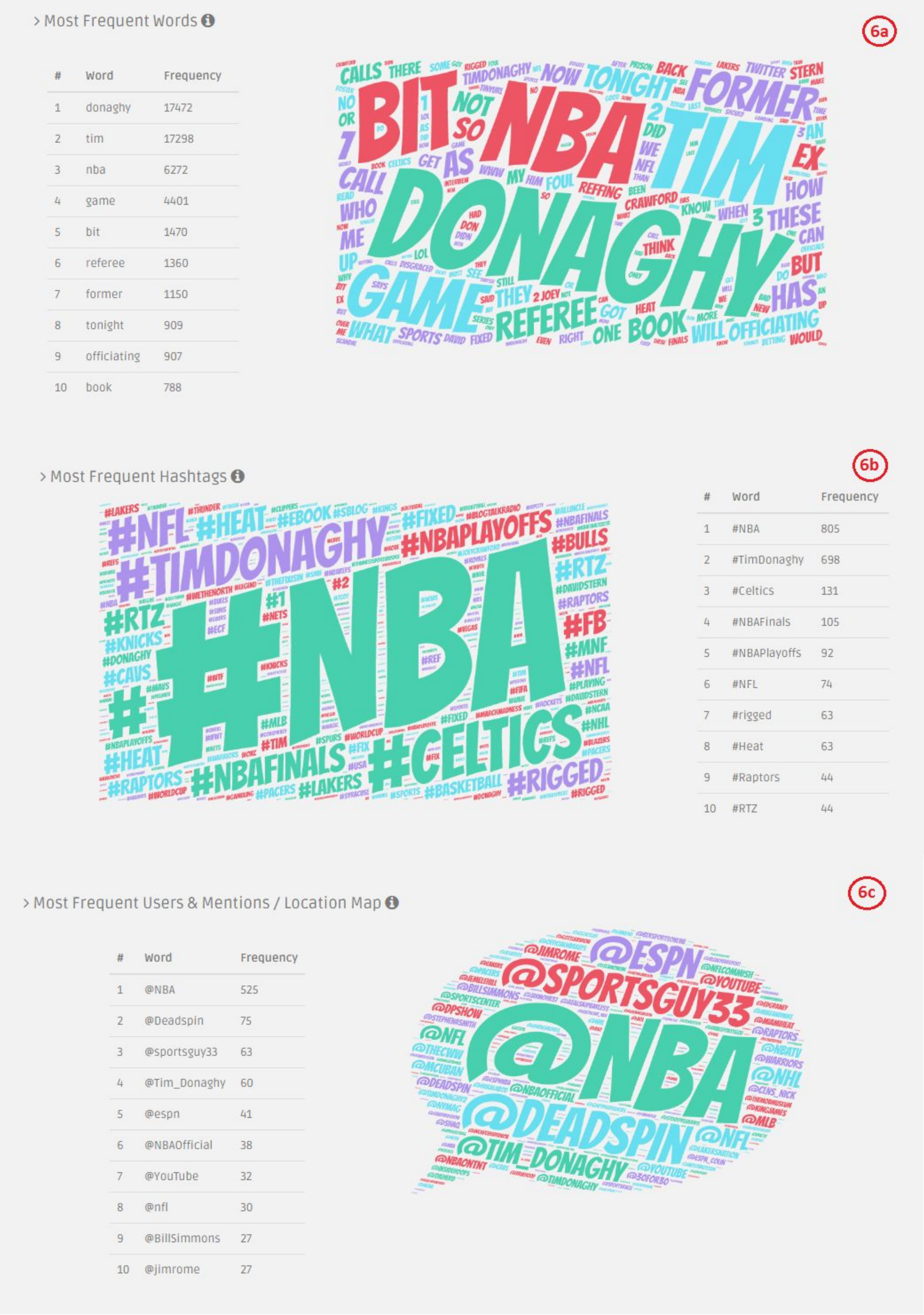
4. A date chart where the user can find the amount of tweets for each month of the dataset time frame, in order to detect any spikes on the amount of people that discussed this case on social media;
5. A sentiment chart where there are monthly results of our sentiment analysis for all six basic sentiments, color coded so that any spikes on any sentiment can be easily detected;
6. The most frequent distinct words (a), hashtags (b) and user mentions (c) of the dataset, presented both in a tag cloud as well as in a table, with the Top10; and
7. A choropleth map with the locations of the users that discussed about the case, both from Twitter and YouTube.



