

Twitter as an Indicator of Food Security

Master Thesis

Alexander Büsser

July 26, 2015

Advisors: Prof. Dr. Karl Aberer, Prof. Dr. B. Egger
Department of Computer Science, EPFL

To my family and friends

Abstract

This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes.

Contents

Contents	iii
List of Figures	1
List of Tables	1
1 Introduction	2
2 Literature Review	4
3 Social Media Data Acquisition	6
3.1 Hyperspace Analogue to Language	7
3.1.1 Motivating a Semantic Approach	7
3.2 Food Lexicon	8
3.2.1 Candidate Food Term Selection	9
3.3 Predictor Lexicon	10
3.3.1 Candidate Predictor Term Selection	11
3.3.2 Annotation of HAL Terms	11
3.3.3 Annotation Results and HAL Precision	12
3.4 Filtering	14
3.4.1 Food Related Tweets	14
3.4.2 Predictor Related Tweets	15
3.4.3 Sentiment Extraction	15
4 Twitter Data Analysis	16
4.1 Quantitative Analysis	16
4.1.1 User Distribution	16
4.1.2 Food Term Distribution	16
4.1.3 Price Correlation	17
4.2 Qualitative Analysis of Conversation Drivers	20
4.2.1 A Visual Analysis of the Social Attention	21
4.2.2 Event Detection	22
4.2.3 Event Annotation	23
4.2.4 Results	23
4.2.5 Discussion	25
5 Forecasting Model	26
5.1 System Overview	26
5.2 Input Model	27

5.2.1	Features	27
5.2.2	Feature Selection	29
5.3	Data Preprocessing	32
5.4	Training the Model	33
5.5	Time Series Model	33
5.5.1	Motivating a Fuzzy Approach for Time Series Modelling	33
5.5.2	Fuzzy Variables and Fuzzy Sets	34
5.5.3	Fuzzy Inference System	34
5.5.4	Adaptive Neuro Fuzzy Inference System	35
5.5.5	Optimization	36
5.6	Experimental Evaluation	37
5.6.1	Prediction Task and Evaluation	37
5.6.2	Evaluation of Smoothing Function and Determining Lag Factor	38
5.6.3	Price Model Results	39
5.6.4	Social Media Model Evaluation	41
5.6.5	Mixture Model	44
5.6.6	General comments to the Results	44
6	Conclusion	47
6.1	Future Work	47
A	Crowdfunder	49
A.1	Categorise: Food Price	49
A.2	Categorise: Food Supply	50
A.3	Categorise: Food Poverty	51
A.4	Categorise: Food Needs	52
B	Price Correlation	53
C	Time Series Modeling	55
	Bibliography	58

List of Figures

3.1	Framework of the Social Media Data Acquisition	6
3.2	Food Lexicon - Hierarchy	9
3.3	HAL Evaluation	13
4.1	User Distribution	17
4.2	Volume of Tweets per Keyword and per Category	17
4.3	Heatplot Meat: Volume of Tweets per Keyword and per Category	19
4.4	Heatplot Cereals: Volume of Tweets per Keyword and per Category	19
4.5	Topic Distribution - Food Security	22
4.6	Annotation Distribution - Relevant to Food Security	25
5.1	Forecasting - System Overview	27
5.2	Characteristics of Commodity Price Data	32
5.3	Fuzzy Variables and Fuzzy Sets - A Colour Object Example	34
5.4	Caption for LOF	36
5.5	Volatility of Social Media Features	38
5.6	Prediction Accuracy - Benchmark	40
5.8	Benchmark Prediction Beef	41
5.9	Benchmark Prediction Milk	42
5.11	Social Media Prediction	44
5.12	Social Media Prediction	46
B.1	Heatplot Dairy: Volume of Tweets per Keyword and per Category	53
B.2	Heatplot Sugar: Volume of Tweets per Keyword and per Category	54
B.3	Heatplot Oil: Volume of Tweets per Keyword and per Category	54
C.2	Social Media Prediction	55
C.1	Social Media Prediction	56
C.3	Social Media Prediction	57

List of Tables

3.1	Toy example of HAL	7
3.2	A Summary of the Evolution of our Food Lexicon	10
3.3	Categorisation of Predictor Lexicon	11
3.4	Keywords of Predictor Categories	14
4.1	Price Correlation	20
5.1	Input Model: Benchmark Prediction	28
5.2	Input Model: Social Media Prediction	29
5.3	Feature Selection: Benchmark Prediction	30
5.4	Feature Selection: Benchmark Prediction	31
5.5	Feature Selection: Benchmark Prediction	31
5.6	Parameter Settings	38
5.7	Sentiment Correlation	39

Chapter 1

Introduction

Despite living in a highly developed world, food security is still a prevailing issue. Around 842 million people are estimated to be experiencing malnourishment and hunger ¹. Especially global soaring food prices seem to have a negative effect by transmitting into rising food inflation rates in domestic markets [45]. The consequences of rising food prices differ among developed and developing countries. Developing countries struggle with accessibility due to poor infrastructure and affordability as most of their income is spent on food. Developed countries on the other hand face increased malnourishment and as a result an increase of health-related diseases such as obesity or diabetes ². The diversity between and within different countries calls for an improved approach in effectively tracking long-term development trends to aid the development of safer policies and ultimately a world where everybody has enough food.

Food security assessments are typically done through household surveys, which are timely and expensive to execute [13]. It takes years to analyse, validate and release. As a result, it is mostly an exercise in history, they fail to provide real time information which hinders an early response [24]. Today however we live in a world where multimedia content is rapidly increasing [48], especially as people continuously use social networks for communicating and documenting their lives. More often than not social media provides a rich source of media content such as images and links to news [48]. As people use those services they leave traces in the data and if their lives change for the better or the worse we see those patterns change as well [15]. As a result social media successfully mirror the well being of our society [49].

Twitter is considered to be one of the largest social media platforms and among one of the most consistent and prevalent topics on the platform is food ³. Cleaned and aggregated this provides a valuable opportunity for *studying food security indicators on Twitter*. Previous research [24] and commercial applications ⁴ in this area have focused on visualising the data but do not provide an in-depth investigation of the forecasting capabilities. Understanding if and how food security objectives are perceived by the population and whether the information is predictive of future price changes or even indicative of the next food crisis is the focus of this work. This dissertation contributes to the community by providing a fully automatic forecasting approach for food prices.

¹<http://www.fao.org/news/story/en/item/204232/icode/>

²<http://www.fao.org/docrep/u9920t/u9920t07.htm>

³<http://www.businessinsider.com/most-discussed-topics-on-social-media-2013-5>

⁴<http://foodsecurity.dupont.com/2014/10/09/the-need-for-a-common-language-to-support-global-food-security/>

First, and most central to this dissertation we try to understand how social media users talk about food security indicators. This inherently comes down to finding contextually similar words to our given food security objectives. Most frequently a local co-occurrence analysis is performed [1], however in this work we make a case that this approach fails to capture a lot of indicative terms [11]. Hence, we model the language used in Twitter to refer to food products and food security indicators by using a semantic space of words co-occurrences (§3.1) and propose a set of metrics that favour the semantic analysis of social media text (§3.3.3). Ultimately, with this analysis, we seek to distinguish between relevant and irrelevant discussions on social media. The analysis of the retrieved tweets showed that raw products (e.g. *corn*) are only very sparsely represented in Twitter conversations (§3.2). To circumvent this problem we use a mixture of manual annotation and automated vocabulary expansion to increase the coverage of our lexicon (§3.2, §3.3).

In a second study we investigate whether social media captures useful signals about food price fluctuations. To address this questions we performed a large-scale analysis of 29M Tweets distributed over 15.5 million users. We investigate to what extent the volume of food discussions can be correlated to the international Food Price Index and commodity price quotes (§??). We found that on an aggregated level (e.g. *meat*) no real correlation exists however on a finer granularity (e.g. *sirloin steak*) certain products exhibit a strong linear relationship of up to 0.7369 with a significance of 0.0001. Considering this evidence we apply a methodology introduced by [3] to automatically detect peaks in volume of the twitter data and investigate the relevance to our desired food security objectives. The results showed that up to 13 % of the peaks can be attributed to discussions around food supply, price and poverty (§5.6.5).

In a third analysis we seek to understand if it is possible to use social media to model commodity prices. Being able to forecast price trends is of great importance as it is widely assumed that the food crisis in 2008 was accelerated by speculative actions on the commodity future markets [18]. The prices of U.S. corn tripled from \$94 to \$281 during a period of only three years. With a Pearson correlation of 0.8436 and a significance of 0.0001 there is a strong correlation between the Food Security Index and the commodity prices making it an important indicator. We hence monitor volatile commodity prices by using an Adaptive neuro fuzzy inference system for time series modelling (§5.5). We show that social media features by itself are not informative enough of explaining the price variance. However, coupled with price data we are able to accurately predict a trend four weeks into the future (§5.5).

Overall this dissertation demonstrates the potential value of a large scale twitter analysis for monitoring food security. We believe there are a range of applications for governmental institutions, healthcare and the private sector alike. In the next Chapter 2 we present similar work to ours. In Chapter 3 we perform an investigation of the word semantics and detail a framework for creating our lexicons. In Chapter 4 we perform a correlation analysis between the tweet volume and the international Food Price Index. Chapter ?? details the time series modelling of different commodities. Lastly, in Chapter 6 we conclude our findings and give directions for future work.

Chapter 2

Literature Review

Relevant work was performed in the fields of economics, social development and the study of dietary patterns framed to the task of predicting real world quantities through text data. We review each of them and lastly detail work similar to ours.

Following the two recent food crises in 2007 and 2010 a lot of research in economics has been centered around defining causes of volatile food prices [23] [34]. Getaw Tadesse et. al. [45] defined a taxonomy for drivers of international food price spikes and differentiated among three different causes namely exogenous shocks, conditional causes and internal causes. Examples of exogenous shocks are extreme weather events, oil price shocks, economic and demand/supply growth, and lastly economic shocks. Conditional causes can originate through political conflict or market conditions. Internal causes on the other hand are speculative activities (driven by price expectations) and declines in world food stocks. Contrary to our work most of these studies try to find causal links between official statistics of oil inflation rates, supply growth and the food price mostly neglecting public sentiment or some quantifiable measure of current affairs (e.g. news). Through the use of Twitter we try to capture the above mentioned indicators and extend on previous studies by investigating the effect of public sentiment on volatile food prices.

Studies on social development investigated how twitter users react during a food crisis and if social media could be used as a tool for food risk communication. Rui Gaspar et. al. [19] performed an extensive study on Twitter communication during the 2011 European EHEC outbreak. They investigated the response of a population when coping with food-related threats. Their findings showed that most people were seeking for information and showed a high degree of uncertainty but also anger was expressed in regards of identifying the contaminated products. Peter Rutsawert et. al. [41] analysed Social Media as a tool for risk communication. Most notably they examine the risks and benefits of such a medium. The experts interviewed for this study mostly agreed that social media is a valuable tool for food crisis communication as it allows to speed up correspondences and provides a platform for interaction. Furthermore they found it to be helpful in understanding public concerns. The most noted objection was that social media might act as an amplifier and escalate a food crisis. Recently a new medium for monitoring social development has emerged. With a mobile phone coverage of more than 90 % in most developing countries mobile phone data has shown to be a valuable indicator for socio-economic problems. Adeline Decuyper et al. [15] correlate the airtime credit purchases to the expenditure on vegetable. People that spend more money on their mobile phones also spend more money on food. This was shown to be true with a Pearson correlation of over 0.8.

Our work falls into the broader scope of predicting real world quantities through text data. A lot of work has focused on predicting movie box-office revenues due to its high attention on

social media. Sitaram Asur et al. [6] showed that twitter outperformed market-based predictions for box-office revenues through a regression model based on twitter volume and sentiment of reviewers. Further examples followed for elections. Andranik Tumasjan et al.'s [47] prediction results were very similar to traditional election poll's accuracy by simply counting the occurrence of the candidates in tweets. Current work has focused on predicting economic and market trends by leveraging social media [57] [14] [43]. While this work generates a lot of attention in the community its applications for trading purposes remain unknown [32]. Similar to above work we use the attention and sentiment of twitter to explain the variance of a known quantity but juxtapose the problem to the topic of food security to understand how to use words in a specific context to predict a measurable quantity.

Natalie Dixon [16] build the foundation for relating tweets about food to known quantities. She analysis global food consumption patterns and tries to understand how certain items are perceived by the population and how the consumption of food differs between countries. In her work she makes assumptions about how food about twitter could be predictive for official GDP statistics and obesity levels. The later was investigated by Sofiane Abbar et al. [1]. By assigning specific nutritional values (calories) to each food item mentioned in a tweet they aggregate the caloric intake per state and correlate it with the official obesity and diabetes rates. A strong correlation was observed of $r = 0.77$ for obesity and $r = 0.66$ for diabetes. This paper makes a strong case on the potential value of twitter as a national surveying tool. Further studies on dietary patterns were performed by Robert West et al. [51]. Web logs of receipes were investigated to study the dietary preferences of a population. They found a significant correlation between the sodium intake displayed in those logs and the number of congestive heart failures reported at the hospital.

Only a few studies have been performed for food and price. Victor Chahuneau et al. ?? studies how the semantics of menus affect the price. In other words what kind of language is indicative of a cheap or expensive menu. They further explore the relationship between the sentiment of a restaurant reviews and the menu price. For instance they showed that menus containing words indicative of locally-grown food are more expensive then menus containing words appealing to tradition (old time favourite, real, fashioned). Lastly, the work by the UN [24] which is most similar to ours illustrates how Twitter reflects changes in food prices. In their analysis the tweet load was correlated to a time series of food prices. Observations showed that there is a linear relationship between food and tweet increase. One of the main conclusions was that despite a strong correlation domain knowledge is needed as speculations sometimes cause the public to panic causing false positives.

Our work extends the above analysis to a range of different foods not only considering the raw materials but also resulting products which are more present in social media. Lastly we do not make any hypothesis about what could have caused the attention spikes but perfume a detailed investigation of the content causing the fluctuations in twitter volume. Although the UN showed the potential in twitter as an indicator of food security no actions have been taken to predict real world quantities.

Social Media Data Acquisition

The topic food is discussed in many contexts on social media. To distinguish the relevant from the irrelevant Tweets we follow a framework illustrated in Figure 3.1. Although food security objectives are well defined it is a challenging task to filter Tweets that fall into such segments. We hence motivate and provide an overview of a semantic analyser HAL [33]. After filtering Twitter data with an initial list of food related terms (e.g. *rice*, *wheat*, *milk*) we proceed to use HAL to increase the number of food related keywords that eventually build our food lexicon. In a second round we collect Twitter data corresponding to the food lexicon and use it as a corpus to study the semantics of the terms price, supply, poverty and needs. We perform the analysis on a number of different random samples and experiment with different configurations of the algorithm HAL. The resulting keywords of the semantic analyser are presented to crowd workers which accept or reject terms based on a set of instructions. Finally the accepted terms (e.g. *pricey*, *cheap*, *available*) build the predictor lexicon which aims to capture Tweets that are relevant to food security. We downloaded 2 TB of Tweets from the internet archive ¹ over a span of October 2011 - September 2014. The filtering process detailed in the last section of this chapter resulted with 29 M food relevant Tweets which corresponds roughly to 4.5 % of the entire sample.

