Analyzing Realtime Tweets to Generate Match Reports

Clouded Minds

Deepak, Sandra, Will, and Michael

Topics We'll Be Covering

Michael - Project Overview

Deepak - Volume Analysis

Will - Sentiment Analysis

Sandra - Poster

Project Goal

- Show game highlights based on tweets posted by users watching the game.
    - Examples
        - Soccer
            - Red Card
            - Penalty
            - Goal

- The analysis is done post-hoc.

Data Sets

- Data Collection
  - Soccer games
    - English Premier League - 10 Game
    - Spanish League - 2 Game
    - International Friendlies - 2 Game
  - Game time and teams playing were known before hand to set up collection.
  - Filtered using official game hashtags, team names and their short nicknames.
  - Troubles collecting
    - Some previous data corrupted
    - Language parameter missed

Data Sets

- Data Preprocessing
  - Stored into MongoDB.
  - Conversion of mongo datasets to csv for better analysis.
  - Tweet timestamp translation into game minutes.
  - Data is formatted into JSON that the sentiment analyzer will accept.
  - Sentiment is then added back to the original csv file.
  - For analysis, data is split into two files for each time.

Data Sets Details

Game	Tweet Count	Avg. Tweet/Min
Real Madrid vs Barcelona	114289	1170
Chelsea vs Arsenal	119029	1200
England vs France	79381	847
Chelsea vs Southampton	91972	878
Manchester City vs Liverpool	83946	877
Swansea vs Bournemouth	5368	55

Data Set Analysis Approaches

- Volume Analysis
- Sentiment Analysis

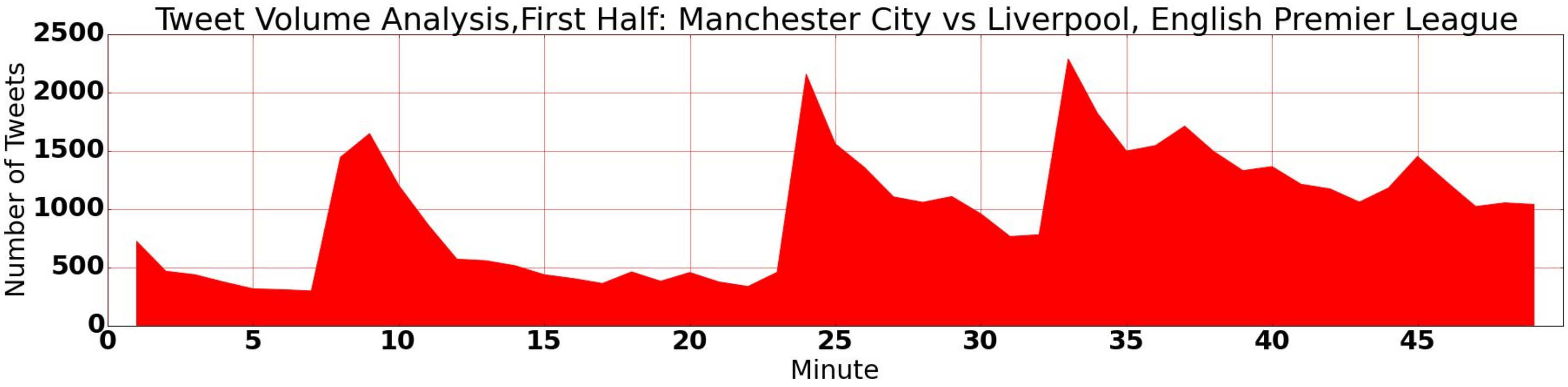
Volume Analysis

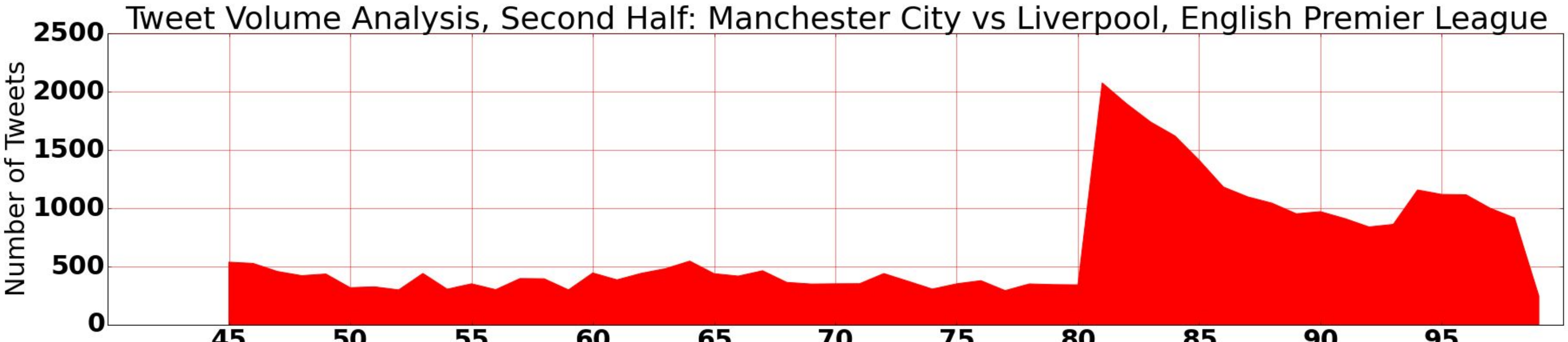
- Highlights Detection
  - Detected by analysing spikes in the volume of tweets.
    - Word analysis is performed on those spike moments to describe the moment (What exactly happened at that moment).

