# Commodity Market Sentiment: Capturing the Outlook of the Agriculture

Commodity Market.

This is an interim report on the current status and future development of the project.

## Introduction

Solving world hunger is the key responsibility of the Food and Agricultural Organisation of the United Nations (UNFAO). This long-term strategic goal and mission is an uncompromisable task to achieve a better prospect of the human society.

The many dimension of this complicated issue makes it one of the most challenging obstacle remains to be solved. Various inter-related factors such as production, price, waste, nutrition and politics are all important in ensuring the steady supply and distribution of food to those who are in need.

One of the most important determinants in resolving the issue is the stability of commodity price and ultimately the affordability of food. In this paper, we propose a market sentiment index aiming to capture the outlook and perception of the commodity market. This general perception will provide a better overview of the market as perceived by professional and traders around the world by mining the textual information and align it with the commodity price.

A market sentiment index has several potential applications. First of all, it provides an overview of the market not from a small group of experts but a comprehensive aggregated views based on knowledge and observations of professional, reporters and field analysts around the world. Secondly, the market sentiment also has the ability to predict the general trend of prices in the future and also the probability of a forthcoming food crisis.

In contrast to most literature related to text mining and sentiment analysis of financial markets, we do not focus on the predictability of price. Price prediction based on sentiments has a very short lifespan and generally has no predictive power more than a couple of days. This has no application nor value to the organisation where our goal is to ensure the long-term stability and survival of people. Rather, our focus is on the perception of the market based on the aggregated and continuous evolving perception of the market. The ultimate goal is to predict potential food crisis given the current status and future prospect of the market.

## Data

In order to make the connection between the text data in the news articles and tweets, several preparation steps are required.

In this section, we outline the necessary steps to prepare the data in order to perform the modeling.

### News and Tweet Scrapping

News articles are scrapped and obtained from various news agencies. Future work includes the integration of additional news outlet and also tweets from Twitter and StockTwits.

Standard text processing such as stemming, and removal of stop-words were performed to transform the data into usable format. Stemming and lemmatisation are two processes to reduce the inflected word into their word stem. For example, running is reduced to run and cars are reduced to car to avoid duplication of topics. Stop-words such as 'the' are removed as they contain little if no information.

A total of 140971 of articles were obtained and processed.

### Commodity Price

Five commodities has been chosen to be the focus of this study, namely wheat, maize, rice, barley and soybean.

The daily price of the selected commodities are obtained from the International Grains Council (IGC).

The data spans from 2000 to 2016 with a total of 4400 daily observations. Trading does not occur on weekends and Christmas.

## Methodology

Various steps of transformation and quantification are necessary for the extraction of essential and relevant information contained in the textual data.

In this section, we briefly describe each component of the project and the current status.

### Sentiment Extraction

In order to quantify the meaning and position of each individual news articles, a sentimental analysis is performed and each article is scored correspondingly.

Due to the fact that sentimental analysis is highly dependent on the financial position and the dictionary used, extensive research has been conducted to provide a sentiment analysis tool set which meets the need and position of the organisation.

Details of the research and analysis can be found in the [Sentence by Sentence Sentiment Extraction](https://github.com/EST-Team-Adam/TheReadingMachine/blob/Alberto/Sentiment_Extraction/sentence_by_sentence.pdf) paper.

An additional set of sentiments based on the Google Natural Language API is also extracted. This is used for comparison and benchmarking.

To capture the spread and retention of information, all sentiments are decayed exponentially. That is, a news article will have a sentiment score not only on the same day it was published but also all subsequent date to represent the permanent modification of the view on the market.

### Tagging and Topic Modeling

It would be naive to assume each piece of information as equally relevant and has the same effect on the perception of the market. Each article can contain information about a particular dimension such as trade, production and research about an individual commodity. The relevance of the piece of information with respect to the prices are different, and thus, it is important to classify the document with respect to the dimension.