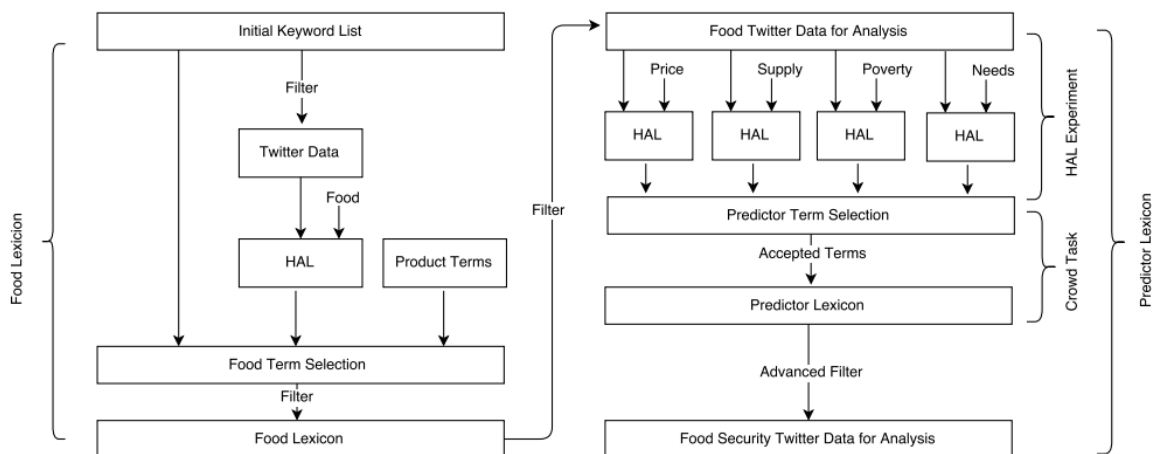


Figure 3.1: Framework of the Social Media Data Acquisition

¹<https://archive.org/details/archiveteam-json-Twitterstream>

3.1 Hyperspace Analogue to Language

“HAL creates a semantic space from word co-occurrences”². By using a sliding window parsing mechanism, the frequency of each term co-occurring within a fixed window size is recorded. It is important to note that HAL only records the terms before the word we wish to analyse the context from. The terms after the word will appear in the column in the matrix that corresponds to that word. The matrix is created by storing a vector for each word with the number of co-occurrences of every other word in the corpus. Hence, if our corpus contains N different words the resulting HAL space would be a square matrix of co-occurrences. Every time a specific word appears in the fixed window size the co-occurrence vectors are updated. For each co-occurrence, HAL applies a scoring function. Words that appear closer receive an inversely proportional score to its distance.

To illustrate the idea [11] gives an example of a simple sentence *The horse raced past the barn fell.* in Table 3.1 with a sliding window of five. Let us consider the first row. *The* precedes *Barn* twice. Once within a distance of five and the other time it directly precedes the word *Barn*. Hence, that cell receives a score of five for the proximate one and a score of one for the word further away resulting in a final score of six.

Following the creation of the matrix we concatenate both the column and row vector of a word, where the former represents the preceding words and the later the following. To compare the distance of the vectors we used the cosine similarity function.

	Barn	Fell	Horse	Past	Raced	The
Barn			2	4	3	6
Fell		5	1	3	2	4
Horse						5
Past			4		5	3
Raced			5			4
The			3	5	4	2

Table 3.1: Toy example of HAL

3.1.1 Motivating a Semantic Approach

HAL allows us to study the relationship between words. More specifically it is an algorithm that aids our goal of understanding what words are represented in the context of *food* and topics targeted around *food security*. To achieve this target we need a methodology for representing the meaning of a word. We analyse the context of a word to identify new words that have a similar meaning or given an identical context express the same thing. The later is concerned with identifying synonyms where as the former looks at the contextual similarity. For example, let us look at the word *mould* and *tons*. Those two words seem unrelated, but given the context of food they express the same thing. Namely an abundance of food. Burgess and Lund [11] motivate that through the context, they possess elements of item’s similarity but by themselves they would never be considered words with similar meaning. Burgers and Lund further note that the words are not similar because they occur frequently locally, but because they occur frequently in similar sentential context. They further argue that a simple local co-occurrence analysis misses to capture a lot of relationships. For example the word street and road are basically synonyms, however they seldom locally co-occur. Road and street do however occur in the same context. Those observations motivated us to deviate from the

²<https://code.google.com/p/airhead-research/wiki/HyperspaceAnalogueToLanguage>

commonly used co-occurrence analysis and take a step further to improve the precision of our filtering framework.

3.2 Food Lexicon

We began the construction of our Food Lexicon by considering a simple list of food related keywords. To avoid ambiguities we will refer to the initial list of keywords as $K_{initial}$. Words included are the most common traded food commodities as listed by IMF ³ along the ten most important staple foods that feed the world ⁴.

We filtered the archive dataset using exact string matching on $K_{initial}$. The distribution of the food related Tweets motivated us to structure our lexicon hierarchically as certain commodities were only represented very sparsely and insufficient for further analysis. Where global keywords such as *food* are highly represented, more specific keywords such as *beef* occur infrequently. To circumvent this problem, we mimic the hierarchical representation of the Food and Agriculture Organisation of the United Nations (FAO) ⁵.

FAO tries to measure the overall food fluctuation by five different food categories namely *meat*, *dairy products*, *cereals*, *vegetable oil* and *sugar*. We further created a category named *Other Food of Interest*. This category contains general keywords (e.g. *food*, *dinner* or *lunch*) and food keywords that cannot be assigned to one of the five categories, but frequently occur (e.g. *coffee*, *tea*). To be considered frequent, the set of Tweets containing the keyword needs to be $> 1\%$ of the total sample. *meat*, *dairy*, *cereals*, *vegetable oil*, *sugar* and *other food of interest* build the top layer of our hierarchical representation as shown in Figure 3.2 (only a subset of the lexicon is displayed).

For the second layer we use subcategories. As the name implies subcategories abstract the categories into different subsets i.e. for *meat* we would have the subsets *beef*, *chicken*, *lamb* and *pork*.

As the third layer and lowest instance, we consider food products. Each subcategory consists of food items which i.) can simply be the name of a category and subcategory (e.g. *meat*, *beef*) or ii.) be a product that is commonly found in markets and stores around the world. An example of the later would be *flour* for the subcategory *wheat*. The intuition and motivation to include such products is simple. Firstly, we were motivated to include such terms because products are more likely to capture the social attention than raw items due to their every day use. For our second argument let us consider the production of the product pasta. The variable cost factors are the raw materials flower and egg, machinery maintenance and the cost of the packaging. Where maintenances are performed on a regular predictable basis the raw materials plastic, egg and flower are market driven. ⁶. If the price of the raw materials increase then it is very likely that the price of pasta will increase as well. Products should hence be just as expressive in explaining the variance of food prices. One however has to be cautious as certain producers hedge themselves against price fluctuations of commodities allowing them to sell the product to the same price despite rising commodity prices.

³<http://www.imf.org/external/np/res/commod/index.aspx>

⁴<http://knowledge.allianz.com/demography/health/?767/the-worlds-staple-foods>

⁵<http://www.fao.org/worldfoodsituation/foodpricesindex/en/>

⁶[http://www.wikinvest.com/stock/FomentoEconomicoMexicanoS.A.B.deC.V.\(FMX\)/IncreasespricesRawMaterialsIncreaseCostofFoods](http://www.wikinvest.com/stock/FomentoEconomicoMexicanoS.A.B.deC.V.(FMX)/IncreasespricesRawMaterialsIncreaseCostofFoods)

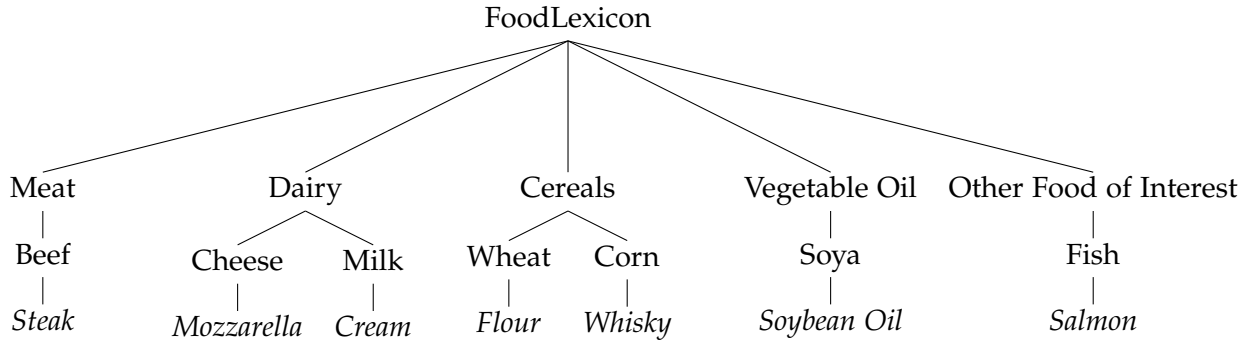


Figure 3.2: Food Lexicon - Hierarchy

Other than the sparsity of the data we further faced the problem of ambiguous keywords. *Soy* is such a keyword that refers in English to the *bean* and in Spanish to the verb *to be*. To avoid such ambiguity we extended the term to make it distinct (e.g. *soy* → *soy bean*). Terms were added to the lexicon by following a framework as explained in the following section.

3.2.1 Candidate Food Term Selection

We initially assume an empty set K_{final} and structure it hierarchically as mentioned in the previous section. The six categories ($c_1, c_2 \dots c_6$) are $\in K_{final}$ where c_i is one of the six categories mentioned above. For interpretability purposes we introduce an axiom in form of a set K_{all} . It only contains five of the above mentioned six categories *meat*, *dairy products*, *cereals*, *vegetable oil*, *sugar* excluding the category *other food of interest*. We assume that K_{all} is a fully populated lexicon containing all possible food items for a specific category (e.g. the subset *dairy* would contain all possible dairy products). It returns *True* if a term is an element of the set and *False* otherwise. For all keywords $k_i \in K_{initial}$ we evaluate if $k_i \in K_{all}$. If *True* we add k_i to K_{final} . For all keywords $k_i \notin K_{all}$ the condition of it being frequent is evaluated and if *True* we add it to the category *other food of interest* c_6 in K_{final} . Infrequent terms were discarded (e.g. *orange*, *cocoa*, *onion*). Lastly, the set K_{final} was further enriched by using food terms of products (e.g. *pasta*, *flower*), let's call those terms p_i , that have been identified by [1]. Again we only add p_i to K_{final} if it is clearly a meat, dairy, sugar, oil or cereal product or in other words an element of K_{all} . To further improve our coverage of the six food categories we filtered for synonyms and contextual similar words using HAL.

We summarize our framework as follows:

- 1.) Add all keywords $k \in K_{initial}$ to K_{final} only if $k \in K_{all}$ or k is frequent
- 2.) Include all p_i to K_{final} only if $p_i \in K_{all}$
- 3.) Create a HAL space using a random subsample of 10% from $K_{initial}$ with all keywords that occur > 100 . $\forall c_i \in K_{final}$ pick the keyword $k \in K_{final}$ that most frequently occurs over the entire sample and retrieve the top 500 terms with the highest vector similarity in the HAL space. The authors hand select those that are $\in K_{all}$ and add it to K_{final} .

Finally the resulting set of keywords in K_{final} concludes our food lexicon. All the accepted terms along with their source are illustrated in Table 3.2.

Lexicon / Subset s	Keywords (i: from initial set, e: from [1] , h: from HAL space)
$K_{initial}$ Food	meal (i), meals (i), food (i), foods (i), wheat (i), rice (i), maize (i), carley (i), soybean (i), soy (i), meat (i), beef (i), cattle (i), chicken (i), poultry (i), lamb (i), swine (i), pork (i), fish (i), seafood (i), shrimp (i), salmon (i), sugar (i), bananas (i), oranges (i), coffee (i), cocoa (i), tea (i), milk (i), yams (i), cassava (i), potatoes (i), sorghum (i), plantain (i), nuts (i), onion (i), salt (i), egg (i), dairy (i), cereals (i)
K_{final} Meat	meat (i), lamb (i), pork (i), swine (i), chicken (i), poultry (i), beef (i), sausage (e), rib (e), pastrami (e), kidney (e), liver (e), ham (e), bacon (e), chorizo (e), salami (e), sheep (e), boeuf (e), oxen (e), kine (e), steak (e), cow (e), brisket (e), veal (e), tenderloin (e), sirloin (e), poulet (e), volaille (e), hot dog (h), hamburgers (h), meatballs (h), burgers (h), goat (h), cattle v, turkey (h), pig (h)
K_{final} Cereals	wheat (i), atta (i), starch (i), farina (i), bran (i), ethanol (i), biofuel (i), rice (i), corn (i), maize (i), ravioli (e), barley (e), scotch (e), whisky (h), oat (h), bread (h), flour (h), gluten (h), pasta (h), noodles (h), beer (h)
K_{final} Oil	coconut oil (i), corn oil (i), olive oil (i), palm oil (i), peanut oil (i), sunflower oil (i), rapeseed oil (i), safflower oil (i), soybean oil (i), sunflower oil (i), soybeans (i), soya (i), soy sauce (i), soja (i)
K_{final} Sugar	sugar (i), sugarcane (i), syrup (e), energy drink (e), cola (e), chocolate (e), nestle (e), cookies (h), cupcakes (h)
K_{final} Dairy	dairy (i), egg (i), milk (i), kefir (e), butter (e), yogurt (e), quark (e), mozzarella (e), cheddar (e), parmesan (e), buttermilk (e), ricotta (e), feta (e), romano (e), provolone (e), colby (e), edam (e), eggnog (e), pimento (e), cheshire (e), roquefort (e), icecream (h), milkshake (h), cheese (h), cream (h)
K_f Other	meal (i), meals (i), food (i), foods (i), fish (i), prawn (i), seafood (i), salmon (i), tea (i), coffee (i), dinner (h), lunch (h), breakfast (h), dish (h), cuisine (h)

Table 3.2: A Summary of the Evolution of our Food Lexicon

3.3 Predictor Lexicon

From our food lexicon K_{final} we proceeded to extract features that we can use to explain events around food security and later use for our price prediction task. We structured our predictor lexicon into categories that capture the main food security objectives. The FAO measures food security based on four dimensions [39] namely *Access*, *Availability*, *Stability* and *Utilisation*. Where *Availability* mostly captures the supply of food, *Access* is concerned with the affordability of the basic goods. *Utilisation* captures the nutritional value of food to indicate whether the basic food needs are met. Lastly *Stability* is a measure of the economic strength.