Volume Analysis

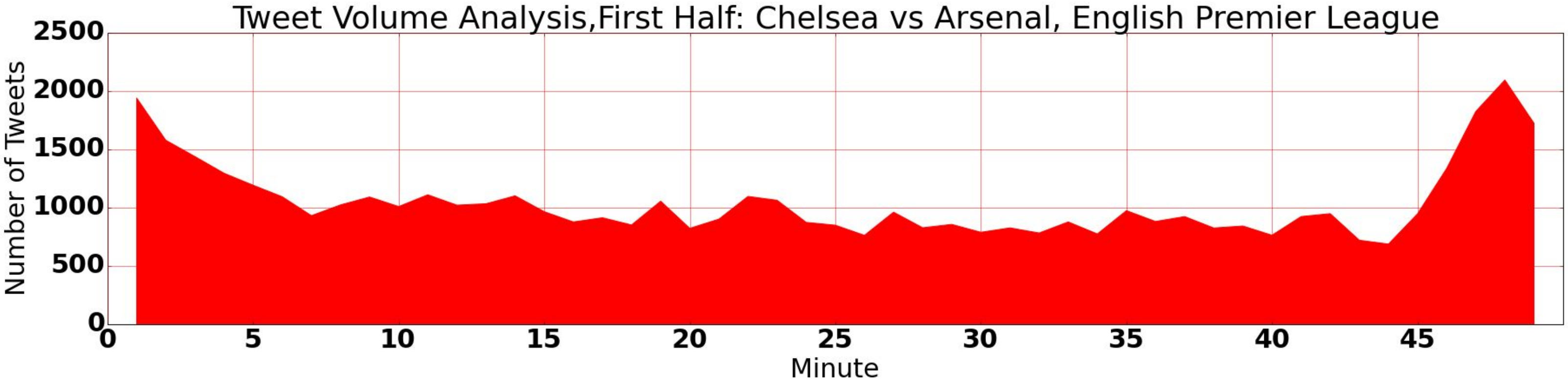
- Translation of timestamp to game halves (Each 45’ + added time)
    - Tweets in each halves grouped by minute
- The tweets per minute (TPM) is calculated
    - Each TPM Spike in a game = A Highlight in the game

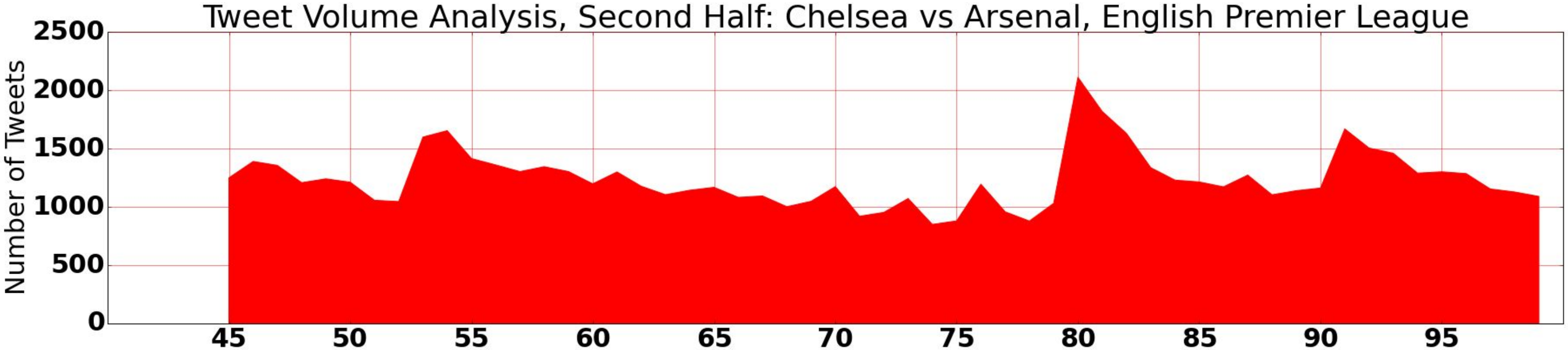
Volume Analysis


Tweet Volume Analysis,First Half: Manchester City vs Liverpool, English Premier League


Tweet Volume Analysis, Second Half: Manchester City vs Liverpool, English Premier League

