Balazs Zankay: Investor Sentiment in Social Media
Project Summary

Project introduction

The project is about to review a possible method of investigating relationship between the price chart of QQQ (Nasdaq-100 Index tracking Exchange Traded Fund) and the sentiment change of Nasdaq related Tweets. The method is based on the tidytext R package developed by Julia Silge and David Robinson.

Background and Hypothesis

Background

Market participants are restlessly seeking Alpha since the very beginning of equity and commodity trading. Looking for methods, tools that would help them to beat the market, i.e. to outperform the competition. Sensing the sentiment of market was always in focus, as the relation between market's price actions and investor sentiment is a long known phenomenon. The hard part was always how to measure sentiment.

Interview based surveys - like the AAII (American Association of Individual Investors) sentiment survey - are very limited in terms of input data, and results are reflecting the past. Some publicly released data like the amplitude and pace of change in portfolio assets, or the volumes and proportions of equity and index options (put/call ratio) are also used for sentiment research, however the reasons behind are not always related to sentiment change.

The exponential growth of data storage and computing capacity in recent years has opened new paths to success: analysts can calculate and define things they had to estimate earlier. The technological development helped companies to implement new methods and processes to gain knowledge from data.

Meanwhile an other revolution is happening: the rapid spreading of social media. Communication channels of society has been changed dramatically worldwide. The number of social media users and the volume of released text had exploded in last few years, people are sharing their views, opinions online, supplying analysts with enormous volumes of unstructured text data.

Hypothesis

If we assume that there is an intersection of the set of people trading / investing in a certain equity and the set of people commenting / blogging about it, then we may assume that the following hypothesis might be true:

Social media reflects investor sentiment and therefore text based sentiment analysis is applicable as a market indicator.

Purpose of Project

The purpose of this project is to apply sentiment analysis on unstructured text of social media extraction, and to review it as a possible method that investigates the relationship between the price movements of certain market instrument and the sentiment of social media text related to this instrument. I have selected QQQ, the Nasdaq-100 index tracking ETF, as the instrument, and Twitter as the source of social media text flow.

Objects, tools and method

The NASDAQ Composite index

It is calculated from the price of the stocks and other securities listed on the NASDAQ stock market - the second-largest stock exchange in the world by market capitalization, and the first electronic stock exchange. It is one of the most important indices globally. Investors worldwide follow and trade derivatives and ETF based on the index. The composition of NASDAQ contains mostly information technology companies.

QQQ

QQQ is the Nasdaq 100 Index tracking exchange-traded fund. The 5 largest holdings of QQQ - Apple, Microsoft, Amazon, Facebook and ex-Google Alphabet - together account for almost half of the value of the ETF.

Twitter

Twitter is a San Francisco based, worldwide popular online social networking company, founded in 2006. It is providing news flow and application for posting and blogging for hundreds of millions of registered users. Twitter is one of the most popular social media company. Hashtags (#) are used to identify collections of tweets connected to the same conversations.

R Programming Language

R is a free software mainly applicable for statistical computing and visualization. The community of R users and developers is continuously expanding the software, implementing additional packages that makes R more efficient and easy to use

Tidytext

It is an R package that allows users to convert unstructured text into processable text data, and by that to adapt methods of data mining, analytics and visualization on natural language text processing. The power of this tool is briefly in transforming unstructured text into a tokenized, one token / row format, so that it can be processed as tidy data set.

Bing Sentiment Lexicon

Bing is a collection of English words (unigrams), categorizing them in a binary (either positive or negative) fashion, and calculates the sentiment of analyzed text by summing up the individual sentiment scores for each unigram in the text. Larger sets of text data are often broken up into equal sections, so that each section’s sentiment can be calculated, to gain a better overview.

Project Description

The project will be done in R. It starts with selecting and installing the necessary R packages (ready to use code units , that will save a lot of coding time).

Data pipeline

To get access to #nasdaq tweets, It is necessary to set up a data pipeline:

1. Creating a Twitter account. Twitter offers an application programming interface (REST API) for user applications with publicly accessible home page.
2. Creating website "ZankayAnalytics"
3. Setting authentication keys, allowing R studio to access data on Twitter.
4. Creating API token. This will launch the authentication

After successfully going through the 4 steps above, the data pipeline is live.

Setting search criteria

Setting search criteria and downloading #nasdaq tweets:

1. Hashtag (#) sign preceding the search criterion is used to identify tweets categorized into a specific topic. Setting q = "#nasdaq"
2. Restrict language to English: results will not contain non-English words and characters.
3. Twitter is limiting search results return rate to 18,000 every 15 minutes per user token. It is possible to exceed this number by setting the "retryonratelimit" argument to "TRUE", however it would make it very time consuming to run the code as Rtweet would then wait for rate limit reset, and then continue downloading tweets...
4. Setting timeframe: Rtweet package can return data only from the past 6-9 days via REST API, due to Twitter’s capacity limits. We can set a timeframe within this time period. There are companies collecting and selling social media the data, however in this project only freely available data is used.