To account for the asymmetry of the information, tagging and topic modeling were performed to further derive additional information from the textual data. Two dimensions of tagging were performed, adding geographical and commodity dimension to the sentiments.

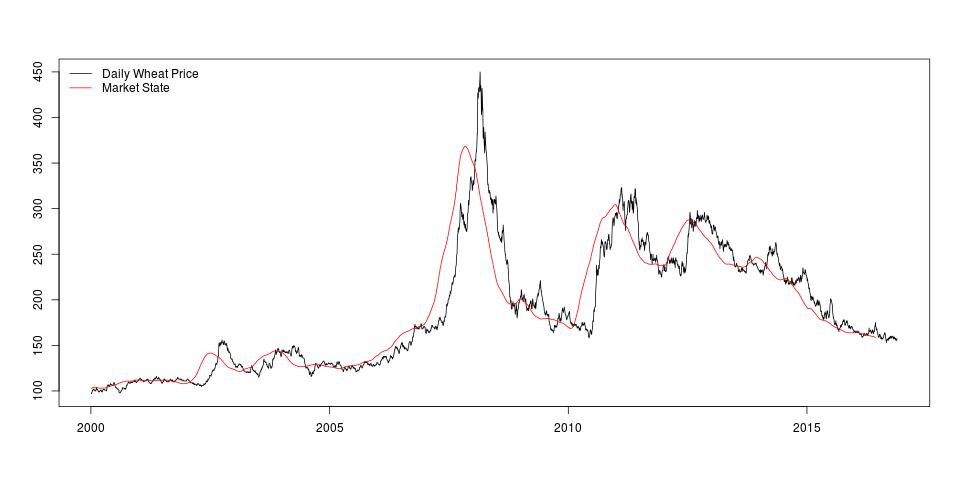
The topics are extracted based on the method non-negative matrix factorisation. The details of the method is also outside the scope of this overview, please refer to the topic modelling documentation for a detailed description.

### Market State

Instead of modeling the commodity price directly, a state-space model was fitted to the time series in order to extract the state of the market.

The purpose of this procedure is two-fold. As we are concerned with the long-term stability of the food prices, the fluctuation of daily price bears little importance to the analysis. The model smoothed the series and enables the analysis to focus solely on the level and trend of the price. Secondly, the derived state leads the price time series, thus, rendering predictive power to the price. If a connection can be made between the textual information and the derived state, then we can forecast the general trend of the price in the future.

Shown below are the price of wheat and one of the estimated state. From the graph, we can see the state is a slow transition time series which lead the movement of the price time series.



### Cumulative Sentiment Index

Each individual article contains a piece of information that changes the perception of the market. This continuous change need to be integrated in order to capture the dynamic of the market. To do this, we take the cumulative sum of the sentiment to capture this nature.

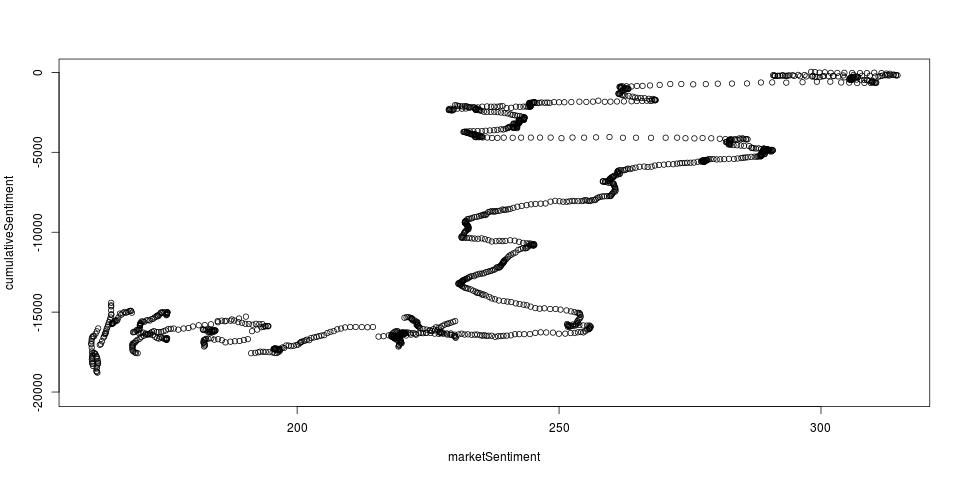
The cumulative sentiment index is simply defined as the continuous summation of the sentiment index.