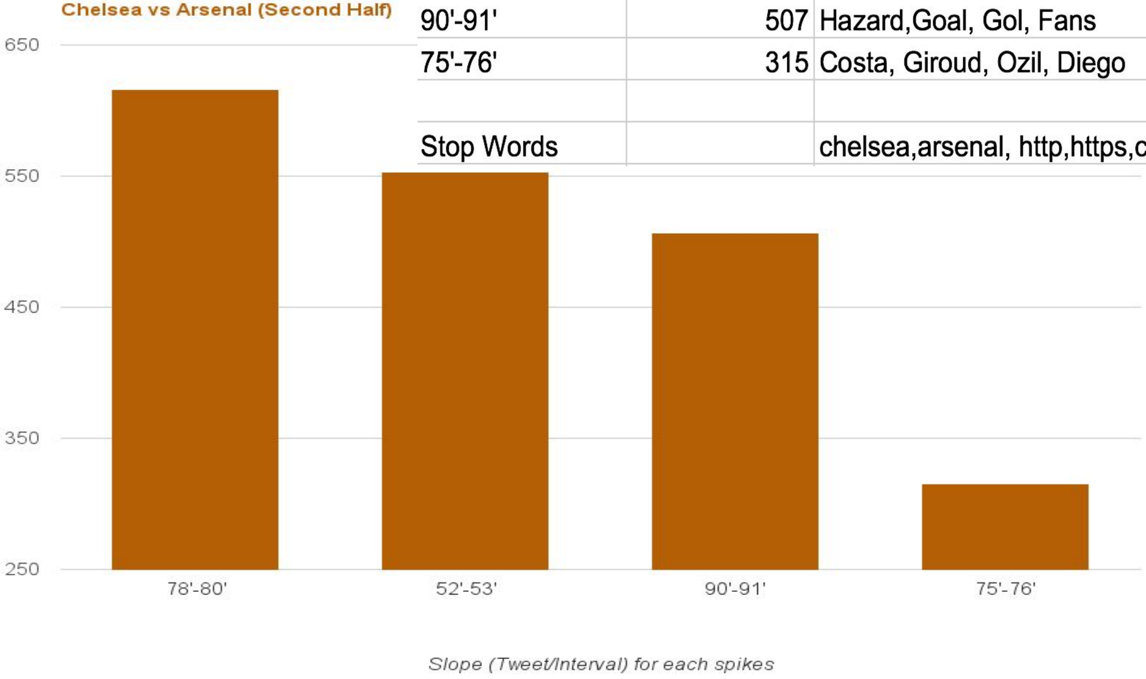
Volume Analysis





Volume Analysis

Spike	Slope Value	Top Words	Real Incident
78'-80'	616	Men, Red, Cazorla, lol	Red Card to Cazorla
52'-53'	553	Zouma, Gol,Goal, Costa	Goal by Zouma
90'-91'	507	Hazard,Goal, Gol, Fans	Goal by Hazard
75'-76'	315	Costa, Giroud, Ozil, Diego	Substitution: Ozil off, Giroud On
Stop Words		chelsea,arsenal, http,https,cfc,afc,chears	



Volume Analysis

- Filtering Top Spikes
  - Threshold: 15% of TPM
- TF-IDF calculation of tweets in top spikes
  - Stop words
    - English Stop Words
    - Keywords that used for pulling tweets from Twitter API

Volume Analysis

- TOP 10 words based on TF-IDF values
  - Keywords about the Highlight
  - Player involved in the Highlight
- Matched keywords with players and events we defined.
  - Highlight - Goal, Red Card, and Penalty
  - Player involved, and which gives team involved

Volume Analysis - Results

- Overall Game Result

Game	Tweet Count	Avg. Tweet/Min	Match Result	Predicted Result
Real Madrid vs Barcelona	114289	1170	0 - 4	0 - 3
Chelsea vs Arsenal	119029	1200	2 - 0	2 - 0
England vs France	79381	847	2 - 0	2 - 0
Chelsea vs Southampton	91972	878	1 - 3	1 - 2
Manchester City vs Liverpool	83946	877	1 - 4	2 - 3
Swansea vs Bournemouth	5368	55	2 - 2	1 - 2

- Accuracy
  - Correct Result: **0.83**
  - Correct Score: **0.33**

Volume Analysis - Results

- Goals

Game	Tweet Count	Match Result	Goal Detected(FP)	Missed Goals
Real Madrid vs Barcelona	114289	0 - 4	3(1)	1
Chelsea vs Arsenal	119029	2 - 0	2(0)	0
England vs France	79381	2 - 0	2(0)	0
Chelsea vs Southampton	91972	1 - 3	3(0)	1
Manchester City vs Liverpool	83946	1 - 4	5(0)	0
Swansea vs Bournemouth	5368	2 - 2	3(0)	1

- Accuracy
  - Precision (Correct among detected ones): **0.95**
  - Recall (Detected among actual total): **0.86**

Volume Analysis - Results

- Red Cards

Game	Red Card ?	Detected Red Cards (FP)
Real Madrid vs Barcelona	1	1(0)
Chelsea vs Arsenal	2	2(0)
England vs France	N/A	0(0)
Chelsea vs Southampton	N/A	0(0)
Manchester City vs Liverpool	N/A	0(0)
Swansea vs Bournemouth	N/A	0(0)