> The Donaghy Case

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.

Query Words: tim donaghy



> Most Frequent Words

6a

#	Word	Frequency
1	donaghy	17472
2	tim	17298
3	nba	6272
4	game	4401
5	bit	1470
6	referee	1360
7	former	1150
8	tonight	909
9	officiating	907
10	book	788

> Most Frequent Hashtags

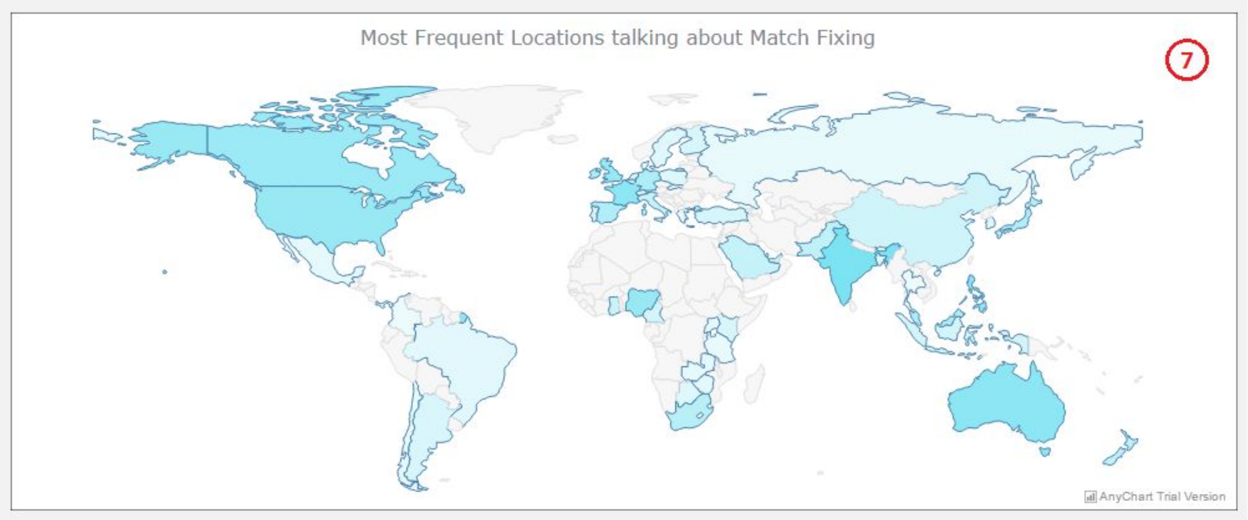
6b

#	Word	Frequency
1	#NBA	805
2	#TimDonaghy	698
3	#Celtics	131
4	#NBAFinals	105
5	#NBAPlayoffs	92
6	#NFL	74
7	#rigged	63
8	#Heat	63
9	#Raptors	44
10	#RTZ	44