To assign relevant keywords to the above categories we aimed to build on previous work. However, for some categories such *Availability* to the best of our knowledge no terms have been proposed in previous research. Furthermore we found [24] attempt to capture Access limited as they simply used the term price. Similarly, only a limited number of keywords were proposed by [42] to capture the category *Stability*. On the other hand we found that measuring *Utilisation* by aggregating caloric values for each food term mentioned in a tweet as in [1] to be out of the scope of this dissertation. In [4] they propose to use the keywords *price*, *supply*, *poverty* and *needs* to capture the above mentioned categories. We try to improve upon [4] by using synonyms that are derived from a corpus of tweets rather than from an

official dictionary. Table 3.3 concludes the new definitions we derived from [4]. We now proceed to detail how we select appropriate terms for each single category using HAL.

FAO	New Terminology
Access	Price
Availability	Supply
Utilisation	Needs
Stability	Poverty

Table 3.3: Categorisation of Predictor Lexicon

3.3.1 Candidate Predictor Term Selection

Our goal of this section is to capture terms that accurately describe the four food security objectives *price, supply, poverty and needs*. As illustrated in Figure 3.1 we run four HAL instances, one for each food security objective. To increase the precision of HAL we evaluated the performance on three different sample sizes (10 %, 20 %, 40 %) constituting a corpus of around 23M, 47M, 93M words from food related Tweets respectively. Our samples are particularly suited for studying the meaning of words. Not only do they cover a wide vocabulary targeted around food but they further give a close approximation of every day speech. This is very different from normal corpora which are usually based on business reports or specialised dictionaries [11].

The initial set of words in our corpus was filtered only to contain those words that appear at least 100 times. Words occurring infrequent were discarded as well as stop words and punctuations. On a test sample of 10 % we observed that around 10 % of the Tweets contain equal or less than four words which could impact the precision of the retrieved keywords as they rarely contain any useful information [38]. Hence, on the 40 % sample we further excluded Tweets that contain less or equal to four terms. Using the words in the Twitter corpus we produced a $N \times N$ matrix with the co-occurrences for three different window sizes namely five, eight and ten to investigate if the window size has an impact on the precision of HAL. According to [33] a window size of eight should yield the best results. However, the nature of a tweet is very different from a classical text so it remains to see if this observation holds for microblogs. Since vector similarity measures are sensitive to the magnitude of the vectors we normalized all the vectors to a constant length.

Each HAL instance retrieved 500 terms. Since for every sample we performed three experiments with three different window sizes for each category the result is a list of 18000 terms.

3.3.2 Annotation of HAL Terms

To determine whether a keyword is indicative of supply, price, poverty or needs we hired workers from the crowdsourcing platform crowdflower⁷ to annotate the HAL results. To perform the annotations we presented the workers with four different tasks, one for each food security objective. For every task, we asked the workers to classify the term as **A. Relevant**, **B. Likely**, **C. Unlikely** and **D. Not in English**. Since overlaps may occur, particularly between the category price and supply as well as poverty and needs we asked the workers to classify those ambiguous terms as **B. Likely** in order to detect to which category the word has a stronger association. We added the instruction along with the four different tasks to the Appendix A.

The crowd task presented a number of challenges. In our first test run, we counted a false positive rate of around 40 %. This was due to the lack of quality control we imposed on

⁷ <http://www.crowdflower.com/>

the workers. We observed a large amount of random guesses and a poor level of English among some workers. Hence, we selected workers from commonwealth countries and regions where the majority are native English speakers. We further created test questions which were manually selected to avoid inattentive workers. The test questions were designed to be unambiguous so that we could classify workers whose result significantly differed from ours as untrusted. We chose the threshold to be 70 % which followed the recommended guidelines of the platform. Lastly, we collected three independent annotations for every word and applied a majority vote to resolve disagreements. Due to the imposed additional costs through the multiple annotations per term we restricted our search for relevant keywords to the top 140 terms suggested by HAL.

3.3.3 Annotation Results and HAL Precision

From the crowdsourcing platform we present some basic statistics. For each annotation task it took a worker on average (measured via the Interquartile Mean IQM) five seconds to complete the job. Every worker received on average 0.2 cents per question and an average agreement of 84 % was observed among all crowd tasks (measured on a sample of 100 rows for each crowd task). The agreement factor serves as a quality measure for the design of the questions. A low agreement factor would mean that the questions are subjective, difficult or ambiguous as noted by the platform.

We manually assessed the annotations produced by crowdflower to check for disagreements between the crowd workers and ourselves. For the category supply we rejected 26 from 69 (39 %), for price 4 (12.5 %) from 32, for needs 8 (7 %) from 113 and for poverty 14 (13.2 %) from 106.

The high disagreement for the supply category was most likely due to the ambiguous design of our question in the crowd task. We asked workers to accept words that can be both indicative for supply and price (e.g. *rise*, *high*) which unfortunately was misunderstood as to include words that can be only indicative of price (e.g. *expensive*). If we however exclude this task we conclude as in [38] that crowd workers provide reliable results for large tasks that are hard to automate.

Similar to [3] we observe that crowdsource annotators applied a more narrow definition of the predictor categories overlooking some keywords associated with the categories. For example the term *market* was missed as a price keyword. Tweets containing the word *market* could provide valuable information regarding the state of food security as it is commonly used to describe the price mood of a commodity⁸.

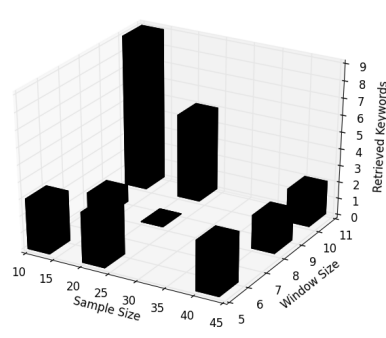
We turn our attention to Figure 3.3 that measures the overall number of relevant keywords (relevance was determined by crowd workers) retrieved for a given configuration (i.e sample size, window size). We observe that for all categories HAL had the highest precision with a window size of 10 which contradicts the findings of [33]. This is rather surprising as Tweets are characterised as being short hence we would have expected a smaller window size to be representative for a Tweet. Statistics about the English language show that on average a word consists of 5.1 characters and 14.3 words [7]. Tweets on the other hand have an average length of 28 characters so roughly 5.5 words⁹. Although it has been shown that the average tweet is relatively short there is a high peak around the maximum of 140 characters constituting a tweet length of around 27 words. Following this investigation it is not surprising that the window size of 5 and the windows size of 10 show the best performance where the later

⁸Nine out of the ten most relevant tweets retrieved on www.topsy.com for the keyword wheat and market contained information about the price mood of the commodity

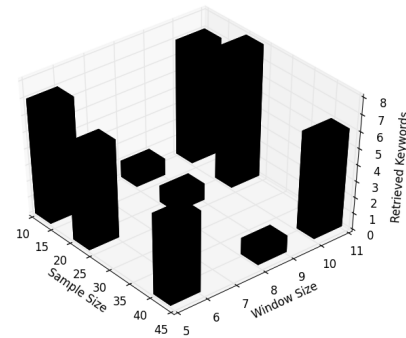
⁹<http://thenextweb.com/twitter/2012/01/07/interesting-fact-most-tweets-posted-are-approximately-30-characters-long/>

successfully captures the short Tweets and the former is representative for the large number of long Tweets. Additionally, we see that the smaller sample sizes (10%, 20%) consistently produce more relevant keywords than the large sample size (40%). A large sample size increases the likelihood of a keyword's occurrence. Since we set a fixed threshold of 100 occurrences across all samples we are more likely to include words with a smaller confidence as the sample size increases, which might explain the poor performance. This might further explain why HAL's vector similarity measure for terms retrieved in the smaller samples are consistently higher than the ones from the large sample. For poverty we recorded an average similarity score of 0.3121 for the 40% sample 0.7970 and 0.7822 for the 20% and 10 % sample respectively. Similar values were observed for the category needs namely 0.6636 for 40% , 0.8972 for 20 % and 0.8874 for the 10 % sample. For supply the values were 0.4187 for 40%, 0.6621 for 20% and 0.5939 for 10% sample. Lastly 0.49563518 for 40%, 0.8660 for 20% and 0.8708 for the 10% sample. There is an intuitive relationship between the number of accepted terms and HAL's vector similarity measure. Terms with lower vector similarity are more likely to be rejected.

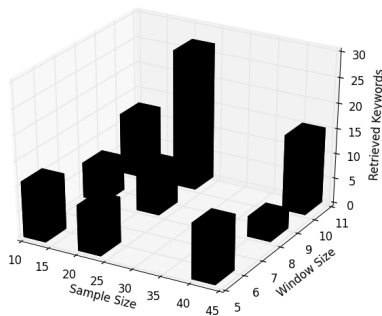
To conclude this chapter, terms of the final predictor lexicon together with the source are presented in Table 3.4. The final lexicon was further enriched by including synonyms from thesaurus¹⁰ for supply, need, poverty, and price.



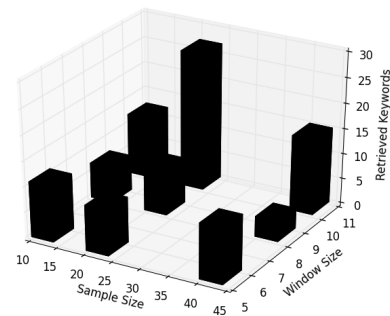
(a) HAL - Price



(b) HAL - Supply



(c) HAL - Needs



(d) HAL - Poverty

Figure 3.3: HAL Evaluation

¹⁰<http://www.thesaurus.com/>

Lexicon / Subset s	Keywords (h: from HAL space, t: from thesaurus)
<i>Food Supply</i>	<i>supply</i> , item (h), stock (h), vendors (h), demand (h), provided (h), feeds (h), delivery (h), supply (h), industry (h), production (h), waste (h), source (h), stash (h), numbers (h), list (h), growing (h), stores (h), distribution (h), delivered (h), policy (h), purchases (h), market (h), processing (h), chain (h), packaging (h), network(h), mart (h), stalls (h), sustainability (h), aplenty (t), bags (t), bulk (t), bundle (t), chunk (t), expanse (t), extent (t), flock (t), chunk (t), gob (t), heap (t), hunk (t), jillion (v), load (t), lot (t), magnitude (t), mass (t), meassure (t), mess (t), mint (t), mucho (t), oodles (t), pack (t), pile (t), scads (t), score (t), slat (t), slew (t), ton (t), volume (t)
<i>Food Price</i>	<i>price</i> , affordable (h), cost (h), rise (h), savings (h), coupons (h), prices (h), label (h), purchase (h), economy (h), discount (h), budget (h), sales (h), benefit (h), target (h), bonus (h), size (h), money (h), better (h), best (h), free (h), buy (h), amount (t), bill (t), , demand (t), estimate (t), expenditure (t), expense (t), fare (t), fee (t), figure (t), output (t), pay (t), payment (t), premium (t), rate (t), return (t), tariff (t), valuation (t), worth (t), appraisal (t)
<i>Food Poverty</i>	<i>poverty</i> , appetite (h), rich (h), shelter (h), homeless (h), shortage (h), control (h), provide (h), feed (h), needy (h), edible (h), nutrition (h), donate (h), expensive (h), economy (h), thought (h), budget (h), poor (h), service (h), supplies (h), crisis (h), demand (h), poverty (h), pantry (h), cravings (h), agricultural, resources, assistance, insecurity, storage (h), issue (h), bank (h), safety (h), prices (h), funding (h), health (h), drug (h), challenges (h), distribution (h), helping (h), government (h), affected (h), scraps (h), fair (h), children (h), support (h), waste (h), program (h), crops (h), restrictions (h), parcels (h), industry (h), healthcare (h), culture (h), catering (h), delicious (h), writer (h), sustainability (h), revolution (h),inflation (h), policy (h), daily (h), bankruptcy (t), debt (t), deficit (t), difficulty (t), famine (t), hardship (t), lack (t), scarcity (t), shortage (t), starvation (t),underdevelopment (t), abundance (t), affluence (t), bounty (t), myriad (t),plenty (t), plethora (t), profusion (t), prosperity (t), riches (t), wealth (t)
<i>Food Needs</i>	<i>need</i> , must (h), loving (h), share (h), like (h), favourite (h), hate (h), ordering (h), eat (h), give (h), much (h), want (h), needs (h), takes (h), beg (h), iwant (h), getting (h), favorite (h), buy (h), 50thingsilove (h), enough (h), ilove (h), whatilovethemost (h), got (h), horrible (h), cookout (h), poor (h), ate (h), deliver (h), neeeeed (h), loooooove (h), neeed (h), neeeed (h), make (h), good (h), 2thingsilove (h), lack, tweetyourweakness, terrible, bring, ineed, lots (h), waiting (h), bit (h), starving (h), gave (h), delicious (h), drink (h), nice (h), cook (h), hungry (h), craving (h), healthy (h), wish (h), awesome (h), really (h), best (h), dearth (t), deficiency (t), drought (t), inadequacy (t), insufficiency (t), lack (t), need (t), omission (t), privation (t), unavailability (t), void (t), want (t),affluence (t), bounty (t), myriad (t), plenty (t), plethora (t), profusion (t), prosperity (t), riches (t), wealth (t), ampleness (t), copiousness (t), fortune (t), opulence (t), plentitude (t), prosperousness (t)

Table 3.4: Keywords of Predictor Categories

3.4 Filtering

The filtering of the Tweets was performed in three rounds. First we filtered for relevant food Tweets. In a second round, we applied our predictor lexicon on the retrieved set of Tweets obtained in the first step. Lastly, we filter by sentiment. The first two filtering steps are illustrated in Figure 3.1 and are labeled as *Filter* and *Advanced Filter* respectively.

3.4.1 Food Related Tweets

The food related Tweets were retrieved through exact term matching, i.e. a tweet containing the term *foods* would not match on the keyword *food* where the reverse is also true. We mimic the term matching Twitter performs. In the initial round we optimized for coverage and hence avoided further filtering steps. The result was a collection of 29 M Tweets posted by 4.2 M users.

3.4.2 Predictor Related Tweets

The first round drastically reduced our dataset to around 90 GB of Tweets. With a smaller dataset we were able to perform a more involved filtering mechanism similar to [4].

For every word in a tweet and every word in our predictor lexicon the stem was computed. This was necessary to capture Tweets that may contain a predictor term that is not in its base form. For example, a Tweet containing the word *pricey* would not match the term *price*. Furthermore the framework also accounts for miss spelt words. To do this efficiently the algorithm computes the edit distance between a given word and terms from the predictor lexicon. We return the predictor term with the minimal edit distance if the error is in a fixed threshold

3.4.3 Sentiment Extraction

Experiments in [4] showed that sentiment analysers such as SentiStrength [46] or Stanford CoreNLP [44] performed poorly on microblog content. Hence in [4] the decision was made to extract the sentiment by having specific terms for each sentiment (polarity). Besides one had to account for changes in polarity through negations such as *never* and *not* which inverted the polarity of a predictor category term.