After R finished collecting tweets, data is ready for viewing, manipulating and analyzing.

Viewing frequency

Takeaways after plotting tweeting frequency throughout the 7.-14. December period and comparing to QQQ price chart of same timeframe:

1. Daily volatility - highs during trading hours, lows outside trading hours
2. Higher average daily volumes on trading days, lower on weekend
3. Highest peaks of tweeting frequency coincide with turnarounds of short term trends on QQQ chart

Data Manipulation

1. Tidytext unnest_tokens function converts row, natural language tweets into a one-token-per-row, R compatible tidy data format by breaking the text into individual tokens (tokenization of text)
2. Anti_join(stop_words) will drop the most common but useless words from data.
3. Filter(!str_detect(word, "[0-9]") drops all words with numeric characters.

Reorder and Visualization of occurrence, Data cleaning

1. Counting the occurrence of every single word in data and visualize it on a column chart by reordering in descending rate of frequency.

2. Filtering words below a certain number of occurrence helps to concentrate on most frequent words.
3. Despite of dropping stop words with anti_join, the data is still full of useless words: brand names, abbreviations, codes, technical terms, buzzwords and other irrelevant expressions.
4. Stepping back to data manipulation phase, adding further filtering criteria to drop useless words: selecting from data manually one by one.
5. Filter value for frequency rate (of occurrence) can be set to lower value meanwhile, in accordance with recurring visualization of remaining words

The Sentiment Analysis

1. Words of the #nasdaq tweets dataset will be matched with words of Bing Sentiment Lexicon with inner_join and classified in binary (positive or negative in terms of sentiment) fashion. Inner_join operation will return only words of dataset with existing matching value in Lexicon
2. The "index = row_number() %/% 100" setting will result equally 100 row long sections through out the text flow in chronological order. Each section's total sentiment will be calculated by adding up the positive and negative words of the section. Each section will have a total sentiment value.
3. Visualization of #nasdaq tweets sentiment on bar chart: each bar represents the total sentiment value of 100 rows of tweets. The bars are in chronological order in the 07.-15. December time period.
4. Interpretation: The period starts with an overwhelmingly positive sentiment phase, but at a certain point it turns to deeply negative in a relatively quick pace. After that, the negative sentiment prevails for a longer period, sparsely mixed with a few slightly positive bars. Last quarter of charts shows the signs of recovery in sentiment again, but fails to stay in positive territory.
5. Comparing the formations on the sentiment and the QQQ price chart, there are some visible similarities: following a brief uptrend, QQQ price chart makes a sharp downturn on the 7. December. This downturn is probably reflected on the sentiment chart with the sudden sink of sentiment to negative territory. After a significant drop of price, correction starts on the 10. December, but seems to fail on the 11. December. Finally the upward correction continues and drives price chart to a new high in the timeframe. Probably this is the corresponding formation to sentiment's short living recovery to positive territory. Price chart's positive correction also ends at the final section and price turns to down too.
6. Repeating the comparison with a more focused sentiment chart on the 07.-09. December period. Following the previously seen sharp drop both in sentiment and in price on the 7. December, there is improvement of sentiment and slightly positive values at the end of chart. Since the end of chart is based on weekend dated tweets, this change of sentiment was not influenced by market price actions. Following the weekend the QQQ price made a bullish turn and significant appreciation. This may tell us that sentiment improvement had anticipated the positive turn on market next day.
7. Repeating the comparison with a sentiment chart focused on an other time period: 10. December-first part 11. December. The trading on the 10. December started with a short drop. After that price started to rise and closed near to daily high. Next day the trading started higher, leaving a gap in chart, but price of QQQ soon starts to fall at the end of

focused period. Sentiment chart of the timeframe also starts with negative values, then it turns to positive, but ends with negative sentiment again.

Conclusions

The reviewed analysis method has obviously limited abilities to reflect investor sentiment or to unveil the nature of relationship between sentiment and price action. Nevertheless we may make some assumptions:

1. Tweeting activity is likely higher when market volatility is higher, or when market is at a trend turning-point
2. Sentiment chart of tweets on QQQ and the price chart of QQQ in corresponding time period, often display similar reactions.
3. In certain cases the change in sentiment may predict according price action on market
4. Social media based sentiment analysis had proven the right to it's existence. Better tools and methods, more data may result really useful outcome.

Page intentionally left blank

Page intentionally left blank

Page intentionally left blank

Page intentionally left blank

Page intentionally left blank