

Sentiment Analysis | Player Reviews

Gathered from players reviews on an online football forum.

NLP Project | 2016

Sneha Nanavati (201456243)
Sriharsh Bhyravajjula (201425019)

RedCafe

RedCafe is the most popular Manchester United football forum online. Frequented by 5,000-10,000 online users at any time, it consists of various 'threads' on players from the football club Manchester United. It has around 1620 average words per page and on an average 40 pages for each player in addition to having 40 posts per page. It provides a plethora of data on players all-round the football season, and has been active since 1999.

Structure of RedCafe

Sign Up

New users sign up

As part of the forum rules, the “newbies” are granted posting right to the major threads after making some basic contribution.

Post on threads

Once a full member

The users are allowed to post on any thread. Each thread has certain rules. Users also have to follow the general posting rules and instructions of the forum to keep posts structured and valuable.

Reply to posts


Follow up on posts

The users can reply to posts and carry on the discussions with follow up pointers.

URL

Average
Rating

www.redcafe.net/threads/juan-mata-2016-17-performances.419643/page-29

 **Juan Mata**
2016-17 Performances
Jump to: [Man Utd 1:1 West Ham](#)
[View full 2016-17 profile](#)

6.2	7.0	18	5	2	3
Average rating	Your rating	Appearances	Goals	Assists	Yellow cards

Page 29 of 30 < Prev 1 ← 25 26 27 28 29 30 Next > Threadmarks

Thread Tools Watch Thread

Saturday at 09:46 Report #1121

AN17
Full Member
Joined: Feb 13, 2015
Messages: 419
Location: Somewhere they can't find me.

Clever little player who knows when exactly to arrive in the box and finish off a move. Against most opposition he's more than adequate as an attacking option due to the sheer number of goals he gets and the kind of passes he sees which can cut open defenses.

+ Quote Reply

Saturday at 22:37 Report #1122

Loublaze
ATLien
Joined: Aug 31, 2009
Messages: 8,897

OnlyTwoDaSilvas said: ↑

Just realised he's scored more goals for United than he did for Chelsea, and in 9 fewer appearances. According to Transfermarkt.

Wow. He deserves to win a PL title with United

+ Quote Reply

User

Post
Numbers

Posts
in a
thread

Sample screenshot of a RedCafe page.

Goals

What we wish to achieve in this project.

RedCafe is the equivalent of a collection of reviews, albeit one for players. We wish to:

- Gather a huge corpus of player reviews.
- Manually mark sentiment value of a portion of data, creating training data
- Train a Naive Bayes classifier on this training data, and then classify the test data into sentiment categories.
- Compare classified sentiment vs trained sentiment, and analyze accuracy.

Data Gathering

We scraped data from the online forum, page by page. The data collected is an indication of the players' performances (and reception by supporters) over two football seasons (2014/15 and 2015/16).

Scraped Database Details

PLAYERS	PAGES	POSTS	WORDS
60	4668	185044	7565033

Average Words/Page : 1620.6

Average Words/Post : 40.8

Average Posts/Page : 39.6



Data Details | 2015-16 Season

36 players. 2045 pages. 80,605 posts. 3,132,945 words.

PLAYER	PAGES	POSTS	WORDS
Januzaj	59	2329	111385
Herrera	67	2643	101775
Pereira	34	1348	48343
Martial	169	6738	223577
Valencia	35	1379	52787
Young	20	763	29729
Schweinsteiger	115	4604	146129
CBJ	23	920	30426
Smalling	67	2680	95593
Blind	80	3192	144113
DDG	43	1192	48107

Love	2	69	2078
Varela	20	780	22305
Wilson	12	444	20499
Chicha	9	349	10752
Lingard	89	3524	144617
Riley	2	53	1572
Evans	3	114	3536
Mata	104	4136	174414
Shaw	28	1107	36525
Rojo	34	1337	38747
Rashford	53	2100	77444
Fellaini	107	4278	205116
Darmian	83	3308	113883
Memphis	190	7577	309005
Carrick	34	1326	46531

Carrick	34	1326	46531
Schneiderlin	73	2889	110998
Powell	8	292	10696
McNair	9	350	12737
Jones	24	926	33298
Johnstone	7	262	11136
Romero	26	1027	34448
TFM	30	1165	40734
Blackett	5	199	5053
Valdes	47	1846	62887
Rooney	334	13359	571970

36 players. 2045 pages. 80,605 posts. 3,132,945 words.

Data Details | 2014-15 Season

24 players. 2623 pages. 104,439 posts. 4,432,088 posts.