- Accuracy
  - Precision (Correct among detected ones): **1.0**
  - Recall (Detected among actual total): **1.0**

Volume Analysis - Results

- Penalties

Game	Penalty ?	Detected Penalties (FP)
Real Madrid vs Barcelona	N/A	0
Chelsea vs Arsenal	N/A	1
England vs France	N/A	0
Chelsea vs Southampton	N/A	1
Manchester City vs Liverpool	N/A	0
Swansea vs Bournemouth	1	1

- Accuracy
  - Precision (Correct among detected ones): **1.0**
  - Recall (Detected among actual total): **0.33**

Volume Analysis - Results

- Player Involvement Accuracy

Game	Accuracy
Real Madrid vs Barcelona	1
Chelsea vs Arsenal	0.75
England vs France	1
Chelsea vs Southampton	1
Manchester City vs Liverpool	0.83
Swansea vs Bournemouth	0.75

- Overall Accuracy: 0.89

Volume Analysis-Future Work

- Problem Faced
  - Correct player of the moment, if multiple players involved
  - Correct team when own goal is scored
  - Minor Events - Yellow Cards, Substitutions
- Improvements and Future Work
  - Proper noise removal technique
  - Extensive tweet analysis to generate match report in the form of text.

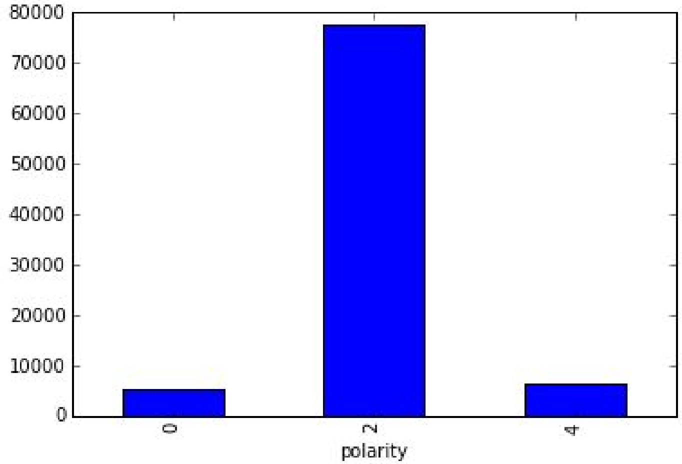
Why Sentimental Analysis?

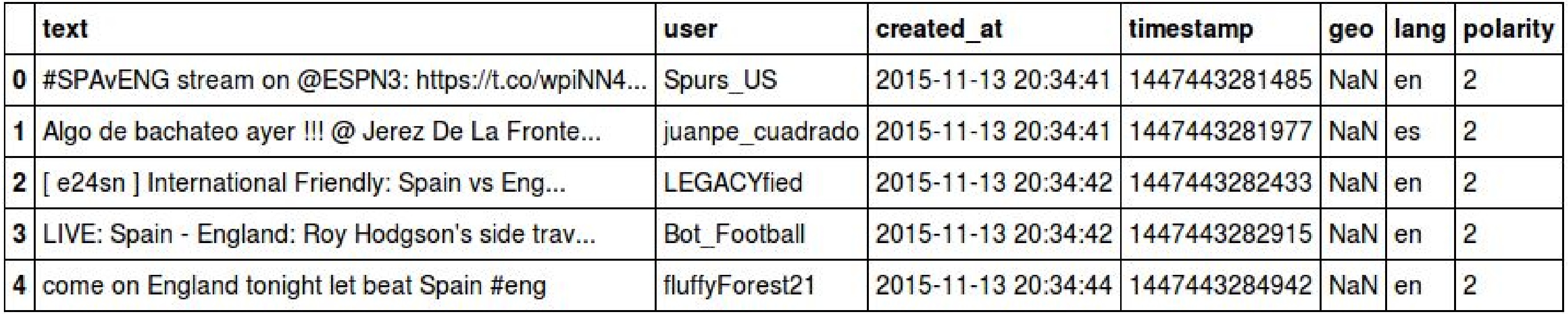
- What is Sentimental Analysis?
  - Attempting to classify a message based on its positive, neutral, or negative connotation
- Why did we choose it?
  - We hoped to see how the overall sentiment of messages correlated to the results of the games and the various match events

How We Did Sentiment Analysis

- Attempted to use Sentiment140’s API
    - Only supported English and Spanish tweets
        - This eliminated roughly half of our data, since we lacked language tags initially
        - Likewise, it reduced at least half of our tweets for some games, since Europe isn’t just English or Spanish
    - The model didn’t fit our data well
        - They only allowed specific classification for the “movie” category, otherwise it was applied to a general model