S\_T = sum\_1^(t<T)(s\_t)

That is, we sum up all the sentiments prior to the current date for each observed date.

## Exploratory Analysis

Before we begin with the modeling, we examine the relationship between the market sentiment extracted from the wheat price time series and the cumulative article sentiment obtained from the news articles.



The scatter plot shows clear positive correlation between the market sentiment and the cumulative article sentiment over long period of time. That is, when the cumulative article sentiment increases, the market sentiment also increases which also lead to an increase in the price.

The correlation between the two time series is 0.77 and if we were to perform a naive regression, we would obtain a significant model with R-squared of 0.597.

However, knowing this is a time series, the relationship is dynamic and this can be observed that the strength of the correlation varies over various segment of the time series. This suggest models with dynamic coefficient is required.

Furthermore, we have not yet incorporate the relevance of the sentimental index according to the article relevance using the score obtained from the topic modeling. This should further improve the correlation and provide more predictive power.

## Model

After deriving all possible information from the textual data, a model is required to make the connection between the market information and the market sentiment.

In the previous section, we explore the raw relationship between the market sentiment and the cumulative sentiment obtained from the news articles. In this section, we propose the following model in order to incorporate the relevance of each article with respect to the price and modify the sentiment in order to improve the performance and the predictive power of the cumulative sentiments.

The following model to estimate the effect of each individual piece of information. The cost function is defined as follow:

C = sum(ms\_i - y\_i)

= ms\_i - sum(d\_i \* s\_i)

= ms\_i - sum(|tw\_i - tw\_hat| \* s\_i)

Where **ms\_i** is the market sentiment filtered from the price series, **tw\_i** are the individual topic score of each article and **s\_i** is the sentiment of the article. The **tw\_hat** are the coefficients to be estimated, they represent the topic of the price time series. Essentially, the **d\_i** represents the relevance of the particular dimension on the price of the commodity and the sentiments are scaled based on the relevance.

After obtaining the coefficients **d\_i**, we can use it to modify the raw sentiment and create the adjusted cumulative sentiments.

S\_adj,T = sum\_1^(t<T) (d\_i,t \* s\_i,t)

The formula is identical to the previous cumulative sentiment index except each sentiment **s\_i** is now weighted by **d\_i** or the relevance of the piece of information.

The adjusted cumulative sentiment index will have a even higher correlation with the market sentiment shown above.

## Conclusion

In this paper, we have demonstrated the potential of utilising textual information available in the public domain to capture the general perception and outlook of the market. Further, we shown that the predictability of the market sentiment may provide us the ability to detect events such as food crisis in the future.

In addition to the main objective described in the introductory section, there are many uses to each and every product that was produced in the process of this project.

The derivation of the sentiments for each article can not only be used to predict the future trend of the commodity prices, we can also use it to identify major events happening around the world in real time that requires the attention of the organisation.

Tagging and topic modeling give us the possibility to build our own information retrieval system where similar event or relevant information to be obtained more readily.

## Future Works

The following is a list of items to be worked on.

* Further expand the set of textual data available publicly.
* Continue on the improvement of sentiment index, further, investigate the potential of multiple sentiment index for various position in the market.
* Incorporate the information derived from topic modeling, and weight the relevance of each piece of article accordingly.
* Implement a dynamic linear model to account for the varying changes in the relationship between the market sentiment and cumulative individual sentiment.