We choose to deviate from this approach and use a sentiment analyser despite the bad results. We give two reasons for doing so. **1.)** Hutto et. al recently published a new sentiment analyser VADER [22] with an F1 Classification Accuracy = 0.96 which outperformed human evaluators. **2.)** Often keywords can not be manually assigned to a polarity without knowing its context.

Besides the above-mentioned benefits “VADER allows us to obtain a degree of sentiment by analysing grammatical and syntactical conventions that humans use when expressing sentiment intensity” [22]. For example it accounts for emoticons which are commonly used to express a sentiment or even acronyms such as *LOL*, *WTF*. It is further worth mentioning that VADER is an unsupervised approach and is well suited for streaming data.

Twitter Data Analysis

In this chapter we provide a quantitative and a qualitative analysis of the retrieved Twitter data. In the former we want to emphasize the general popularity of food among the Twitter users and describe the term distribution of the different commodities. We further investigate if and to what extent social media data can be found to correlate with the international Food Price Index (FPI) and the commodity price quotes. This is accomplished by analysing 29 M Tweets related to food. In the qualitative analysis we focus on examining the content of Tweets and investigate the relevance to our desired food security objectives. For this purpose we automatically detect peaks in our twitter conversations and manually investigate the cause of 352 conversation spikes.

4.1 Quantitative Analysis

To facilitate the comprehension of our Twitter dataset we draw some basic statistics of the user and term distribution. The later illustrates the volume increase that we achieved through iteratively improving our lexicon in Section 3.2. Lastly we perform a through correlation analysis between food conversations and the FPI.

4.1.1 User Distribution

Twitter is a social network and in general such networks follow a power law distribution [52]. We see in the below Figure 4.1a and Figure 4.1b that the distribution of the number of Tweets per user deviates from a normal power law. A lot of individuals send only a few Tweets about the subject and only a small number of users transmit a large amount of Tweets. Unlike [8] suggest the contribution participation level of 80 %, 20 % does not seem to apply to Tweets about food. In Figure ?? we can see that the curve is almost linear. About 50 % of the Tweets are caused by 50 % of the users. This deviates highly from the normally observed 80 %, 20 % ratio. We assume that this is due to the wide spread interest of the topic.

4.1.2 Food Term Distribution

Our framework for the data acquisition successfully increased the total volume of food related Tweets. From an initial 13.7 M Tweets we raised the entire volume by 110% to a total of 29.9 M food related Tweets. The distribution of the volume per food term is displayed in Figure 4.2a. We illustrate in light grey the added volume alongside the initial size in dark grey. The most popular food terms on Twitter are general terms such as *food*, *dinner* and *lunch*. Within the 10 most popular terms we found that three beverages (coffee, beer, tea) were represented. The most popular traded commodity term on social media is chicken. We further show the distribution of the categories in 4.2b. By far the highest contribution has the category *other*

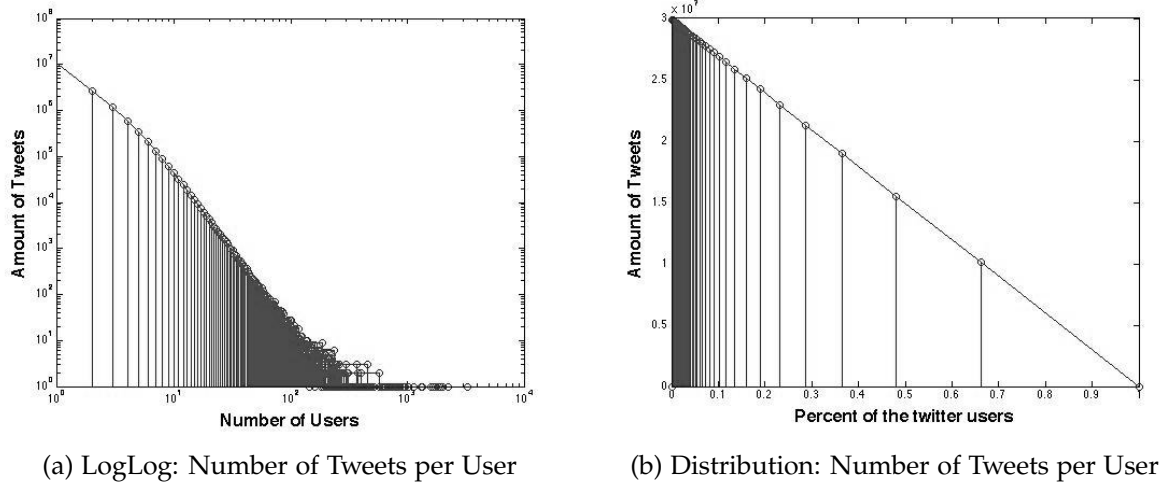


Figure 4.1: User Distribution

food of interest due to general food related keywords such as *dinner* or *food*. It builds the absolute majority with 51 %. Meat related keywords have the second highest contribution with around 15 % followed by 12% sugar, 11% cereals, 10 % dairy and lastly 0.2 % vegetable oils. We would like to note that the volume roughly follows the economic importance of the different categories with the only outlier being sugar [39]. We assume this is due to the highly popular products *coca cola* and *chocolate* which caused alone 70 % of the sugar related Tweets.

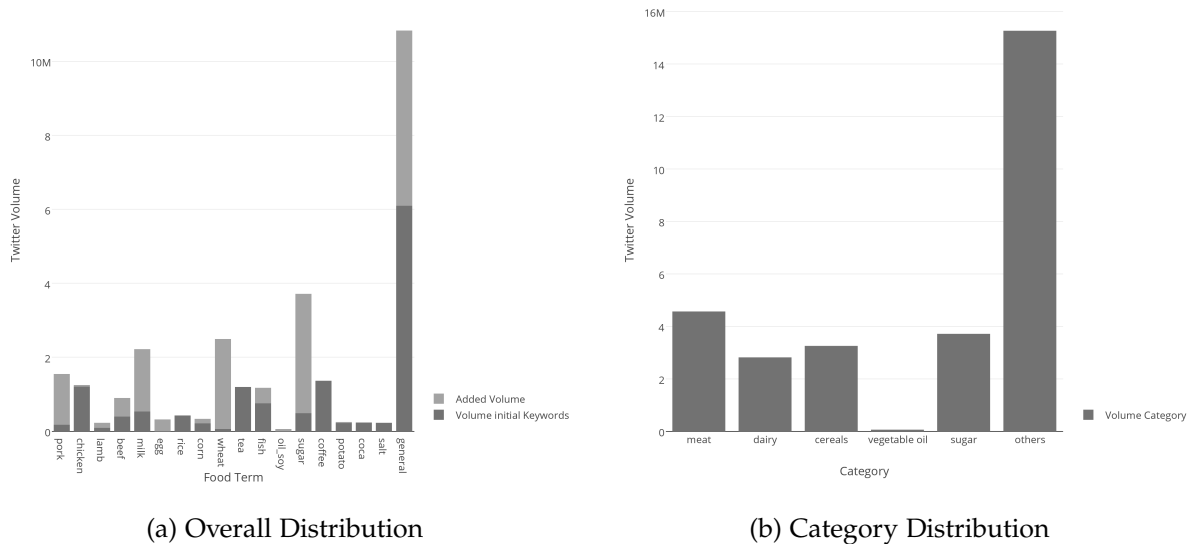


Figure 4.2: Volume of Tweets per Keyword and per Category

4.1.3 Price Correlation

We observed a general popularity of food in our initial analysis and that certain food categories have a much stronger presence than others. There is however still a concern on whether the sampled data is useful to detect difference in price fluctuation and lastly can be

used as medium to determine food security. For the purpose of this correlation analysis we used the price quotations of the Food and Agriculture Organisation of the United Nations¹ and commodity quotes from candle². FAO differentiates between a Category Food Price Index (CFPI) and a universal FPI. The CFPI is specific to a food category (e.g. meat, cereals) so different among all categories, whereas the FPI is a general indicator and the same for all categories. Unfortunately daily commodity quotes could only be obtained for meat, dairy and cereals.

For each food category (e.g. *meat, dairy*) we correlated the tweet volumes of the subcategories (e.g. *beef, chicken for meat*), products (e.g. *bacon, salami*) and the price quotes for each category. These subcategories mirror the categorisation of the FAO [39]. Since the price quotes of the FAO are based on a monthly average, we aggregated the daily tweet volumes per food term over a month and calculated the daily average volume. We only included food terms that have an average of greater than 10 Tweets per day. The internet archive did not contain Tweets for certain days. We approximated those values by taking the average of the previous and the following day.

Results

Between the meat subcategories there is a positive linear relationship in the range of 0.7264 to 0.9361. This means if the volume of chicken Tweets increases so do the volumes of beef and pork conversations. A p value of 0.0001 suggest that we can reject the idea that the correlation is due to random sampling. No clear relationship exists between the tweet volume of the meat category and the three price indices. A slight correlation for the meat subcategories was observed where all of them were negatively correlated to the price quotes meaning that if the volume increases the price will most likely decrease. Only a few sub products showed a significant correlation with the price quotes. A positive relationship can be seen between the term goat and the commodity price with a correlation of 0.7369 and a p value of 0.0001. A possible explanation might be its popularity among developing countries. 95% of the world's 475 million goats are located in developing countries³. People consuming goat meat would be more sensitive towards price fluctuation making it potentially a valuable feature in measuring food prices. By correlating the price indices we see that there is a strong positive relationship between the FAO meat price index and the commodity quotes. This analysis supports Abbott et al's theory [2] that the commodity markets have a strong influence on the rising food prices and are a strong indicator for measuring food security.

¹<http://www.fao.org/worldfoodsituation/foodpricesindex/en/>

²<https://www.quandl.com/>

³<http://www.fao.org/docrep/009/ah221e/AH221E13.htm>

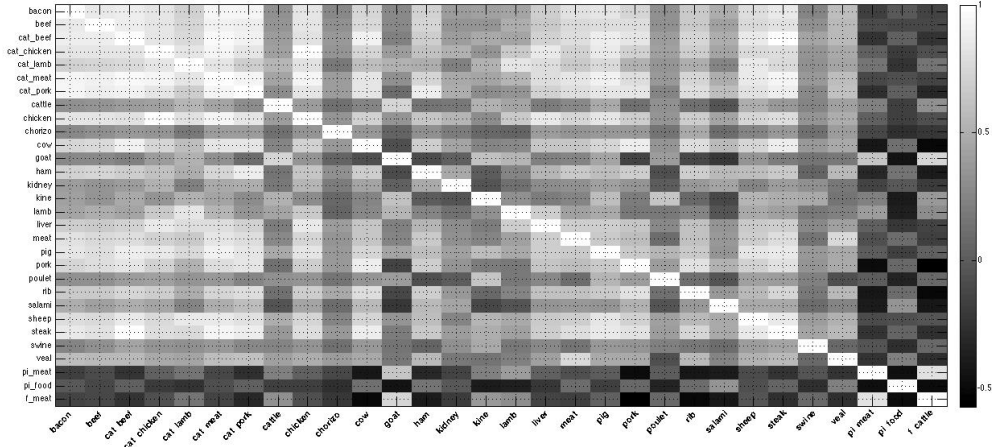


Figure 4.3: Heatplot Meat: Volume of Tweets per Keyword and per Category

For cereals similar to meat we likewise see a high correlation in volume of around 0.82 between the different cereal categories. The products *beer*, *barley*, *bread*, *atta* and *pasta* show a strong positive relationship to the cereal categories. Unlike meat, the cereal category price index and the commodity price show a strong positive relationship with the universal FPI. This is somewhat surprising as meat prices have a stronger influence on the universal FPI than cereals do [39]. Furthermore the product pasta has a strong linear relationship with the commodity price of 0.7212.

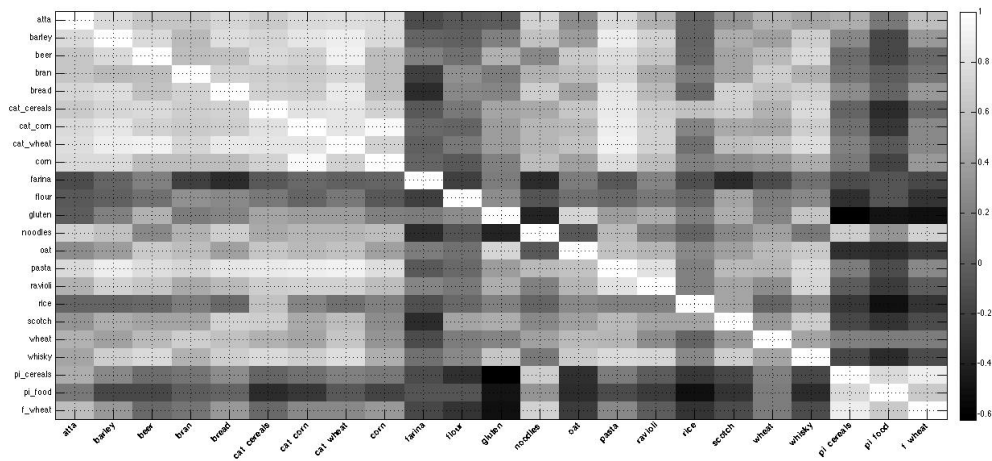


Figure 4.4: Heatplot Cereals: Volume of Tweets per Keyword and per Category

The heat plots of dairy, sugar and oil show yet again no clear linear relationship between the Twitter volume and the food price indices. More so than in other food categories the subcategories of dairy can be clearly distinguished through its strong correlation with the different products (i.e. mozzarella has a strong relationship with the category cheese and only a weak correlation with milk products). The heat plot for dairy, sugar and oil have been added to the appendix B.

Discussion

Our analysis did not show a significant correlation between the raw attention on food and the price quotes. Nonetheless the insights gained from this analysis will help us improve our features. For example the category meat shows a number of products that have a strong negative correlations. By only including such terms we are hoping to strengthen the relationship between the meat category and the price quotes.

Although we can not provide any scientific evidence there might be a nonlinear relationship between social media and the commodity market. We hence will experiment with a nonlinear model to predict price quotes in Chapter ???. According to [14] such models are better suited to utilise social media for predictions.

A smilier correlation analysis has been made by the UN [24]. They however used contextual sensitive Tweets i.e. instead of only using Tweets containing food they performed an n-match on different criteria. The tweet had to contain a food item, the word price and a quantification such as high or low. Overall a Pearson correlation of around 0.42 was detected with a significance of 0.04. By exploiting our predictor lexicon to filter Tweets that contain keywords such as supply and price we were able to improve the linear relationship and found similar results as in [24]. Although the UN concluded a linear relationship they simply provided assumptions about what might have caused the volatility of price conversations. We hence explore the conversation drivers in the next section.