> Most Frequent Users & Mentions / Location Map

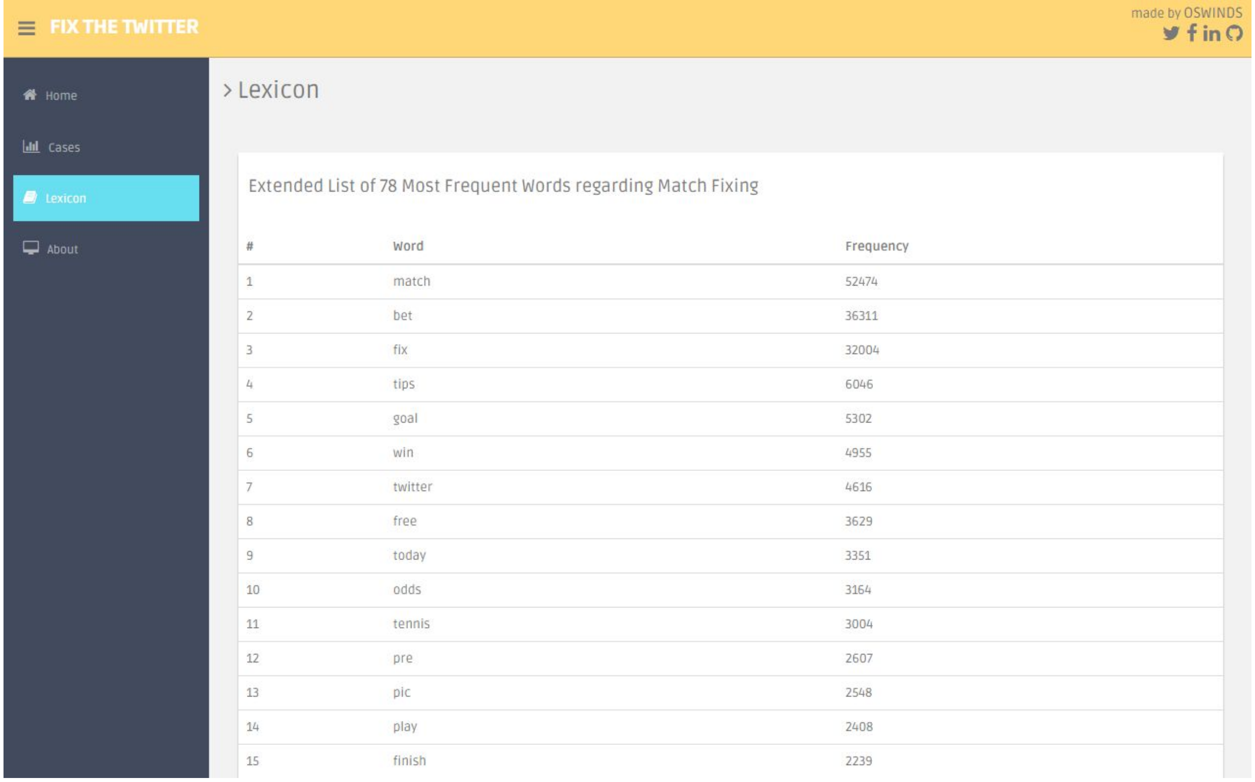
6c

#	Word	Frequency
1	@NBA	525
2	@Deadspin	75
3	@sportsguy33	63
4	@Tim_Donaghy	60
5	@espn	41
6	@NBAOfficial	38
7	@YouTube	32
8	@nfl	30
9	@BillSimmons	27
10	@jimrome	27



Lexicon page

Finally, the lexicon page presents the 80 most frequent distinct words and their frequencies, which were derived from the general case, and are also presented in the Appendix D of this report.



› Lexicon

Extended List of 78 Most Frequent Words regarding Match Fixing

#	Word	Frequency
1	match	52474
2	bet	36311
3	fix	32004
4	tips	6046
5	goal	5302
6	win	4955
7	twitter	4616
8	free	3629
9	today	3351
10	odds	3164
11	tennis	3004
12	pre	2607
13	pic	2548
14	play	2408
15	finish	2239

CONCLUSIONS

To conclude, our research is divided into 3 basic parts:

1. Data Collection, Case and Sentiment Analysis
2. Lexicon Expansion
3. Creation of Web Tool

During the first part, we collected data from 2 different crowdsourcing applications - Youtube and Twitter - in order to make conclusions about different cases of fixed sport events. More specifically, we analyzed the cases of Djokovic (2015), Donaghy (2007), Australian Football Case (2013) and Pakistani Cricket Case (2010) and created plots, tag clouds and maps that helped us better understand each case and the implications it had on each person's career. Moreover, we conducted sentiment analysis on the data collection in order to plot the sentiment of the crowd during the case revealing. Finally, we analyzed the results, made conclusions for every case separately and presented them in this report. Those results can be used to present to athletes, coaches, referees etc. the effects of match fixing and prevent them from taking part in similar events.

The second part of our research is linked to the first as it contains new words on match and sports event fixing. During our research, we mined new words that give us the opportunity to better understand the world of sports fixing. We collected those words and created a new "Fixing Lexicon" that contains words related to sports fixing and exclusively mined from Crowdsourcing applications like Twitter and Youtube. By studying our expanded lexicon, future researchers can better understand match fixing or even create tools that detect suspicious cases.

The third and final part is the creation of our Web tool. This tool is created to present our project and research to the outer world. More specifically, we analyze each case separately by presenting the produced graphs in a more user-friendly format to help people interested in our research understand both the cases and their results. Our web tool can be found in http://oswinds.csd.auth.gr/tweetfix/.

REFERENCES

1. Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. “Why We Twitter: Understanding Microblogging Usage and Communities.” Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (2007): 56-65.
2. Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. “Breaking News on Twitter.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2012): 2751-2754.
3. Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. “How Useful are Your Comments? - Analyzing and Predicting YouTube Comments and Comment Ratings.” Proceedings of the 19th International Conference on World Wide Web (2010): 891-900.
4. Daniel E. O'Leary. "Twitter Mining for Discovery, Prediction and Causality: Applications and Methodologies." Intelligent Systems in Accounting, Finance and Management 22.3 (2015): 227-247.
5. Meena Nagarajan, Amit Sheth, and Selvam Velmurugan. "Citizen Sensor Data Mining, Social Media Analytics and Applications." Proceedings of the 20th International Conference Companion on World Wide Web (2015): 289-290.
6. Chong Oh, Sheila Sasser, and Soliman Almahmoud. “Social media analytics framework: the case of Twitter and Super Bowl ads.” Journal of Information Technology Management (2015).
7. Dominique Haughton, Mark-David McLaughlin, Kevin Mentzer, and Changan Zhang. "Can We Predict Oscars from Twitter and Movie Review Data?." Movie Analytics: A Hollywood Introduction to Big Data (2015): 41-54.
8. Despoina Chatzakou, and Athena Vakali. “Harvesting Opinions and Emotions from Social Media Textual Resources.” IEEE Internet Computing 19.4 (2015): 46-50.
9. Chenyan Xu, Yang Yu, and Chun-Keung Hoi. “Hidden In-Game Intelligence in NBA Players’ Tweets.” Communications of the ACM 58.11 (2015): 80-89.
10. Ryan M. Rodenberg, Elihu D. Feustel. “Forensic Sports Analytics: Detecting and Predicting Match-Fixing in Tennis.” Journal of Prediction Markets 8.1 (2014): 77-95.
11. Kevin Carpenter. “Match-Fixing - the Biggest Threat to Sport in the 21st Century?.” International Sports Law Review (2012): 13-24.
12. Daniel J. Friedman. “Social Media In Sports : Can Professional Sports League Commissioners Punish ’Twackle Dummies’?.” Social Media In Sports : Can Professional Sports League Commissioners. Pace I.P. Sports & Entertainment Law Forum 2.1 (2012): 72–102.
13. Robert P. Schumaker, Osama K. Solieman, and Hsinchun Chen. “Sports knowledge management and data mining.” Annual Review of Information Science and Technology 44 (2010): 115–157.
14. Paul Ekman, and Wallace V. Friesen. “Unmasking the face - A guide to recognizing emotions from facial cues.” Journal of Personality (1975).