How Sentiment Analysis Failed Us





	text	user	created_at	timestamp	geo	lang	polarity
0	#SPAvENG stream on @ESPN3: https://t.co/wpiNN4...	Spurs_US	2015-11-13 20:34:41	1447443281485	NaN	en	2
1	Algo de bachateo ayer !!! @ Jerez De La Fronte...	juanpe_cuadrado	2015-11-13 20:34:41	1447443281977	NaN	es	2
2	[e24sn] International Friendly: Spain vs Eng...	LEGACYfied	2015-11-13 20:34:42	1447443282433	NaN	en	2
3	LIVE: Spain - England: Roy Hodgson's side trav...	Bot_Football	2015-11-13 20:34:42	1447443282915	NaN	en	2
4	come on England tonight let beat Spain #eng	fluffyForest21	2015-11-13 20:34:44	1447443284942	NaN	en	2

Some Examples of the Junk Positive Tweets

Hi! Good stuff! cheering on @MarcoRubio all the way from out here in England! Hope he's the next President of the US

Today is #WorldKindnessDay! Show someone you care with our Spa Perfect Relax and Rejuvinate Tote!

"Hi from Spain @MrsdogC is there some mail where i can make u few questions? I need some advice... Thanks a lot, a big hug!"

Just saw the @yogscast Skobbels Gameshow, and feeling a pinch of sympathy for Salt Films. Like watching England in Eurovision. @hat_films

Coleman Lay-Z-Spa 77"" x 28"" Inflatable Spa Portable 4-Person Hot Tub - Open Box

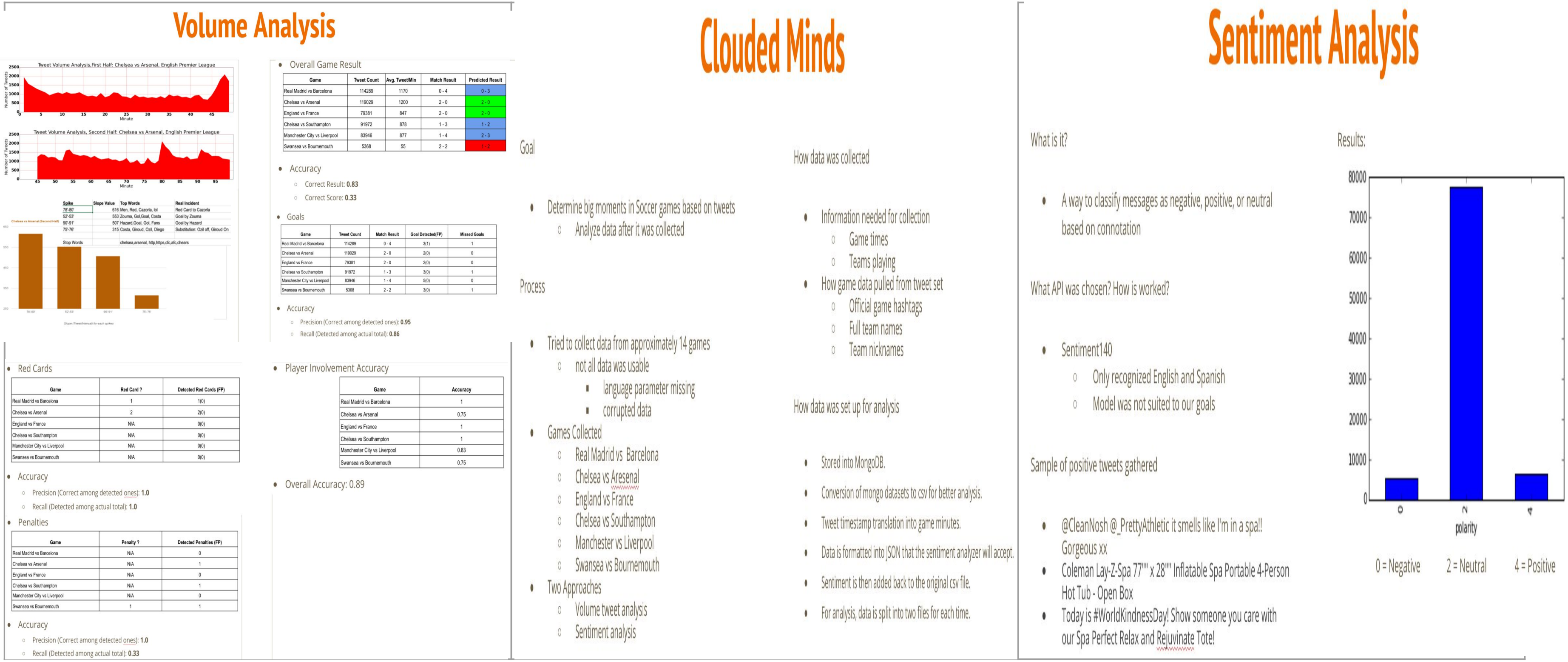
@england Da#### is Lalalalanananana & Delph playing, neither good enough. Id of also played Stones. So he can build a partnership with small

@CleanNosh @_PrettyAthletic it smells like I'm in a spa!! Gorgeous xx

How It Could Be Improved

- Noise Elimination
  - Attempt to cull advertising that was unrelated to the game
- Text Cleanup
  - Remove unnecessary ASCII characters that the sentiment analysis couldn’t account for
- Pick a better Sentiment Analyzer
  - Less generalized
  - More language support
- Automate the Volume Analysis through Machine Learning in real-time
- Data Redundancy