	Category Price Index	Food Price Index	Commodity Price Index
Meat	-0.0112	-0.0653	- 0.1489
Dairy	-0.2166	0.1314	-0.0676
Cereals	0.0357	-0.3360	0.0594
Oil	-0.2484	-0.2382	-
Sugar	-0.2000	-0.1019	-

Significance: $p < .0005$ ***, $p < 0.005$ **, $p < 0.05$ *

Table 4.1: Price Correlation

4.2 Qualitative Analysis of Conversation Drivers

Following our quantitative analysis we proceed with a detailed investigation of Twitter conversations relevant to food security to uncover events that trigger conversations. We found that our contextual sensitive Tweets (i.e. such tweet that contain a food term and a predictor term such as price) have a stronger Pearson correlation than the raw volume. Encouraged by this observation we want to investigate further to which extent the Tweet content is related to food security. By means of a visual analysis we first characterise the temporal behaviour of our twitter conversations and manually investigate the major peaks to verify that such spikes are caused by real life events. With the intent to automate the detection of real life events (we will interchangeably use events or spikes to refer to peaks) we introduce a methodology and introduce a semi automated approach to identify the conversation driver. Lastly, we present the annotation categories and the results of this section.

4.2.1 A Visual Analysis of the Social Attention

We commence our investigation of the conversation drivers by a visual and manual investigation of the most prominent spikes. To gain an overview about the social attention of our food topics we plotted the relative distribution of food supply, price poverty and needs in Figure 4.5. By far the highest attention is attributed to food needs with around 70 % , poverty and supply receive a similar attention distribution with price taking the smallest interest among Twitter users.

To visually categorise the activity, Lehman et al. [31] defined three categories of temporal behaviours. “Continuous activity, periodic activity or activity concentrated around an isolated peak”. Continuous activities are topics that are of daily interest such as weather. On the other hand periodic activities reoccur with a fixed pattern such as the release of a popular Tv show. The latter is event driven and usually occurs once during a very short period such as a national holiday.

For price and supply we observe a similar temporal pattern. Both show a continuous activity with one extremely prominent isolated peak. The activity is concentrated symmetrically around those two events , showing abnormal activities for around 9 days before and after.

We manually investigated the two isolated peaks to see if we can attribute them to any discussions relevant to food price or food supply. Surprisingly, the content in the price discussion corresponds to a popular Korean pop band *T-ara* . *T-ara* released a music video on the 10th of September which caused the first anomaly, reaching a global maximum on the 16th when they announced to collaborate with a famous European DJ ⁴. Similarly, in our supply conversation the peak was not caused by supply indicators but was driven by conversations centered around health & life style topics.

The topics needs and poverty do not exhibit any extreme outliers and similar to price and supply can be categorised according to Lehman et al's. framework as of continuous interest.

⁴<http://www.kpopstarz.com/articles/112632/20140916/t-ara-sugar-free.htm>

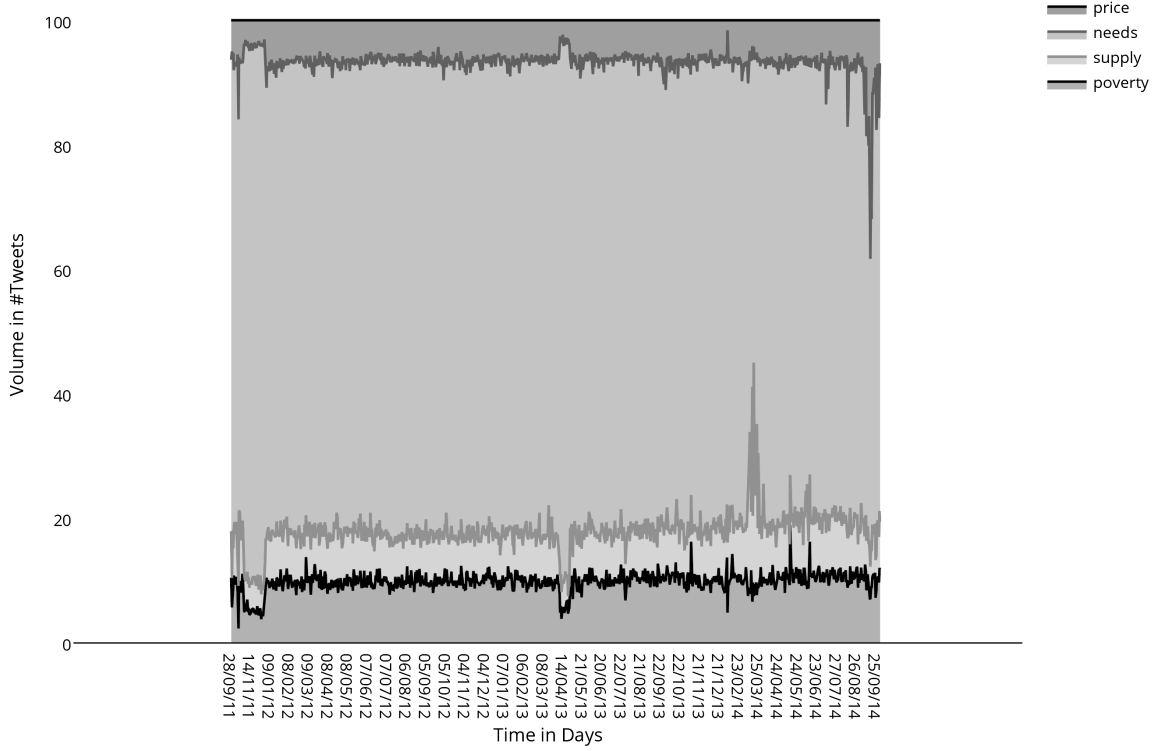


Figure 4.5: Topic Distribution - Food Security

4.2.2 Event Detection

As described in the previous section the prominent peaks could not be attributed to any discussion around indicators that are relevant to food security. To facilitate the automatic detection of peaks we now introduce an automated approach similar to [3] [31]. We evaluate the four food categories' temporal behaviour on a granularity of one day. This scale was chosen in order to be in accordance with the temporal quotations of the commodity market. To detect peaks in our food topics we used a fixed window size of $2m + 1$ where $m = 15$ giving us a month long window. Within the window we identified the median and calculated the mean of the Twitter volume. From those values we calculated the Median absolute deviation (MAD) as in Equation 4.1:

$$\overline{MAD} = \text{median}_i(|X_i - \text{median}_j(X_j)|) \quad (4.1)$$

X_j is the set of data points within the fixed window and $X_i \in X_j$

A peak is declared if v_i deviates more than 2 MAD from the mean. For this analysis we only consider positive peaks and ignore anomalies in form of a steep descent.

The discussion centered around food price showed 82 peaks. Tweet activities for food supply resulted in 91 peaks. 80 peaks were detected for food needs and lastly 99 for food poverty.

To identify what topics spike the attention we follow the approach discussed in [3]. We computed the top 50 unigrams and top 10 bigrams of all Tweets occurring during a peak. We then manually investigate the Tweets that contain the most frequent n-grams. Some peaks

could be attributed to multiple events. If two conversation drivers could be identified, both of them were used to label the peak. Else, if most likely more than two events caused the peak we marked it as ambiguous.

4.2.3 Event Annotation

We annotate each peak according to the definitions given below. Our classification mostly mimics the main dimensions of food security but also includes categories from the taxonomy of Tadesse et al[45]. There is a strong overlap between the two taxonomies where the later focuses more on Economic Access and the former has a stronger orientation towards Food Utilisation. This categorisation is not extensive i.e. there are a range of further categories we could consider. However given the sparsity of relevant events this classification gives a good overview of the discussed topics.

Some events show causal relationships i.e. a breach in the food supply can be a cause for riots and political unrests. In such cases we annotated both.

Food Supply

Events entered around the food supply chain are considered including indicators of food waste. We define Food Loss and Food waste according to Parfitt et al. 's [26] definition. Food Waste refers to Food Loss that occurs at the retailers and consumers side whereas the term Food Loss refers to the decrease in food volume that leads to edible food for consumption.

Economic Access

We define Economic Access according to FAO's [39] definition. Price, expenditure or market indicators fall into this domain.

Government

The classification Government takes topics such as legislation and policy changes into account. An example is restrictive trade policies such as export or import restrictions [45].

Stability

Poverty, political unrest and topics concerning extreme weather [39] fall into this classification. Factors that cause insecurity such as riots or severe draughts are considered.

Unrelated

Viral jokes, advertisements, health & lifestyle are example topics that we consider unrelated.

Our findings showed that for price only 7 (8.5 %) out of 82 fell into the above given categories, for supply 4 (4.3 %) out of 91 for poverty 13 (13 %) out of 99 and finally for needs no relevant topics were found.

4.2.4 Results

The distribution of the annotations is visualised in Figure 4.6. Surprisingly the conversations mostly peaked outside their domain, i.e. the price conversation was more intrinsic for supply indicators than for economic access indicators. We now give examples to each annotation topic of events that we classified as food security relevant to illustrate what kind of discussion caused a peak .

Food Supply

Topics that caught the social media audience were especially safety threats to the food supply. In April 2012 a newly discovered case of cow disease threatened the safety of

America's beef supply and heavy import restrictions were imposed from major beef importers such as South Korea ⁵.

Economic Access

In 2014 sharp rising food prices caused a lot of discussion on Twitter. Wholesale prices were suffering due to a severe drought in the previous year, which thinned the cattle herds and increased consumer prices ⁶. As a consequence there was also a sharp increase in discussion around food banks. The UK observed a 51 % increase in food bank users ⁷.

Government

Most discussions around legislation changes were focused on Food Bank reforms. A high amount of attention can be attributed to the UK rejecting the European Union food bank funding. The population heavily criticised the British government to deny EU fund to be spent on the poor ⁸

Stability

Discussions around stability were usually headlined by extreme poverty causing riots. A food program that provided free lunch to underprivileged school kids used poisoned crops in their dishes. 20 children died as a consequence causing riots and closed shops all over the city. ⁹

Unrelated

Unrelated topics cover a vast amount of domains. Most often peaks are caused by viral Tweets posted by online celebrities that contain a food term. Public holidays, such as Easter, Thanks Giving are also frequently captured. Furthermore public figures such as Ray Rice, a famous football player, caused a lot of hype in the social media community ¹⁰. Often it was very hard to extract the conversation drivers in the unrelated topics. There is a considerable amount of noise in our conversations centered around food security, making it very challenging to extrapolate meaning from an event. This might be attributed to the general popularity of food we identified in previous chapters.

⁵<http://www.theguardian.com/science/2012/apr/25/mad-cow-disease-us-mutation>

⁶<http://www.cnbc.com/id/101588110>

⁷<http://www.bbc.com/news/business-27032642>

⁸<http://www.theguardian.com/society/2013/dec/17/government-under-fire-eu-funding-food-banks>

⁹<http://www.usatoday.com/story/news/world/2013/07/17/india-children-deaths/2523727/>

¹⁰<http://www.nytimes.com/2014/09/09/sports/football/ray-rice-video-shows-punch-and-raises-new-questions-for-nfl.html>

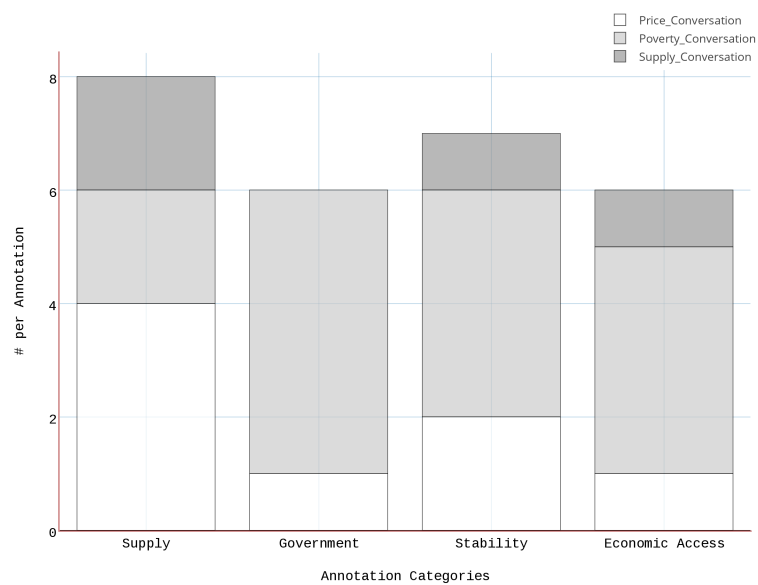


Figure 4.6: Annotation Distribution - Relevant to Food Security

4.2.5 Discussion

Extracting information from noisy and unstructured text has shown to be a very difficult task. Despite our various filtering attempts there is still a substantial amount of noise in our food security discussions. As most of the volatility was attributed to unrelated topics we will have to rethink the way we use the Twitter data as a feature in our model. Unrelated does not necessarily mean irrelevant. Even Twitter discussion outside our food security objectives can provide valuable information. Most commonly this is achieved by analysing the attitude of a tweet’s author. We will hence explore the entropy of sentiment in explaining volatile food prices.

Forecasting Model

This chapter applies the insights gained in Chapter 4 and makes a practical case on how we can use Twitter as an early warning system for food security threats by means of time series forecasting. First we outline the methodology of our time series prediction followed by a discussion on the different input models we used. We further highlight details towards the data preprocessing and how we train our model. Lastly, we present some basic concepts that will aid the reader in understanding the dynamics of the predictions and ultimately help to comprehend the results.

5.1 System Overview

In Figure 5.1 we present a conceptual overview of our proposed system. A clear understanding of the problem and knowing which variables are capable of explaining the variance is important for accurate predictions. We create different features from the historical price of the commodities and the filtered social media data. To assess the predictive value of the features we run a feature selection algorithm Relief. The proposed set is used as our input and later served to the prediction model. First however we normalize the data and in a second instance perform a smoothing of the input signal. Not illustrated in this image is an additional optimisation procedure we performed for the social media data. We assume that a causal relationship exist between the twitter data and the price. Given this assumption we investigate at which point in history the social media data correlates the most with the given price data. We discuss the particular case of the social media features in Section ?? . It is common practice to divide the data into three distinct sets namely train, validation and test. Due to the sparsity of the available social media data we restrict ourselves to a train and test set. The reader should note that the value of the train set is not static. As we discuss in Section ?? the training set is updated every 28 days to include recent values. The fourth and final step constitutes the actual model. Roughly the prediction model ANFIS performs four steps. Fuzzification corresponds to the process of transforming a value between 0 and 1 into grades of membership of linguistic terms. Since the number of linguistic rules is exponential to the number of input values we perform a clustering of the rules. In the third step those rules are evaluated and eventually transformed into a scalar value between 0 and 1. The output is the predicted value of the price of a commodity.

