|  |  |
| --- | --- |
| Name: | James Barrett |
| Supervisor(s): | 葉志立 Yip Chi Lap [Beta] |
| Dissertation Title: | Sequential Associations: Market reactions to large-scale global |
|  | events. |
| Planned Submission Semester: | Summer, 2017/18 |

**Aim**

|  |
| --- |
| Financial markets react to global events in different ways. I would like to identify various market’s reactions to large scale events. I will use natural language processing to identify unpredictable events and then classifying the events into broad concept hierarchies either natural disasters (for example earthquake or flood) or man-made disasters (for example, terrorist attack or plane crash).  Twitter has only been around since 2006 but on average in 2016 there were over 300,000 tweets per minute. Because of the short form of tweets, and the speed of dissemination throughout the network through retweets, Twitter as a platform is often considered to be a news media itself. If Twitter can be used as an aggregator of news sources, to quickly identify when a major world event has occurred, it can be used to give a trading signal.  Different events will affect the various asset classes in various ways, and sometimes on different time horizons. Major market index responses can be measured using popular indexes over short and medium time frames. Investment managers can use this information to help position portfolios’ asset allocation to best protect their client’s money, and profit from expected swings. |

**Brief Literature Review**

|  |
| --- |
| Even though Twitter has only been around since 2006, it has drawn significant interest from the academic community for its powers as a sentiment analysis tool. Different approaches have been attempted such as hashtags, parts of speech and emoticons(a). Kouloumpis et. al. found that hashtags and emoticons were useful approaches to collection of training data, while using parts of speech were not.  Kwak et. al. looked at the dissemination of headline information on Twitter, with a tweet once re-tweeted, quickly reaching 1000 retweets. The one caveat they mentioned is that there is some level of homophily.  Kireyev et. al. examined some of the challenges associated with using microblogging sites such as Twitter. Specifically, the lack of detail and the casual format are two of the challenges mentioned. In this project, using official accounts of news sites, many of the issues with typos and colloquialisms can be ignored as they are more likely to use formal language, and be proof-read for spelling errors and typos.  Keogh and Kasetty tested over 20 different data mining algorithms on over 50 different data sets with a view to verifying their utility. Their conclusion was that the data mining community needs to reduce the occurrence of implementation bias in their experiments and introduce objective benchmarks to verify algorithms.  Zhang and Zhou examined the application of various data mining techniques to the finance industry. Specifically, they looked at Stock market predictions, portfolio management, bankruptcy prediction, foreign exchange markets and fraud detection. They highlighted that the selection of appropriate variables and data mining algorithms, and model assessment and refinement are key components of this process. Also, although neural network modelling is the most widely used method in data mining applications in finance, the optimal design of neural networks for various financial engineering problems remains open. Whilst they didn’t provide suggestions for appropriate algorithms and neural network further research can be undertaken to find more recent research on the topic.   1. E. Kouloumpis, T. Wilson and J. Moore, 2011, Twitter Sentiment Analysis: The Good the Bad and the OMG!, In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* 2. H. Kwak et. al., 2010, What is Twitter, a Social Network or a News Media? 3. K. Kireyev, L. Palen, K. Anderson, 2009, Applications of Topics Models to Analysis of Disaster-Related Twitter Data 4. E. Keogh and S. Kasetty, 2003, On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration 5. Dongsong Zhang and Lina Zhou, 2004, Discovering Golden Nuggets: Data Mining in Financial Application, In *IEEE TRANSACTIONS ON SYSTEMS, Nov 2004* |

**Proposed Methodology**

|  |
| --- |
| The first part of the project will be to build a system to recognise when a critical mass of twitter posts has been made about an event. Twitter has an in-built Streaming API(a), that allows you to access all public posts and run analysis on them in real-time. These are fed in JSON format for easy read access in any programming language.  Using a methodology like the one described by Luce(b), I propose to classify the training set of tweets as natural disaster, man-made disaster or neutral. Given the volume of data on twitter, with over 300,000 tweets per minute, I will focus initially on registered accounts of news organisations globally, which are most likely to be reliable and fast sources of breaking incidents.  Using a list of already classified disasters from insurance companies(3) tweets from the news organisations shortly after the disaster can be collected and classified. If required, using the damage information provided by the insurance companies, the list can be filtered only on major disasters, or only in certain geographical areas with large public markets. Because there is not any existing data sets of classified data, I will need to manually classify the tweets collected.  After the dates have been established, market movements can be calculated over various short term time frames: such as 1d, 2d, 1w, 2w, 1m. I will attempt to gather data on FX, equities, fixed income and commodities. FX movement will need to analyse the local currency against a basket of major world currencies. Equities will look at local as well as regional market movements, and, where available, sector movements. Fixed income instruments will include corporate and government bonds, all along the yield curve. Commodities such as metals, energy, agricultural and livestock will all be analysed as well. Real estate will be excluded from this analysis in the absence of a sufficient market to measure.   1. https://dev.twitter.com/streaming/public 2. http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/ 3. http://media.swissre.com/documents/sigma2\_2017\_en.pdf |

**Milestones**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Start Date** | **Complete Date** | **Duration** |
| Background Reading & Literature Review | 20/Mar/17 | 01/Sep/17 | 120 |
| NLP | 01/Jun/17 | 01/Aug/17 | 30 |
| Deep Learning | 15/Jun/17 | 15/Aug/17 | 30 |
| Data Mining | 01/Jul/17 | 01/Sep/17 | 31 |
| Project | 01/Jul/17 | 01/Dec/17 | 110 |
| NLP System | 01/Jul/17 | 01/Oct/17 | 65 |
| Data Mining Analysis | 01/Sep/17 | 01/Jan/18 | 87 |
| Dissertation Webpage | 01/Mar/18 | 30/Mar/18 | 22 |
| Intention to Submit Dissertation | 01/Mar/18 | 01/May/18 | 44 |
| Poster Exhibition | 01/Jun/18 | 30/Jun/18 | 21 |
| Draft Dissertation | 01/Jan/18 | 01/Jul/18 | 130 |
| Oral Examination | 01/Jul/18 | 13/Jul/18 | 10 |
| Final Version | 01/Jul/18 | 01/Aug/18 | 23 |

**Deliverables**

|  |  |
| --- | --- |
| ***Items*** | |
| 1 | Sets of data for training, testing and system parameter tuning.   1. Twitter feed data to run NLP system over. 2. Market price data for price data mining. |
| 2 | NLP program to analyse twitter feed for key words. Possibly incorporating deep learning. |
| 3 | Market price data mining tool for assessing asset allocation strategy. |