Poster



Tweet Volume Analysis,First Half: Chelsea vs Arsenal, English Premier League
Number of Tweets
0
500
1000
1500
2000
2500
Minute
0
5
10
15
20
25
30
35
40
45
Tweet Volume Analysis, Second Half: Chelsea vs Arsenal, English Premier League
Number of Tweets
0
500
1000
1500
2000
2500
Minute
45
50
55
60
65
70
75
80
85
90
95
Chelsea vs Arsenal (Second Half)
650
550
450
350
250
78'-80'
52'-53'
90'-91'
75'-76'
Slope (Tweet/Interval) for each spikes
polarity
0
10000
20000
30000
40000
50000
60000
70000
80000
0
2
4

Volume Analysis

Spike	Slope Value	Top Words	Real Incident
78'-80'	616	Men, Red, Cazorla, lol	Red Card to Cazorla
52'-53'	553	Zouma, Gol,Goal, Costa	Goal by Zouma
90'-91'	507	Hazard,Goal, Gol, Fans	Goal by Hazard
75'-76'	315	Costa, Giroud, Ozil, Diego	Substitution: Ozil off, Giroud On
Stop Words		chelsea,arsenal, http,https,cfc,afc,chears	

- Overall Game Result

Game	Tweet Count	Avg. Tweet/Min	Match Result	Predicted Result
Real Madrid vs Barcelona	114289	1170	0 - 4	0 - 3
Chelsea vs Arsenal	119029	1200	2 - 0	2 - 0
England vs France	79381	847	2 - 0	2 - 0
Chelsea vs Southampton	91972	878	1 - 3	1 - 2
Manchester City vs Liverpool	83946	877	1 - 4	2 - 3
Swansea vs Bournemouth	5368	55	2 - 2	1 - 2

- Accuracy
  - Correct Result: **0.83**
  - Correct Score: **0.33**
- Goals

Game	Tweet Count	Match Result	Goal Detected(FP)	Missed Goals
Real Madrid vs Barcelona	114289	0 - 4	3(1)	1
Chelsea vs Arsenal	119029	2 - 0	2(0)	0
England vs France	79381	2 - 0	2(0)	0
Chelsea vs Southampton	91972	1 - 3	3(0)	1
Manchester City vs Liverpool	83946	1 - 4	5(0)	0
Swansea vs Bournemouth	5368	2 - 2	3(0)	1

- Accuracy
  - Precision (Correct among detected ones): **0.95**
  - Recall (Detected among actual total): **0.86**

- Red Cards

Game	Red Card ?	Detected Red Cards (FP)
Real Madrid vs Barcelona	1	1(0)
Chelsea vs Arsenal	2	2(0)
England vs France	N/A	0(0)
Chelsea vs Southampton	N/A	0(0)
Manchester City vs Liverpool	N/A	0(0)
Swansea vs Bournemouth	N/A	0(0)

- Accuracy
  - Precision (Correct among detected ones): **1.0**
  - Recall (Detected among actual total): **1.0**
- Penalties

Game	Penalty ?	Detected Penalties (FP)
Real Madrid vs Barcelona	N/A	0
Chelsea vs Arsenal	N/A	1
England vs France	N/A	0
Chelsea vs Southampton	N/A	1
Manchester City vs Liverpool	N/A	0
Swansea vs Bournemouth	1	1

- Accuracy
  - Precision (Correct among detected ones): **1.0**
  - Recall (Detected among actual total): **0.33**

- Player Involvement Accuracy

Game	Accuracy
Real Madrid vs Barcelona	1
Chelsea vs Arsenal	0.75
England vs France	1
Chelsea vs Southampton	1
Manchester City vs Liverpool	0.83
Swansea vs Bournemouth	0.75

- Overall Accuracy: 0.89

Clouded Minds

Goal

- Determine big moments in Soccer games based on tweets
  - Analyze data after it was collected

Process

- Tried to collect data from approximately 14 games
  - not all data was usable
    - language parameter missing
    - corrupted data
- Games Collected
  - Real Madrid vs Barcelona
  - Chelsea vs Aresenal
  - England vs France
  - Chelsea vs Southampton
  - Manchester City vs Liverpool
  - Swansea vs Bournemouth
- Two Approaches
  - Volume tweet analysis
  - Sentiment analysis

How data was collected

- Information needed for collection
  - Game times
  - Teams playing
- How game data pulled from tweet set
  - Official game hashtags
  - Full team names
  - Team nicknames

How data was set up for analysis

- Stored into MongoDB.
- Conversion of mongo datasets to csv for better analysis.
- Tweet timestamp translation into game minutes.
- Data is formatted into JSON that the sentiment analyzer will accept.
- Sentiment is then added back to the original csv file.
- For analysis, data is split into two files for each time.

Sentiment Analysis

What is it?

- A way to classify messages as negative, positive, or neutral based on connotation

What API was chosen? How is worked?

