Project Design Writeup

# Problem Statement & Hypothesis

The aim of this project is to determine whether the measurements of collective mood states derived from Twitter feeds are correlated to, and predictive of, the value of the FTSE 100 index over time. I will be collecting a daily grouping of a sample of tweets from the UK along with the sentiment and mood of these tweets, and the daily closing price of the FTSE 100 index. I will be using a Neural Network to predict the future value of the FTSE 100 index based on the Twitter mood, measuring the mean absolute percentage error (MAPE), the absolute percentage difference between the actual value and the predicted forecast. My hypothesis that twitter will be in some way predictive of movements in the FTSE 100 index. The benchmark is set in a paper by Johan Bollen et al. who focussed on predicting the movement of the Dow Jones Industrial Average (DJIA), and achieved 1.79% MAPE, and 86.7% direction accuracy.

# Assumptions & Questions

Some of the assumptions that I am making:

* The specific words that are tweeted by each user will matter in the model
* Tweets where the author has explicitly stated their mood (i.e. “I feel”, “I am feeling” etc.) are the best measure of the overall countries mood / sentiment
* The mood and sentiment of each tweet must be captured and grouped across the days
* To see correlations between sentiment and movement in the FTSE 100 we may have to lag the tweet data
* There may be some non-linear effects that need to be explored to better understand the relationships.

One of the questions which I am not so sure about:

* What to do about the market closing on the weekend. Do I drop weekends from the data? This seems a better solution than linearly extrapolating the weekend movement

# Motivation & Solution Use

Similar problems have been solved using Twitter mood to predict movements in the DJIA index. One of the motivations for revisiting this problem is to try and replicate the same problem in the UK, and to see how strong the IBM Watson Tone Anlayzer is in creating categories for time series forecasting.

The initial output will be an iPython notebook detailing the process to get to the results. If there is any significant predictive power in the model, I will work on operationalising the solution, and building an automated trading bot, or perhaps even better, selling the data to traders who can add it to their toolset.

# Step by Step Methodology

1. Collect a random sample of 25k tweets per day from the past 12 months using the Twitter API
2. Collect daily closing prices of the FTSE 100 index from the past 12 months using Pandas datareader to connect to Yahoo! Finance
3. Clean the data by removing the stop words and punctuation
4. Group all tweets that were submitted on the same date, only considering tweets that match the expressions, “I feel” etc., and tweets that don’t include hyperlinks
5. Use NLTK to find the positive vs. Negative mood from the text context
6. Pass the text to IBM Tone Analyzer through the API to get measurement of the different mood dimensions
7. Convert sentiment and mood to Z scores
8. Cross validate mood vs. sentiment by regression to show that sentiment is not enough to capture the public mood
9. Grainger causality of sentiment, mood, and lagged FTSE price vs FTSE closing price to test for causality
10. Passed the lagged variables through a Neural Network and measure the Mean Absolute Percentage Error, and direction accuracy
11. Test the prediction on test period, last 3 months

# Alternate Hypothesis

The problem could be rewritten as a classification problem; On any given day, is the Twitter sentiment predictive of the closing price of the FTSE 100 index increasing (1/0)?

# Alternate Goal Metric

The goal metric could be if we were to start with £100 in the bank, where would we be by the end of the test period?

# Data Dictionary

|  |  |
| --- | --- |
| date | Date |
| ftse | FTSE 100 closing Price |
| sentiment | The ratio of positive vs. negative tweets |
| anger | The level of anger predicted by IMB Watson |
| disgust | The level of disgust predicted by IMB Watson |
| fear | The level of fear predicted by IMB Watson |
| joy | The level of joy predicted by IMB Watson |
| sadness | The level of sadness predicted by IMB Watson |