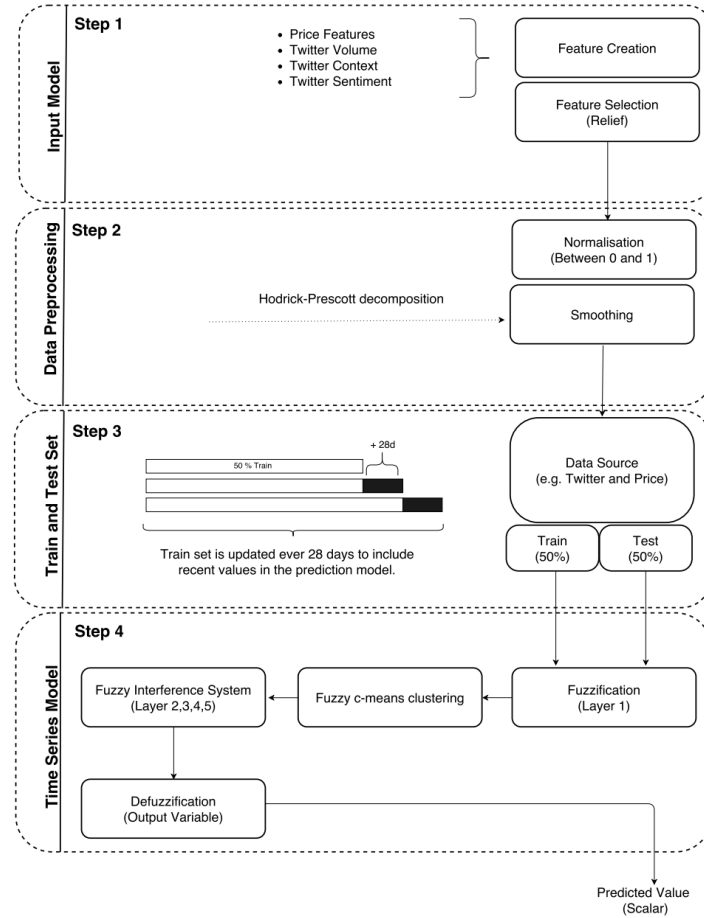


Figure 5.1: Forecasting - System Overview

5.2 Input Model

Critical to the success of the model is choosing the right set of features to represent the sample. This is however an extremely challenging task which is one of the reason why we use ANFIS, namely to detect complex nonlinear relationships. Nonetheless economic theory can help to create representative features. Generally speaking good features should be informative and for our particular task be able to discriminate between rising and falling prices. Given the above named challenges we try to motivate the choice of features by either referring to past work or relying on economic theory. An input model selection algorithm will aid our attempt of selecting a representative set of features. For benchmarking purposes we create three different models. The first model consists of price features, the second model uses social media features and the last model constitutes a combination of the both. In this section we first describe the features we extract and later define the three input models.

5.2.1 Features

In the following sections we present two types of price features (spot price and moving average) and three different social media features (attention, context and sentiment ratio).

Price Feature

Yao et al. ?? suggests that the market can be categorised using the major trends, the intermediate trends and the minor trends. By major trends we refer to trends that last more than a year, and by intermediate trends anything between three weeks to three months. We capture such trends by taking the moving average of one week, two weeks and one month respectively similarly to Yao's model. Given the limited time frame we chose to exclude major trends. We further consider all days preceding the horizon 30 days into the past as potential features. Moving averages are used to remove the day to day instability and extract the underlying trend[5]. The below Table 5.1 illustrates the values assigned to each feature. x is the value of today and $x - 1$ is a value one day in the past. $\mu(\lfloor x - y; x \rfloor)$ symbolises the moving average of the past y days.

input #	name	value
1	D1	x
2	D2	$x - 1$
:	:	:
:	:	:
:	:	:
30	D30	$x - 29$
31	W1	$\mu(\lfloor x - 7; x \rfloor)$
32	W2	$\mu(\lfloor x - 14; x \rfloor)$
33	M1	$\mu(\lfloor x - 30; x \rfloor)$

Table 5.1: Input Model: Benchmark Prediction

Social Media Features

In Section 4.1.3 we rejected the hypothesis that a linear relationship exists between the volume of food tweets and the price of a commodity. However, John Bollen et al. [9] showed that Twitter conversations have a nonlinear relationship with the movements of the financial markets. Hoping to find a similar relationship for commodities we consider social attention on different granularities by measuring the volume of the category, subcategory and the product 3.2. We further include products that are not specific to a commodity but show a strong linear relationship with the commodity price (i.e. we consider goat as a feature for beef despite it not being a beef product). For beet we consider the twitter volume of the products salami and goat, for wheat we consider the twitter volume of gluten and noodles and lastly for milk we consider the twitter volume of the products mozzarella and cheshire. All of those products were found to have a significant correlation with the commodity price.

In general microblogs capture one topic due to its 140 keyword limitation. The topics can usually be inferred by capturing one or two keywords. However certain terms are more intrinsic than others i.e. the keyword *IBM* can unambiguously be related to the company whereas the term *break* has multiple meanings and given the context, could be unrelated to a desired topic. We hence introduce the notion of contextual sensitive Tweets. Tweets considered contextual sensitive match both a term in the Food Lexicon and the Predictor Lexicon. We consider the context of food price, supply, poverty and needs 3.3.

Despite having identified a significant correlation for contextually sensitive Tweets our analysis in Section 4.2 concluded that unrelated topics mostly drive the attention volatility. Public mood states or sentiment might hence be a more valuable indicator. This intuition is supported by [37] namely that emotions and mood do not uncommonly drive financial decisions. For the purpose of this analysis, we consider different ratios of sentiment. The ratio between

the numbers of positive and negative Tweets [36], the proportion of neutral and total Tweets, the ratio between the numbers of non-neutral and total posts [56], the ratio between the number of positive and negative discussions and lastly the ratio between the numbers of neutral and non-neutral messages.

Table 5.2 concludes our input model. We measure the sentiment for both Twitter buzz and contextually sensitive Tweets. To uniquely identify from which category we are measuring the sentiment # refers to a unique identifier. The result is an input model with 51 features.¹

Feature Type	Name	Value
Attention	AC#	#Commodity
Context	CP#	#Price f. Commodity
	CS#	#Supply f. Commodity
	CP#	#Poverty f. Commodity
	CN#	#Need f. Commodity
First Sentiment Ratio	SR#	#positive : #negative Tweets
Second Sentiment Ratio	SR#	#positive : #(positive + negative) Tweets
Third Sentiment Ratio	SR#	#negative: #(positive + negative) Tweets
Fourth Sentiment Ratio	SR#	#neutral : #(positive + negative) Tweets
Fifth Sentiment Ratio	SR#	#(positive+negative) : #all Tweets
Sixth Sentiment Ratio	SR#	#neutral : #all Tweets

Table 5.2: Input Model: Social Media Prediction

5.2.2 Feature Selection

We aim to find the most significant features among the 33 price and the 51 social media features to decrease the complexity and lastly improve the prediction accuracy. We perform the model selection for all three models (e.g. price, social media and mixture model). For that purpose, we used a feature selection algorithm Relief [28] [29] [40] that allows us to distinguish relevant from irrelevant features where relevance is measured on how much influence a feature has on the output value. Relief finds relevant features using a heuristic where the evaluation is performed on a distance measure. The data instances of the features belong to different classes. We measure the relevance of a feature f_x by measuring the distance from the nearest neighbour of each class to f_x . The nearest neighbour of the same class (same class as f_x) is called nearest hit, and the nearest miss is the closest neighbour of a different class. The heuristic is computed for each feature in Equation 5.1.

$$W_i = W_i - (x_i - nearestHit_i)^2 + (x_i - nearestMiss_i)^2 \quad (5.1)$$

A feature is considered irrelevant or redundant if $W_i \leq 0$, that is if the difference between the same class is higher than the one of a different class. We measure the most significant features with respect to the horizon of four days, seven days and 14 days. Furthermore, we consider three different commodities, wheat (w), beef (b) and milk (m).

Price Model

The model selection is performed on the input values illustrated in Table 5.1. The configuration $w-4$ in the Table 5.3 refers to wheat with a prediction horizon of four days. The most

¹ (1 × attention for each food category (3) + 1 × attention for products with high correlation (2) + 1 × attention for each context (4) + 6 × sentiment ratio of each food category (18) + 6 × sentiment ratio of each context (24) = 3 + 2 + 4 + 18 + 24 = 51)

predictive features are 1.) $\varphi_1 : D30$, 2.) $\varphi_2 : D29$, 3.) $\varphi_3 : D28$, 4.) $\varphi_4 : W1$, 5.) $\varphi_5 : D27$ and 6.) $\varphi_6 : D26$.

The numbers in the Table 5.3 refer to the relative importance of the feature for a specific configuration. The higher the value, the more predictive the feature is considered to be. It is interesting to observe that Relief consistently suggested the same features irrespectively of the commodity type and the prediction horizon. This gives us a strong confidence in the relative importance of the prediction task at hand, as it generalises well over different models i.e. horizons and also different commodities. We can see that the values follow a clear trend. With an increasing horizon, the features lose its importance as the prediction task becomes more and more challenging.

For our prediction task, we consider the top 10 features². Those additional features ($\varphi_7 - \varphi_{10}$) not displayed in Table 5.3 are for most commodities and horizons the same but ranked in different orders depending on the prediction task. Interestingly D25, D24, D23 and W2 rank highly along the other values in Table 5.3 which are all placed in the intermediate past. This observation suggests that models based on the intermediate past poses more predictive power and outperform strategies based on features in the recent past. Robert Marx also observed this effect and investigated the hypothesis in [20]. His conclusion was that models with intermediate variables tend to outperform such only considering recent values. A possible explanation mentioned in the paper is that such variable best captures the momentum of a commodity price.

Feature	w-4	b-4	m-4	w-7	b-7	m-7	w-14	b-14	m-14
φ_1	0.0624	0.0212	0.0310	0.0586	0.0197	0.0300	0.0487	0.0155	0.0271
φ_2	0.0558	0.0193	0.0281	0.0513	0.0175	0.0268	0.0411	0.0132	0.0237
φ_3	0.0495	0.0174	0.0253	0.0444	0.0153	0.0236	0.0343	0.0111	0.0205
φ_4	0.0440	0.0157	0.0227	0.0388	0.0135	0.0209	0.0291	0.0093	0.0179
φ_5	0.0434	0.0156	0.0225	0.0381	0.0133	0.0206	0.0283	0.0091	0.0175
φ_6	0.0378	0.0138	0.0198	0.0324	0.0114	0.0178	0.0231	0.0074	0.0149

Table 5.3: Feature Selection: Benchmark Prediction

Social Media Model

In this section we select among the features displayed in Table 5.2 to build the input set of our social media model. Again we try to find features that generalise well among different commodities and horizons. The acronyms (e.g. AC50) used in this section correspond to the ones introduced in Section 5.2.1. The number is used as an index to distinguish sentiment ratios of different food categories.

In the previous section we showed that input variables for the price model generalise well among all commodities and price horizons. On the other hand social media features seem to be more specific to a given task and commodity. For the top 16 features, we only found three features that were present in all commodities and all horizons or in other words were able to generalised well. Among the three were the attention count of the selected products that showed a high correlation in our analysis (AC50), the first sentiment ratio of poverty (SR32) and the first sentiment of the subcategory (SR15). Other features that generalised well for all prediction tasks of a given commodity were the second sentiment of the subcategory (SR16), the second sentiment of poverty (SR33), the fourth sentiment of poverty (SR35), the

²We performed experiments with five, 10 and 20 features and found that 10 features yield the highest prediction accuracy. A better predictions can be expected with more than 10 features however at some point the model will overfit. We leave an extensive search of the optimal number of features for future work.

first sentiment of price (SR38), the second sentiment of supply (SR45), the sixth sentiment of supply (SR49) and lastly the attention count of another selected product that scored well in our correlation analysis (AC51). The top features along with the Relief score for each given commodity are illustrate in Table 5.4. The sentiment of the tweets together with the attention counts of products that showed a high correlation in Section 4.1.3 seem to be expressive for the price variation. Especially the sentiment features for discussions centered around poverty rank highly. Our qualitative analysis in 4.2.4 showed that twitter conversation for poverty discussion had the highest relevance. We would like to point out that no features built on discussions centered around food needs were selected. This again correlates with our analysis in previous sections where we concluded that non of the volatility in the needs discussion was caused by topics relevant to food security. Admittedly, it is not a decisive indicator of the quality of the feature. However, it might explain why poverty features rank so highly.

Feature	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
<i>wheat</i>	AC50 (0.2119)	SR32 (0.2294)	AC51 (0.2121)	SR35 (0.1356)	SR16 (0.1385)	SR15 (0.1385)
<i>beef</i>	AC50 (0.2210)	SR45 (0.1389)	SR32 (0.1323)	SR33 (0.1339)	SR38 (0.1339)	SR44 (0.12826)
<i>milk</i>	AC50 (0.3339)	AC51 (0.1774)	SR45 (0.1774)	SR38 (0.1413)	SR35 (0.1313)	SR49 (0.1013)

Table 5.4: Feature Selection: Benchmark Prediction

Mixture Model

Combining the price and social media features from Table 5.1 and Table 5.2 we want to investigate if Relief proposes as similar set as for the separate models or if the combined model favours a different group of input variables. Only a few features generalised well over all commodities which is why we focus our search of finding input variables that are specific to beef, milk and wheat but generalise well over different prediction horizons. As in our price model, days in the intermediate past were proposed for all three commodities. The ratio varies strongly among the commodities. Beef favours price variables where milk and wheat have a majority of social media features. We observe an interesting trend, namely that all commodities favour supply and needs feature over price and poverty which is very different to the pure social media model. We assume that the price data made the conversations about price more obsolete. To conclude the proposed features are the fourth and the fifth sentiment ratio of the product (SR11, SR12), the fourth, fifth, sixth sentiment ratio of needs (SR29, SR30, SR31), the third sentiment ratio of poverty (SR34), the first, third sentiment ratio of price (SR38, SR40) and lastly the second, third sentiment ratio of supply (SR45, SR46). The price features are all days of the last week i.e. (D24, D25, D26, D27, D28, D29, D30) and the moving average of the past week (W1).