- Sentiment140
  - Only recognized English and Spanish
  - Model was not suited to our goals

Sample of positive tweets gathered

- @CleanNosh @_PrettyAthletic it smells like I'm in a spa!! Gorgeous xx
- Coleman Lay-Z-Spa 77"" x 28"" Inflatable Spa Portable 4-Person Hot Tub - Open Box
- Today is #WorldKindnessDay! Show someone you care with our Spa Perfect Relax and Rejuvinate Tote!

Results:

0 = Negative 2 = Neutral 4 = Positive

Clouded Minds

Goal

- Determine big moments in Soccer games based on tweets
  - Analyze data after it was collected

Process

- Tried to collect data from approximately 14 games
  - not all data was usable
    - language parameter missing
    - corrupted data
- Games Collected
  - Real Madrid vs Barcelona
  - Chelsea vs Aresenal
  - England vs France
  - Chelsea vs Southampton
  - Manchester vs Liverpool
  - Swansea vs Bournemouth
- Two Approaches
  - Volume tweet analysis
  - Sentiment analysis

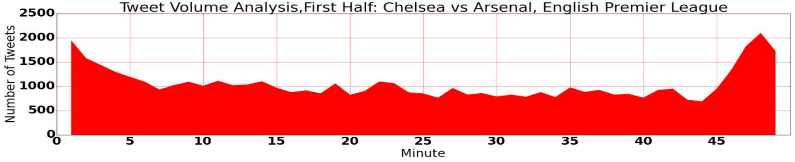
How data was collected

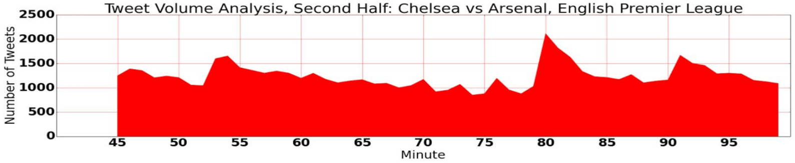
- Information needed for collection
  - Game times
  - Teams playing
- How game data pulled from tweet set
  - Official game hashtags
  - Full team names
  - Team nicknames

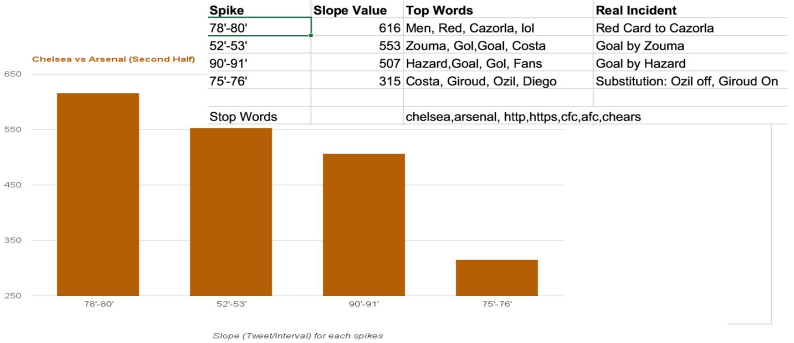
How data was set up for analysis

- Stored into MongoDB.
- Conversion of mongo datasets to csv for better analysis.
- Tweet timestamp translation into game minutes.
- Data is formatted into JSON that the sentiment analyzer will accept.
- Sentiment is then added back to the original csv file.
- For analysis, data is split into two files for each time.

Volume Analysis







Spike	Slope Value	Top Words	Real Incident
78'-80'	616	Men, Red, Cazorla, lol	Red Card to Cazorla
52'-53'	553	Zouma, Gol,Goal, Costa	Goal by Zouma
90'-91'	507	Hazard,Goal, Gol, Fans	Goal by Hazard
75'-76'	315	Costa, Giroud, Ozil, Diego	Substitution: Ozil off, Giroud On
Stop Words		chelsea,arsenal, http,https,cfc,afc,chears	

Chelsea vs Arsenal (Second Half)
650
550
450
350
250
78'-80'
52'-53'
90'-91'
75'-76'
Slope (Tweet/Interval) for each spikes

- Overall Game Result

Game	Tweet Count	Avg. Tweet/Min	Match Result	Predicted Result
Real Madrid vs Barcelona	114289	1170	0 - 4	0 - 3
Chelsea vs Arsenal	119029	1200	2 - 0	2 - 0
England vs France	79381	847	2 - 0	2 - 0
Chelsea vs Southampton	91972	878	1 - 3	1 - 2
Manchester City vs Liverpool	83946	877	1 - 4	2 - 3
Swansea vs Bournemouth	5368	55	2 - 2	1 - 2