APPENDICES

Appendix A - A short introduction to Web APIs[17]

Web APIs are the defined interfaces through which interactions happen between an enterprise and applications that use its assets. An API approach is an architectural approach that revolves around providing programmable interfaces to a set of services to different applications serving different types of consumers. When used in the context of web development, an API is typically defined as a set of Hypertext Transfer Protocol (HTTP) request messages, along with a definition of the structure of response messages, which is usually in an Extensible Markup Language (XML) or JavaScript Object Notation (JSON) format. While "web API" historically has been virtually synonymous for web service, the recent trend (so-called Web 2.0) has been moving away from Simple Object Access Protocol (SOAP) based web services and service-oriented architecture (SOA) towards more direct representational state transfer (REST) style web resources and resource-oriented architecture (ROA). Part of this trend is related to the Semantic Web movement toward Resource Description Framework (RDF), a concept to promote web-based ontology engineering technologies. Web APIs allow the combination of multiple APIs into new applications known as mashups.

Web use to share content

The practice of publishing APIs has allowed web communities to create an open architecture for sharing content and data between communities and applications. In this way, content that is created in one place can be dynamically posted and updated in multiple locations on the web:

- Photos can be shared from sites like Flickr and Photobucket to social network sites like Facebook and MySpace.
- Content can be embedded, e.g. embedding a presentation from SlideShare on a LinkedIn profile.
- Content can be dynamically posted. Sharing live comments made on Twitter with a Facebook account, for example, is enabled by their APIs.
- Video content can be embedded on sites served by another host.
- User information can be shared from web communities to outside applications, delivering new functionality to the web community that shares its user data via an open API. One of the best examples of this is the Facebook Application platform. Another is the Open Social platform.

[17] "Application programming interface - Wikipedia, the free encyclopedia." 2011. 11 May. 2016 <https://en.wikipedia.org/wiki/Application_programming_interface>

- If content is a direct representation of the physical world (e.g., temperature at a geospatial location on earth) then an API can be considered an "Environmental Programming Interface" (EPI). EPIs are characterized by their ability to provide a means for universally sequencing events sufficient to utilize real-world data for decision making.

Appendix B - A short explanation of MongoDB[18]

MongoDB is a free and open-source cross-platform document-oriented database. Classified as a NoSQL database, MongoDB avoids the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster. MongoDB is developed by MongoDB Inc. and is free and open-source, published under a combination of the GNU Affero General Public License and the Apache License. As of July 2015, MongoDB is the fourth most popular type of database management system, and the most popular for document stores.

Main features

Some of the features include:

Ad hoc queries

MongoDB supports field, range queries, regular expression searches. Queries can return specific fields of documents and also include user-defined JavaScript functions.

Indexing

Any field in a MongoDB document can be indexed – including within arrays and embedded documents (indices in MongoDB are conceptually similar to those in RDBMSes). Primary and secondary indices are available.

Replication

MongoDB provides high availability with replica sets. A replica set consists of two or more copies of the data. Each replica set member may act in the role of primary or secondary replica at any time. The all writes and reads are done on the primary replica by default. Secondary replicas maintain a copy of the data of the primary using built-in replication. When a primary replica fails, the replica set automatically conducts an election process to determine which secondary should become the primary. Secondaries can optionally serve read operations, but that data is only eventually consistent by default.

Load balancing

MongoDB scales horizontally using sharding. The user chooses a shard key, which determines how the data in a collection will be distributed. The data is split into ranges (based on the shard key) and distributed across multiple shards. (A shard is a master with one or more slaves.). Alternatively, the shard key can be hashed to map to a shard – enabling an even data distribution.

[18] "MongoDB - Wikipedia, the free encyclopedia." 2011. 9 May. 2016 <https://en.wikipedia.org/wiki/MongoDB>

MongoDB can run over multiple servers, balancing the load and/or duplicating data to keep the system up and running in case of hardware failure. MongoDB is easy to deploy, and new machines can be added to a running database.