<i>milk</i>	SR38	SR40	SR45	D30	D29	SR34	SR12	SR47	SR46	SR11
<i>beef</i>	SR38	D30	D29	D28	D27	W1	D26	SR45	D25	D24
<i>wheat</i>	D30	D29	SR22	SR29	SR30	SR31	SR15	SR16	SR33	SR14

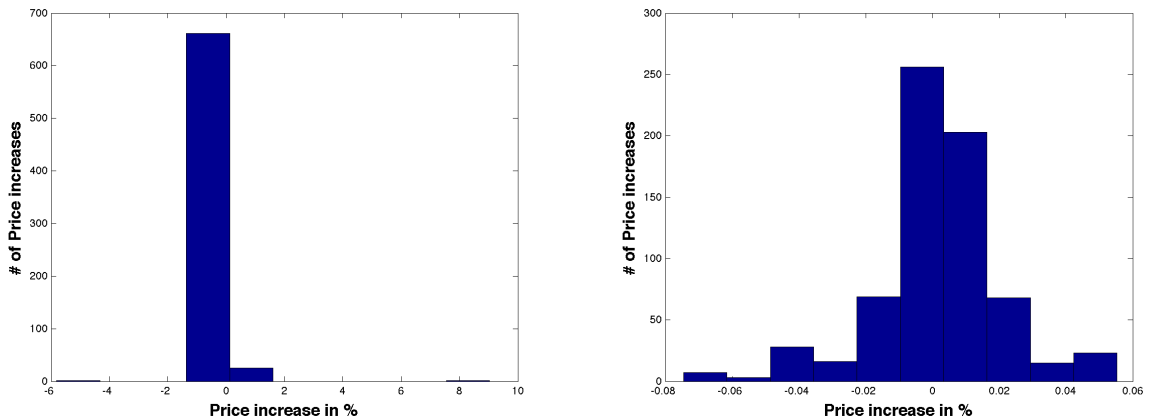
Table 5.5: Feature Selection: Benchmark Prediction

5.3 Data Preprocessing

Before feeding the data into our prediction model, we analysed and cleaned the data. The representation of the data is of great importance to assist ANFIS in learning the relevant patterns. In a first instance, we had to match both the time series of the Twitter data with the time series of the price data. During weekends and national holidays the markets are closed, hence we removed such instances from the Twitter data. The commodities beef, milk and wheat are traded on different commodity markets. We observed that some of the markets had different closing days i.e. some markets considered a day a holiday, some others not. For wheat and cattle we removed the 12/11/12 and the 8/10/12 which are the Veteran day and the Columbus day respectively to match the price data of milk. Secondly, we proceeded to interpolate zero values. Some of the price data sets showed values of zero on days that were neither weekends nor holidays. Similarly, the Twitter archive did not contain Twitter data for some time periods. We linearly interpolated such missing values by solving and approximation to the partial differential equations [17]. Fuzzy Interference Systems expect an input of a unit interval i.e. between 0 and 1. We hence normalised the data as illustrated in Equation ?? . Min and Max are the lower and upper bounds of data set where α is a small constant we introduced to avoid zero divisions. The Fuzzy system expects a unit interval to avoid losing its sensitivity towards smaller scaled features.

$$\bar{y} = \frac{x - \min}{\max - \min} + \alpha \quad (5.2)$$

Lastly, we performed a smoothing of the data which is in agreement with our objective, namely to be more sensitive to long-term than to short-term fluctuations [21]. As we can see in Figure 5.2 there are some extreme price increases and drops. By applying the Hodrick-Prescott decomposition [21] filter, we receive a distribution that is more normal by avoiding such outliers. Additionally the Figure ?? illustrates nicely the characteristics of commodities, namely that it is mostly driven by small price changes.



(a) Price Increase Distribution without Scaling

(b) Price Increase Distribution with Scaling

Figure 5.2: Characteristics of Commodity Price Data

5.4 Training the Model

For training and testing our model, we choose the period 03.01.2012 - 26.09.2014. We thought of numerous ways to test and train our model. Different approaches have been suggested by Ibeling Kaastra [27]. Most commonly the data is split into a train set, validation set and lastly a test set. The model is trained once in a batch fashion and used for all future predictions. Such an approach can be dangerous particularly when one considers historical data reaching far into the past. Market conditions might have been different then and might not apply to future predictions. Choosing a particular time frame can be a bad idea as well. Consider the training data only exhibiting an upwards trend the model will then not generalise well for declining prices. Yao et al. [54] proposed to use statistical methods to investigate the best period for training the network. However to make any statistically significant claims we need more than 2.5 years of historical data. Lastly, Ibeling Kaastra describes a method called the walk-forward or sliding window approach. This approach involves creating overlapping sets of a train and test data. Each set is moved forward through the time series to test the robustness of the model. This framework addresses the concern raised regarding including data from far in the past that might not reflect the current market conditions and is widely used for commodity predictions [27]. We decided to apply a variant of the sliding window approach and train our model in an online fashion. Given the limited amount of data we choose not to exclude any data from future prediction. However, to be able to adapt to new market conditions we increase the training and validation window as we move along the timeline. We used the first 50 % of the data to train the model.

5.5 Time Series Model

In this section, we present the terminologies and concepts around Fuzzy logic, focusing on the basics of Fuzzy variables and Fuzzy sets. We further derive a Neural network with a Fuzzy inference system (FIS).

5.5.1 Motivating a Fuzzy Approach for Time Series Modelling

Compared to other approaches, Fuzzy logic has seen only a few applications in forecasting despite its promising results. We hence want to motivate the use of this technique in further detail. Fuzzy logic was initially proposed to provide a framework for imprecise reasoning. Zadeh [55] introduced the concept to describe real world phenomena that do not have a precise description of a membership class. Another branch of mathematics that deals with uncertainty is the field of probability. However, there is a distinct difference between the two. Probability theory is based on Bivalent logic, which means every proposition is either true or false. Only certainty is a matter of degree, which brings us to an important distinction. In Fuzzy logic, everything is a matter of degree, which is ultimately how we perceive the real world. "This form of reasoning allowed the development and analysis of systems by expressing the qualitative aspects of human reasoning without using any complex mathematical models" [25]. In some areas such as time series prediction techniques such as ARMA and AR, have shown clear limitations [10]. Nonlinear approaches, such as Adaptive Neuro Fuzzy Inference System (ANFIS), have proven to be more successful [12]. Prediction accuracy is however not the only concern in forecasting models. Understanding the behaviour and gaining insight into the underlying dynamics is equally important [50]. This makes ANFIS especially appealing. Not only does it pose strong predictive capability but as a consequence of its rule-based design it allows for interpretability of the predictions. The interpretability is particularly important as the results might help us better understand the determinants of food security risks in social media.

5.5.2 Fuzzy Variables and Fuzzy Sets

Zadeh [55] defined Fuzzy variables as attributes that distinguish between elements of some universe of discourse. He uses the colour of an object as an example. Each colour has a wavelength, which is a precise numerical definition. In natural language, we tend to classify colours not by its numerical value, but by colour objects (e.g. red, blue, green). Colour objects fall within the scope of a specific wavelength or how it is commonly referred to in Fuzzy logic, a Fuzzy set. Red or blue describe the object's colour, but it is by no means a precise definition. By applying a membership function, we can precisely define the colour in a range between zero and one. This whole process is commonly referred to as Fuzzification and is the initial step of the system as illustrated in Figure 5.1. The most common membership functions are Gauss function, Trapezium function and the Triangle function. We choose to use a Gaussian membership function defined by Equation 5.3, as they are differentiable and desirable for optimisation purposes [53]. We will refer to a_i, b_i, c_i as the parameter set.

$$\mu_A(x) = \frac{1}{1 + \left| \frac{x - c_i}{a_i} \right|^{2b_i}} \quad (5.3)$$

Concerning our Fuzzy logic system, we will distinguish between an input variable and output variable. The output variable depends on the input value's corresponding membership function and a decision matrix, where we will explain the later in the following Section 5.5.3.

To illustrate the two concepts Fuzzy logic and Fuzzy sets let us consider Figure 5.3. The three triangular functions precisely define the scope of each colour object (red, green, blue). The colour object green is exactly green at the point x_2 and partly green between x_1 and x_2 and likewise x_2 and x_3 .

For this exercise, we fuzzily the colour object of a tangerine that is orange. We define orange to be a little red (where little is a value between 0 and 1) corresponding to y_r and very green corresponding to y_g . The membership function would take orange as input corresponding to x_i and output y_g and y_r which formulate a degrees of memberships to the fuzzy set green and red respectively. This toy example generalises to any other object. The object price can be modelled with the triangular functions high, medium low. Sentiment, on the other hand, would use positive, negative or neutral instead.

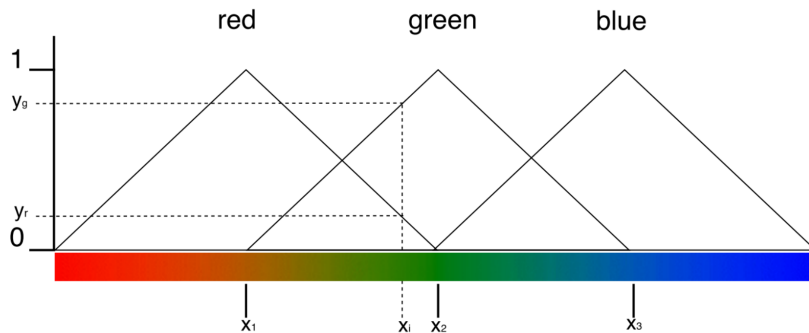


Figure 5.3: Fuzzy Variables and Fuzzy Sets - A Colour Object Example

5.5.3 Fuzzy Inference System

We first describe the components of the Fuzzy inference system (FIS) and then give intuition on how the system is modelled as a Generalised Neural Network (GNN).

FIS evaluates a decision matrix composed of rules in the following semantic:

$$\text{IF } \langle A \rangle \text{ AND } \langle B \rangle \text{ THEN } \langle \text{Conclusion} \rangle$$

The rule base grows exponentially with the number of variables. Hence, we have to consider carefully the input variables to minimise the complexity and to make the inference reliable. A multivariable system uses two different kinds of connectives to combine the fuzzified values. The union is defined in Equation 5.4 as the *multiplication* of the membership function of set A $\mu_A(x)$ and B $\mu_B(y)$. The result is a weight w_i . x and y are the input variables and could be anything from the attention count of tweets to a sentiment ratio.

$$w_i = \mu_{A \cup B}(x,y) = \mu_A(x) \times \mu_B(y) \quad (5.4)$$

The intersection is defined in Equation 5.5 as the *probabilistic OR* of the membership function of set A $\mu_A(x)$ and B $\mu_B(x)$.

$$w_i = \mu_{A \cap B}(x,y) = \mu_A(x) + \mu_B(y) - \mu_A(x) \times \mu_B(y) \quad (5.5)$$

The output of one Fuzzy Rule is computed by Equation 5.6.

$$z = dx + ey + f \quad (5.6)$$

In case the output is preferred to be constant rather than linear we set d and e to zero and the constant f to the desired output of our system. The final output is the weighted average of all rules computed by Equation 5.7 where N is the number of rules.

$$\text{Output} = \frac{\sum_{i=1}^N w_i z_i}{\sum_{i=1}^N w_i} \quad (5.7)$$

5.5.4 Adaptive Neuro Fuzzy Inference System

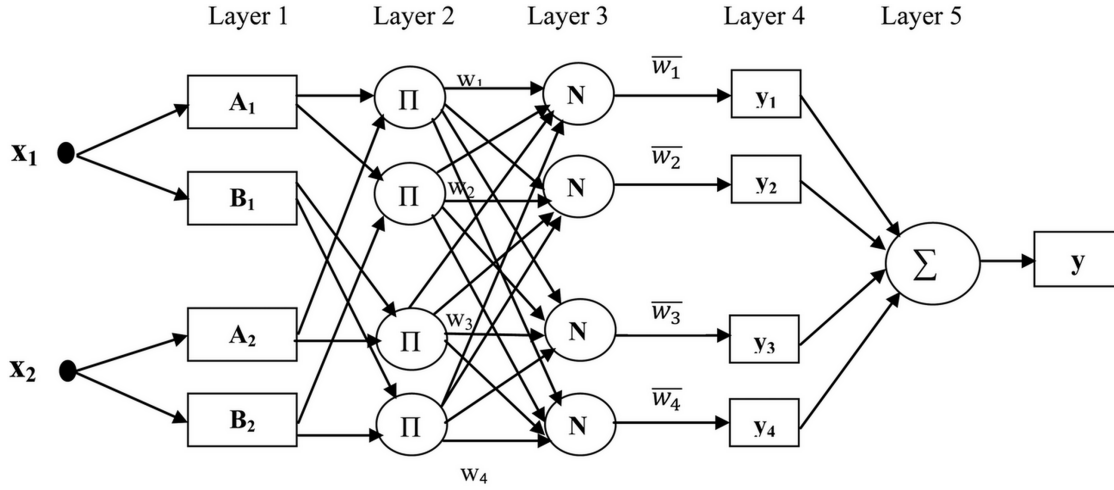
We now model the above-described process as a Neural network illustrated in Figure 5.4. In **Layer 1** every node maps the input variables x_1 and x_2 via the membership function in Equation 5.3 to a Fuzzy Set. **Layer 2** applies Equation 5.4 or Equation 5.5 which multiplies the incoming signals and forward the product to the next layer. **Layer 3** calculates the ratio of the rule's strength to the sum of all rule's strength. We apply the normalisation Equation 5.8 in this Layer.

$$\overline{w}_i = \frac{w_i}{w_1 + w_2}, i = 1, 2 \quad (5.8)$$

Layer 4 applies Equation 5.9, which is similar to Equation 5.6 but multiplied by the factor obtained in **Layer 3** from Equation 5.8.

$$\overline{w}_i f_i = \overline{w}_i (dx + ey + f) \quad (5.9)$$

Finally, **Layer 5** computes the overall output as the summation of all incoming signals through Equation 5.7. The construction yields a network with 5 layers, 16 nodes and 24 parameters (12 in Layer 1, 12 in Layer 4).

Figure 5.4: Adaptive Neuro Fuzzy Interference System³

5.5.5 Optimization

ANFIS optimises the parameters through a hybrid algorithm (least-squares & gradient descent) with respect to a given input, output training data pattern. While the network calculates the output value in a forward pass, the system uses least-squares to find the best parameter values in **Layer 4**. In the backwards pass the errors are propagated backwards and the parameters in **Layer 1** are changed by gradient descent to reflect best the input, output data. ANFIS performs a further optimisation procedure between the first and the second layer to handle high dimensional data. Generating rules by enumerating all possible combinations of membership functions would make the system computationally unfeasible for models that contain more than six features⁴. For three membership functions and x features this leads to 3^x possible combinations. Instead of enumerating all possible rules we use Fuzzy c-means clustering. The number of centroids is indicative of the number of membership functions. Each data point has a degree of membership to every centroid in the space. The summation of the membership degrees equals one for every data point. Each point is assigned through optimising an objective function which is defined in Equation 5.10 where $d_{ij} = ||x_i - v_j||$ is the Euclidian distance between the data point i and the cluster centre j . m on the other hand, is a variable to determine the fuzziness of the clusters. In other words, how much the clusters overlap. If we set $m = 1$ the membership μ_{ij} would correspond to 0 or 1 implying a strict partitioning.

$$J(U, V) = \sum_{i=1}^n \sum_{j=1}^c (\mu_{ij})^m ||x_i - v_j||^2 \quad (5.10)$$

μ_{ij} corresponds to the degree of membership of data point i to cluster centre j . This is captured by Equation 5.11. This involves taking the fractional distance from the point to the cluster centre raising the fraction to the inverse fuzzification parameter. We divide it by the sum of all fractional distances to ensure the sum of all memberships is 1.