- Accuracy
  - Correct Result: **0.83**
  - Correct Score: **0.33**

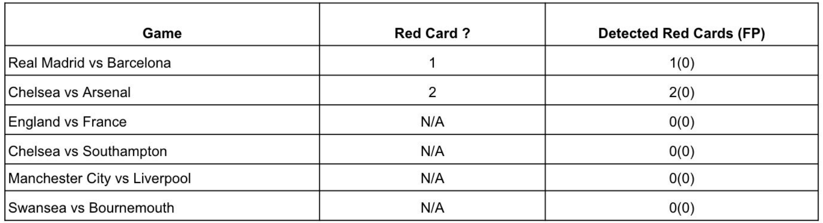
- Goals

Game	Tweet Count	Match Result	Goal Detected(FP)	Missed Goals
Real Madrid vs Barcelona	114289	0 - 4	3(1)	1
Chelsea vs Arsenal	119029	2 - 0	2(0)	0
England vs France	79381	2 - 0	2(0)	0
Chelsea vs Southampton	91972	1 - 3	3(0)	1
Manchester City vs Liverpool	83946	1 - 4	5(0)	0
Swansea vs Bournemouth	5368	2 - 2	3(0)	1

- Accuracy
  - Precision (Correct among detected ones): **0.95**
  - Recall (Detected among actual total): **0.86**

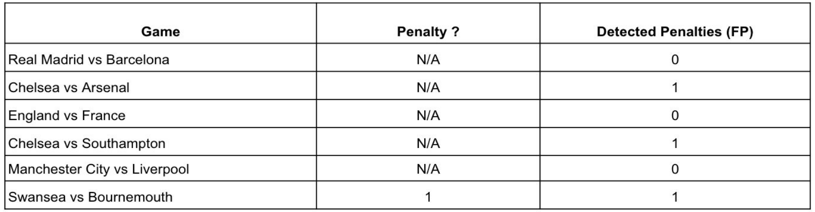
Volume Analysis

- Red Cards



Game	Red Card ?	Detected Red Cards (FP)
Real Madrid vs Barcelona	1	1(0)
Chelsea vs Arsenal	2	2(0)
England vs France	N/A	0(0)
Chelsea vs Southampton	N/A	0(0)
Manchester City vs Liverpool	N/A	0(0)
Swansea vs Bournemouth	N/A	0(0)

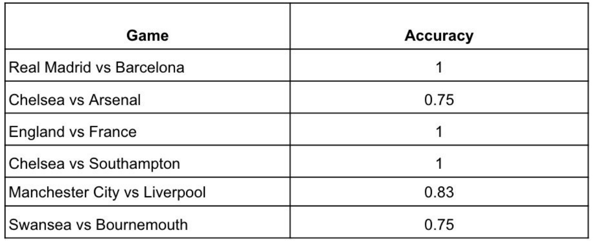
- Accuracy
  - Precision (Correct among detected ones): **1.0**
  - Recall (Detected among actual total): **1.0**
- Penalties



Game	Penalty ?	Detected Penalties (FP)
Real Madrid vs Barcelona	N/A	0
Chelsea vs Arsenal	N/A	1
England vs France	N/A	0
Chelsea vs Southampton	N/A	1
Manchester City vs Liverpool	N/A	0
Swansea vs Bournemouth	1	1

- Accuracy
  - Precision (Correct among detected ones): **1.0**
  - Recall (Detected among actual total): **0.33**

- Player Involvement Accuracy



Game	Accuracy
Real Madrid vs Barcelona	1
Chelsea vs Arsenal	0.75
England vs France	1
Chelsea vs Southampton	1
Manchester City vs Liverpool	0.83
Swansea vs Bournemouth	0.75

- Overall Accuracy: 0.89

Sentiment Analysis

What is it?

- A way to classify messages as negative, positive, or neutral based on connotation

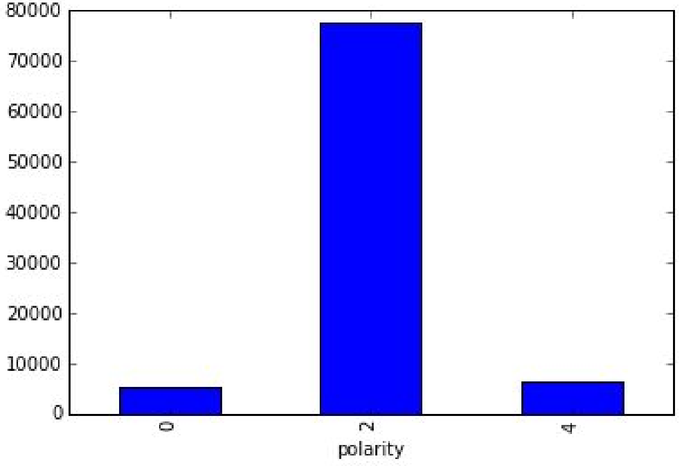
What API was chosen? How is worked?

- Sentiment140
  - Only recognized English and Spanish
  - Model was not suited to our goals

Sample of positive tweets gathered

- @CleanNosh @_PrettyAthletic it smells like I'm in a spa!! Gorgeous xx
- Coleman Lay-Z-Spa 77"" x 28"" Inflatable Spa Portable 4-Person Hot Tub - Open Box
- Today is #WorldKindnessDay! Show someone you care with our Spa Perfect Relax and Rejuvinate Tote!

Results:



0 = Negative　2 = Neutral　4 = Positive

Questions?