Sentiment Analysis of Twitter on English Premier League Soccer

Robert P. Schumaker[1], A. Tomasz Jarmoszko[1], Chester Labedz Jr.[2] and David Freeman[2]

[1]Management Information Systems, [2]Management
Central Connecticut State University, New Britain, Connecticut 06050, USA
rob.schumaker@gmail.com, jarmoszko@ccsu.edu, labedzchs@mail.ccsu.edu and
freemand@mail.ccsu.edu

Abstract

Can the sentiment contained in tweets serve as meaningful proxy to predict match outcomes and if so, can the magnitude of these outcomes be similarly predicted based on the degree of sentiment?

Soccer wagering, like many other domains including the stock market, trades on publicly available information (i.e.; past performance and a knowledge of the participants). Therefore, assuming rational betting behavior, everyone should have a fair chance of achieving success. However, these markets can possess inequalities through withheld information, discounting of certain information or weighting it incorrectly. These informational inequalities lead to arbitrage opportunities that can be unlocked using data mining techniques.

One area of this publically available information is the sentiment contained within tweets. In 2010, Hong and Skiena studied the twitter feeds of the NFL and found Twitter to be a better predictor of game outcomes than experts. This surprising result was theoretically grounded in Surowiecki's idea that crowd wisdom can be a better predictor than individuals. To take Hong and Skiena's research further we seek to apply their technique to European Premier League (EPL) soccer and examine if the magnitude of directional sentiment (positive versus negative) is a good predictor of point spread (i.e.; if the tweets of one team are measurably more positive than the other team will that translate into a larger point spread and just how accurate is tweet polarity?)

To answer these questions we present the CentralSport system that uses the sentiment analysis of tweets to predict EPL outcomes. We apply our system to three months of data (120 matches) and analyze

the sentiment of tweets prior to each match. This corresponded into a rich dataset of 18,454,362 tweets from 2,224,895 unique tweeters with an average tweet length of 105.0 characters.

We first gather club-specific tweets, fixture results and betting lines for each match. From there we examine the sentiment characteristics of each tweet using the OpinionFinder tool which marks each tweet according to tone (objective versus subjective) and polarity (positive versus negative). Each fixture then has its sentiment results aggregated into eight different models based on tone and polarity combinations (1 – only subjective negative tweets, 2 – only objective negative, 3 – subjective positive, 4 – objective positive, 5 – only subjective, 6 – only objective, 7 – only negative and 8 – only positive) and compared against betting lines and fixture results.

We feel that this will yield interesting results and will grab the attention of MIT Sloan Sports Analytics Conference participants. Further extensions of this work include an analysis of post-match sentiment aligning with expectations given an outcome and a longitudinal analysis of tweet term usage.