Initial Report

Predicting association football match outcomes using social media and existing knowledge.

Student Number: C1148334
Author: Kiran Smith
Supervisor: Dr. Steven Schockaert
Module Title: One Semester Individual Project
Module Number: CM3203
Credits due: 40

Project Description

Association football (hereafter referred to as “football”), is generally agreed to be the world’s most popular sport [1], with over 270 million people actively involved with football [2] and as such the football betting industry in the UK was worth over £1.1bn in 2014 [3]. Specifically regarding sport, the creation of social media platforms such as Twitter has led to a new age of digital journalism, increasing the accessibility of the wider populations public opinions. This wider use of Twitter to express opinions is evidenced by the total of 32.1 million tweets published during the 2014 World Cup final [4].

We are increasingly able to use social media such as the aforementioned Twitter to identify trends and predict future activity [5]. By performing sentiment analysis on Twitter data, that is, analysis to determine the positive/negative sentiment of a tweet based on its text content and sentence structure, we are able to subjectively classify a tweet as positive or negative, and assign it a numerical value associated with this classification. Existing research has previously shown a correlation between tweet content sentiment and changes in stock markets [6].

This project aims to investigate the relationship between Twitter data, and the outcomes of football matches, using data generated from Twitter over a number of weeks, and related to British Premier League match outcomes. Taking into account existing probability models in this field, a model will be generated that will determine any correlation between sentiment and football match outcomes.

Project Aims

The aims and objectives to be completed within the scope of this project are listed below.

Core Objectives:

Betting odds retrieval: Retrieving odds for British Premier League football matches throughout the data collection period, listed below in the work plan. These odds will need to be either retrieved as part of a set of requests to a preferred single UK-based betting company, or through web scraping. These odds will then be plotted as a time-series graph to investigate how they change over time.

Twitter data retrieval: Retrieving data through the Twitter Streaming API throughout the data collection period, listed below in the work plan. These tweets will be retrieved using keywords relating to the 20 British Premier League football clubs, in order to retrieve the most accurate data.

Twitter data analysis: Using sentiment analysis tools to derive a quantitative value of a tweet’s positive/negative sentiment. This data will then be analysed and graphed alongside British Premier League football match outcomes, and the retrieved match outcome betting odds.

Explore existing models: Exploration and discussion of existing probability and statistical models for football matches that predict match outcome.

Design and implementation of model: Designing and implementing a model that can accurately and consistently predict the outcome of a football match using the aforementioned data and models.

Desirable Objectives:

Further data analysis: Taking into account football specific information to find terms that are frequently/infrequently associated with football match outcomes, to be incorporated into our model for future improvement. Additional analysis and exploration of the impact that sentiment analysis of Twitter data has upon betting companies football match outcome odds.

Advanced development of model: Development of an extensive model that is trained in a larger data set, incorporating data as it is processed, and can accurately and consistently bet and win against betting companies published odds using the aforementioned data and models.

Development of a web interface: Designing and developing a clean and attractive web interface to effectively display data and predictions for football matches.

Work Plan

The table below depicts how I expect to break down my aims and objectives into weeklong slots, and the timeframe in which I expect to achieve these aims and objectives. It has been agreed that one meeting will take place each week with my project supervisor.

Week 1 – w/c 26th January 2015:

- Initial Project Plan
- Initiate retrieval of betting odds from companies

Week 2 – w/c 2nd February 2015:

- Set up machine to store tweets.
- Initiate retrieval of tweets from Twitter streaming API.
- Begin further research of sentiment analysis tools.
- Continue to retrieve betting odds.

Week 3 – w/c 9th February 2015:

- Conduct literature review of existing or related works.
- Continue further research of sentiment analysis tools and finalise selection of tool.

- Continue to retrieve tweets.
- Continue to retrieve betting odds.

Week 4 – w/c 16th February 2015:

- Begin to filter and analyse initial Twitter data set.
- Continue to retrieve tweets.
- Continue to retrieve betting odds.
- Begin background research for the final report.
- Begin to design match outcome prediction model.

Week 5 – w/c 23rd February 2015:

- Continue to filter and analyse Twitter data set and initial frequent terms.
- Using initial Twitter data sets and sentiment analysis tool, and begin to further analyse data set.
- Begin writing final report (Introduction, Background, Approach).
- Continue to retrieve tweets.
- Continue to retrieve betting odds.
- Provisional interim review meeting with supervisor.

Week 6 – w/c 2nd March 2015:

- Continue analysis of Twitter data set using sentiment analysis tool.
- Continue writing final report (Introduction, Background, Approach).
- Analysis of existing models and their usefulness in this approach.
- Continue to retrieve tweets.
- Continue to retrieve betting odds.

Week 7 – w/c 9th March 2015:

- Continue analysis of Twitter data set using sentiment analysis tool.
- Continued analysis of existing models and their usefulness in this approach.
- Continue to retrieve tweets.
- Continue to retrieve betting odds.

Week 8 – w/c 16th March 2015:

- Collate and analyse all data sets. Categorise tweets into different time periods and different team names.
- Continue to write final report (Implementation, Results etc).

Week 9 – w/c 23rd March 2015:

- Continue to collate and analyse all data sets.
- Continue to write final report (Implementation, Results etc).
- Provisional interim review meeting with supervisor.

Easter recess: 30th March – 19th April 2015:

- This time period is a contingency period where the main focus will be on continuing to write the final report, however can be used for any over running work, if applicable.

- Should completion of all necessary objectives happen early, this time will be used to work on the desirable objectives to complete.

Week 10 – w/c 20th April 2015:

- Overflow period – a period of time, which will allow for any work that runs over from previously.
- Finalise writing of final report.

Week 11 – w/c 27th April 2015:

- Overflow period – a period of time, which will allow for any work that runs over from previously.

Week 12 – w/c 4th May 2015:

- Submission of final report by 5th May deadline.

References

[1] - Dunning, E (1999). *Sport Matters: Sociological studies of sport, violence and civilization*. London: Routledge.

[2] - Anon. (2007). *Big Count.* Available: http://www.fifa.com/worldfootball/bigcount/. Last accessed 30th January 2015.

[3] - Anon. (2014). *Industry Statistics.* Available: http://www.gamblingcommission.gov.uk/Gambling-data-analysis/statistics/Industry-statistics.aspx. Last accessed 30th January 2015.

[4] - Wiltshire, L. (2014). *The roar of the crowd for the #WorldCupFinal.*Available: https://blog.twitter.com/2014/the-roar-of-the-crowd-for-the-worldcupfinal. Last accessed 2nd February 2015.

[5] - Thelwall, Buckley and Paltoglou. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*. 62, p406-418.

[6] - Bollen, Mao and Zeng. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*. 2, p1-8.