File storage

MongoDB can be used as a file system, taking advantage of load balancing and data replication features over multiple machines for storing files.

This function, called Grid File System, is included with MongoDB drivers and available for many development languages (see "Language Support" for a list of supported languages). MongoDB exposes functions for file manipulation and content to developers. GridFS is used, for example, in plugins for NGINX and lighttpd. Instead of storing a file in a single document, GridFS divides a file into parts, or chunks, and stores each of those chunks as a separate document.

In a multi-machine MongoDB system, files can be distributed and copied multiple times between machines transparently, thus effectively creating a load-balanced and fault-tolerant system.

Aggregation

MapReduce can be used for batch processing of data and aggregation operations.

The aggregation framework enables users to obtain the kind of results for which the SQL GROUP BY clause is used. Aggregation operators can be strung together to form a pipeline – analogous to Unix pipes. The aggregation framework includes the $lookup operator which can join documents from multiple documents.

Server-side JavaScript execution

JavaScript can be used in queries, aggregation functions (such as MapReduce), and sent directly to the database to be executed.

Capped collections

MongoDB supports fixed-size collections called capped collections. This type of collection maintains insertion order and, once the specified size has been reached, behaves like a circular queue.

Appendix C - A short explanation of JSON format[19]

JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others. These properties make JSON an ideal data-interchange language.

JSON is built on two structures:

- A collection of name/value pairs. In various languages, this is realized as an *object*, record, struct, dictionary, hash table, keyed list, or associative array.
- An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence.

These are universal data structures. Virtually all modern programming languages support them in one form or another. It makes sense that a data format that is interchangeable with programming languages also be based on these structures.

In JSON, they take on these forms:

- An object is an unordered set of name/value pairs. An object begins with { (left brace) and ends with } (right brace). Each name is followed by : (colon) and the name/value pairs are separated by , (comma).
- An array is an ordered collection of values. An array begins with [(left bracket) and ends with] (right bracket). Values are separated by , (comma).
- A value can be a string in double quotes, or a number, or true or false or null, or an object or an array. These structures can be nested.
- A string is a sequence of zero or more Unicode characters, wrapped in double quotes, using backslash escapes. A character is represented as a single character string. A string is very much like a C or Java string.
- A number is very much like a C or Java number, except that the octal and hexadecimal formats are not used.
- Whitespace can be inserted between any pair of tokens. Excepting a few encoding details, that completely describes the language.

[19] "JSON." 2003. 9 May. 2016 <http://www.json.org/>

Appendix D - Extended List of 80 Most Frequent Words regarding Match Fixing

Word	Frequency	Word (cont.)	Frequency (cont.)
match	52474	tonight	1049
bet	36311	gambling	1037
fix	32004	life	1025
tips	6046	integrity	1024
goal	5302	money	996
win	4955	day	993
twitter	4616	bonus	979
free	3629	team	969
today	3351	city	929
odds	3164	livescore	902
tennis	3004	inplaybetting	900
pre	2607	broken	885
pic	2548	betfair	863
play	2408	england	857
finish	2239	problem	838
corruption	2145	deposit	829
live	2095	world	810
prediction	2045	things	768
chance	1997	draw	766

inplay	1846	news	757
challenge	1842	package	752
game	1810	slots	740
soccer	1622	place	735
returns	1604	make	722
preview	1551	score	680
time	1538	cricket	670
sport	1443	offer	664
suspicious	1406	problems	656
football	1277	players	651
people	1265	lol	622
watch	1217	open	599
man	1206	start	597
good	1177	half	589
united	1172	home	586
back	1168	real	571
league	1168	djokovic	556
fraud	1156	fc	556
saturday	1109	fstinplay	554
illegal	1084	great	552