PLAYER	PAGES	POSTS	WORDS
Januzaj	104	4135	200961
Angel di Maria	237	9457	401607
Herrera	234	9325	369081
Pereira	37	1478	60640
Valencia	87	3449	154689
Young	80	3185	126325
Smalling	85	3367	129961
Blind	106	4243	176201
DDG	93	3689	130680
Fletcher	25	963	37453

Wilson	59	2360	95922
Evans	79	3139	142909
Mata	139	5560	269823
Shaw	80	3166	110997
Rojo	89	3526	125497
Fellaini	173	6922	356162
Carrick	43	1699	70411
McNair	57	2253	84693
Jones	62	2453	99571
Falcao	272	10852	445234

Rafael	80	3188	136850
RvP	107	4257	196944
Cleverley	86	3423	137947
Blackett	49	1958	80406
Rooney	160	6392	291034

24 players. 2623 pages. 104,439 posts. 4,432,088 words.

Manual Annotation of Data

After the data was gathered, we manually read through the first 350 reviews of three player datasets - **Januzaj_15-16**, **Rashford_15-16** and **DeGea_15-16**.

PLAYER	PAGES	WORDS	POSTS / REVIEWS	TRAIN/TEST POSTS RATIO
Januzaj	59	111,385	2329	350/1979
Rashford	53	77,444	2100	350/1750
De Gea	43	48,107	1711	350/1360

Manual Annotation of Data

Each user review was rated on a scale of 1 through 5, with each point based on the following rating criterion:

- 1 - Very Negative
- 2 - Negative
- 3 - Neutral
- 4 - Positive
- 5 - Very Positive



Naive Bayes Classification

In our context:

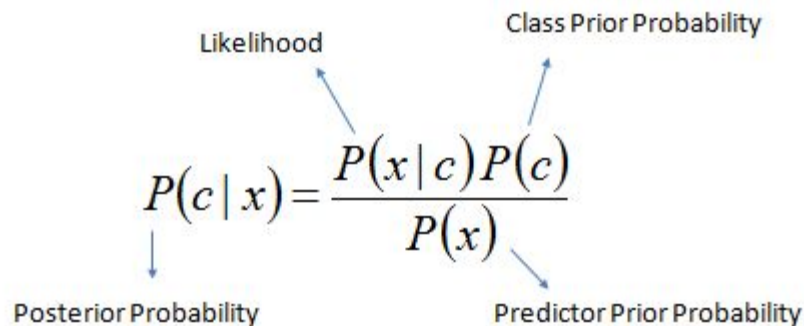
Posterior: $P(\text{sentiment} \mid \text{post})$

Likelihood: $P(\text{post} \mid \text{sentiment})$

Prior: $P(\text{sentiment})$

Evidence: $P(\text{post})$

$x_1, x_2, x_3 \dots$: Words in post



The diagram shows the Naive Bayes formula with arrows pointing from labels to the corresponding parts of the equation:

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Labels and their corresponding parts in the formula:

- Likelihood** points to $P(x \mid c)$
- Class Prior Probability** points to $P(c)$
- Posterior Probability** points to $P(c \mid x)$
- Predictor Prior Probability** points to $P(x)$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \dots \times P(x_n \mid c) \times P(c)$$

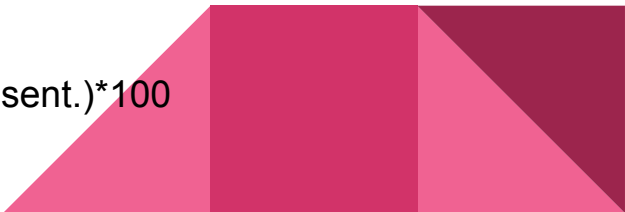
Results

The observed average sentiment and accuracy for each player:

PLAYER	AVERAGE SENTIMENT	ACCURACY
Januzaj	3.85	74.43
Rashford	3.95	87.21
De Gea	3.67	88.73

Where:

$Accuracy = 100 - ((Avg. \text{ classified sent.} - avg. \text{ trained sent.}) / avg. \text{ trained sent.}) * 100$

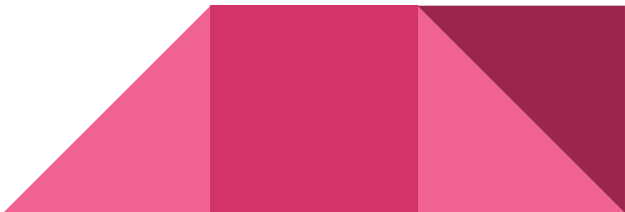


Observations and Conclusions

- The forum consists of United supporters, mostly. Hence there seems to be a positive bias for all three players.
- Despite the high accuracy percentages, this is a very rudimentary sentiment analyzer, since the POS Tagger (based on Penn Tree Bank, using the NLTK library in Python) did not deal effectively with image links, emoticons, tweets, media, et al.
- The sentiment analyzer does not deal well with sarcasm, references, comparisons, slang, et al.
- The annotated data should be in greater percentages.



Improvements Possible

- Making the parser sensitive to the forum posts, and using POS tags in context.
 - Developing an aspect based model which can understand football context and performance cliché expressions.
 - Using SentiWordNet or similar to give a default rating to words in the posts, rather than the default equal value to each word.
 - Understanding sarcasm, hyperboles, emoticon emotions, et al.
 - Greater size of manually annotated data for training.
 - Using tools beyond Naive Bayes.
- 



Thank You