³ Image credit to: <http://pubs.rsc.org/en/content/articlehtml/2014/ra/c4ra02392g>

⁴ <http://www.cs.nthu.edu.tw/~jang/anfisfaq.htm>

$$\mu_{ij} = \frac{\frac{1}{d_{ij}^{\frac{1}{m-1}}}}{\sum_{k=1}^c \left(\frac{1}{d_{ik}}\right)^{\frac{1}{m-1}}} \quad (5.11)$$

The fuzzy center on the other hand is computed by Equation 5.12 $\forall_j = 1, 2, \dots, c$.

$$v_j = \left(\sum_{i=1}^n (\mu_{ij})^m x_i \right) / \left(\sum_{i=1}^n (\mu_{ij})^m \right) \quad (5.12)$$

5.6 Experimental Evaluation

5.6.1 Prediction Task and Evaluation

The goal of our research is to predict the price of a commodity at some day x in the future y_{t+x} where t is considered to be the present. The prediction model takes different features from the present y_t to values M days in the past y_{t-M+1} as input. We define fp to be our fuzzy prediction model. The problem of predicting the future value y_{t+1} can be formulated as:

$$y_{t+1} = fp(y_t, y_{t-1}, y_{t-M+1}) \quad (5.13)$$

Consider the case where we predict y_{t+4} , so four days into the future. A recursive way would be to predicted values y_{t+2} and y_{t+3} and then use them as regressors in predicting y_{t+4} . However, this approach accumulates prediction errors. The further the prediction value is the more prediction outputs are used as regressors. We deviate from this approach by building a direct prediction model, so for each prediction horizon one direct model. Let us consider a concrete example. If the prediction task is to predict y_{t+2} and for illustration purposes we assume today is Monday, then the prediction goal translates to approximating the true value on the following Wednesday. Assuming the feature selection process yields y_{t-3} and y_{t-4} this corresponds to using the price on the previous Friday and Thursday. Finally our model with two membership functions High and Low gives us the following Fuzzy Model.

IF $\langle \text{Friday} \in \text{High} \rangle$ **AND** $\langle \text{Thursday} \in \text{Low} \rangle$ **THEN** $\langle \text{Wednesday} \in \text{Increase} \rangle$

High, Medium and Increase are Fuzzy Sets. The set Increase will be defuzzified and translated into a numeric value.

To compare different prediction models we measure the difference between the actual value and the predicted value by computing the Root Mean Square Error (RMSE) defined in Equation 5.14. It is an aggregation of all prediction errors for different time stamps. \bar{y}_i is the forecasted value and y_i is the actual value. In this form the error is measured on the actual price not the normalised form. Such absolute error metrics are not ideal if we average over instances with different scales. The price of a kilogram beef will be much higher then the price of a kilogram wheat. As a result the error measure for beef is likely going to be higher then the one for wheat even though the prediction accuracy might be comparable. Hence it is important that the input and output variables among different commodities are normalised to be able to compare the results across the different products. In all our experiments the RMSE of our normalised price data will be given.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (5.14)$$

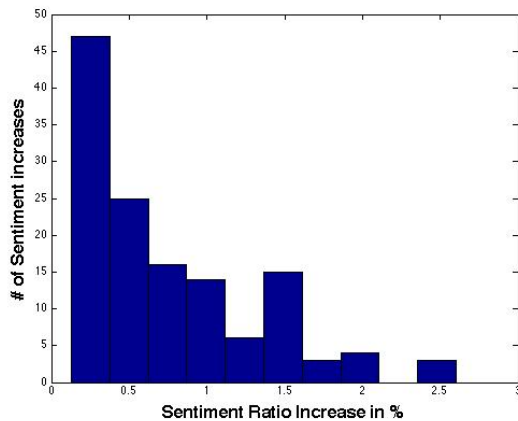
For all experiments we used the parameters in Table 5.6 unless otherwise noted. We motivated the choice of the membership functions in Section 5.5.2. The following four parameters are default values for optimisation methods discussed in 5.5.5. Lastly the retrain rate was set to be 28 days and 50 % of the entire data was used as an initial training set.

name	value
Input membership function	Gaussian
Output membership function	Linear
FIS generator	Fuzzy C-Means
Training epochs	500
Initial step size	0.01
Step size decrease rate	0.9
Retrain rate	28 days
Initial training window size	50 %

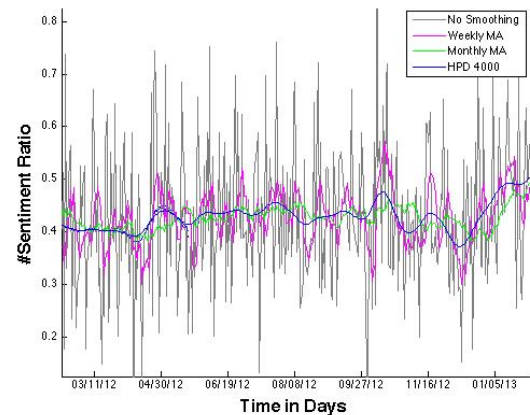
Table 5.6: Parameter Settings

5.6.2 Evaluation of Smoothing Function and Determining Lag Factor

In Chapter 5.3 we motivated the use of a smoothing function to aid our goal of predicting a trend. This kind of preprocessing becomes especially important for social media data as it is characterised by an extreme day to day volatility as illustrated in Figure 5.5a. We experimented with three different smoothing techniques 1.) the weekly moving average, 2.) the monthly moving average and 3.) the Hodrick-Prescott decomposition which we previously used to smoothen our price data. The effect of the smoothing functions is displayed in Figure 5.5b. Where the moving average seems to have, a delayed approximation of the true trend Hodrick-Prescott decomposition shows an exacter approximation of the underlying motion. We compared the different smoothing techniques by measuring the correlation increase of the respective smoothing functions in Table 5.7. Hodrick-Prescott decomposition improved the correlation between the commodity price and the sentiment features and was able to increase the correlation more than the moving average approaches.



(a) Price Increase Distribution of Sentiment Ratio



(b) Comparison of Smoothing Functions

Figure 5.5: Volatility of Social Media Features

We further investigated to what extent Twitter lags as an indicator. Heaven Kwak showed that

Smoothing	S1 Product	S2 Subcategory	S5 Poverty
No Smoothing	0.1463*	0.1340*	0.1719**
Weekly MA	0.2528***	0.1830**	0.3111***
Monthly MA	0.2889***	0.1466*	0.4847***
HPD	0.3586 ***	0.2673***	0.5957***
Significance: $p < .0005$ ***, $p < 0.005$ **, $p < 0.05$ *			

Table 5.7: Sentiment Correlation

Twitter can be seen similar to CCTV providers as a media for news where in some instances Twitter showed to be ahead of live broadcasting shows such as CNN [30]. If a causal relationship exists, how fast do prices react to Twitter conversations? Graphically this corresponds to shifting the sentiment ratio to the left. For each social media feature (51) we measured the correlation between one day and three days lag similar to [43]. We choose a window of three days because none of the attention spikes showed to last longer than three days (Section 4.4). Furthermore it has been shown that most trending topics only last for maximum seven days [30]. The experiment was performed on all three commodities. In case of a disagreement, we solved the tie through a majority voting. A lag of three days clearly favoured the correlation with 44 out of 51 (0.86 %) features and a one day lag for 7 out of 51 (14 %) features.

5.6.3 Price Model Results

The prediction results across different time horizons are highlighted in Figure 5.6. These results are obtained using the model described in Section 5.6.1 with the top 10 features obtained through the feature selection procedure explained in the previous section. The features used in the form of y_{t-2} are the top 10 obtained from the feature selection process. Looking at the results ANFIS performs exceptionally well for day to day predictions and as expected decreases its performance as the horizon increases. The commodity beef seems to be the easiest prediction task and further analysis will clearly show why. We observe that RMSE linearly increases over the period until prediction horizon 20. Prediction accuracy rapidly decreases and becomes unstable from then onwards. To analyse the results in more detail and to put the RMSE into context we turn our attention towards Figure ?? 5.8 5.9. For all three commodities, the predictions within one week are accurate. We can now also observe why beef had the lowest RMSE among the three commodities. Beef follows a clear upwards trend. Such long lasting motions are much easier to predict and as discussed in Section ?? our input variables from the intermediate past are descriptive if there is a clear underlying motion. For the model y_{t+20} ANFIS clearly overestimates a decrease and increase in price but still manages to capture the underlining trend. The overestimation seen in Figure ?? and Figure 5.9d strongly approximates the observed pattern in the training sample explaining the deviation from the actual prediction. The only commodity excluded from this behaviour is beef due to its strong underlying motion.

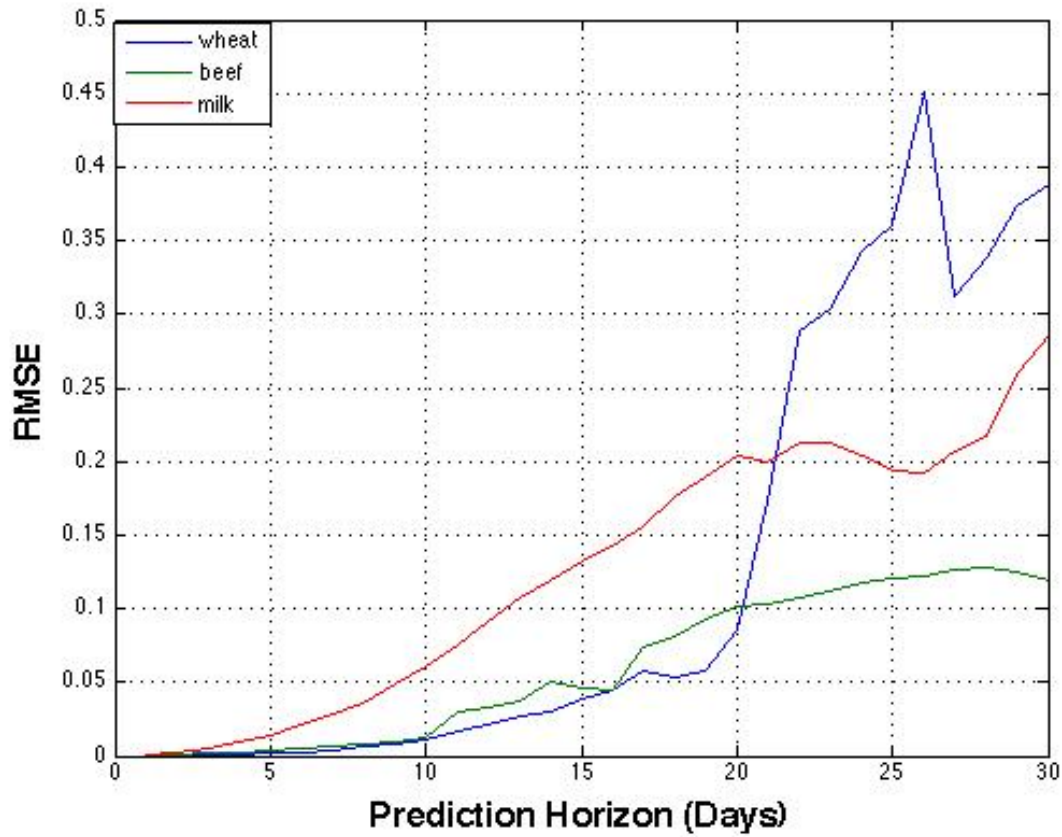
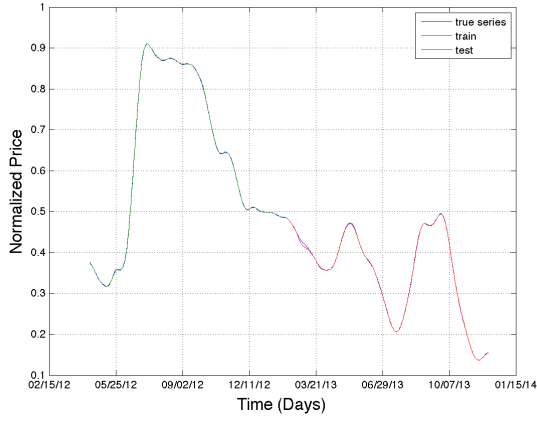


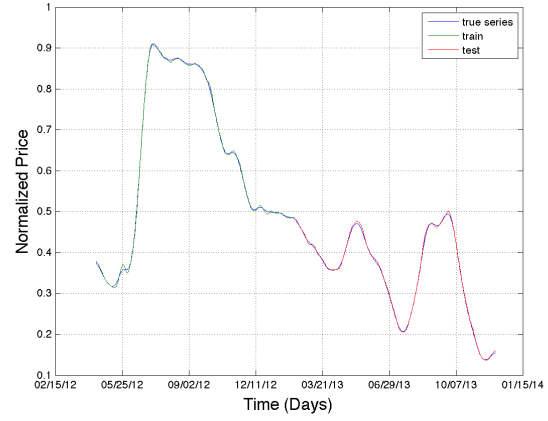
Figure 5.6: Prediction Accuracy - Benchmark

5.6.4 Social Media Model Evaluation

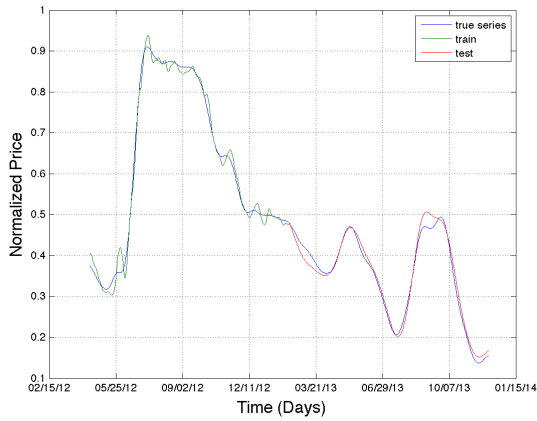
Unlike our benchmark paradigm, the social media model did not exhibit an increasing error rate with increasing horizon. We hence focused our attention on analysing day to day predictions. From Figure 5.11 it is apparent that our strengthening attempts have improved the prediction accuracy across all commodities. However, there is a clear difference in the quality of the predictions between the three commodities. For wheat, we were able to follow the real price line, not so for milk. Why does the prediction model favour the commodity wheat? Our first attempt was to investigate the linear relationship between the features and the price. The correlation analysis favoured the commodity milk however the difference is negligible. Our second intuition was to check whether the training data generalises well for the entire data set. We observed that the training and test dataset exhibited different behaviours. More specifically we found that certain features showed a negative correlation in the train set, however, a positive relationship in the test set. We experimented with various sizes of train sets but none of them notably improved the prediction results. Lastly, we were motivated to examine the content of the Tweets. The set of wheat Tweets might be more relevant to our given topic than the set of milk Tweets. For the purpose of this analysis, we computed the 10 most occurring nouns of the Tweets containing at least one term of the subcategory wheat, beef or milk along the 10 most frequently occurring bigrams. We use the number of keywords as a proxy of the overall number of tweets. This approach has limitations as we can not account for Tweets that contain more than one frequent term. For each of the frequent nouns we manually investigated the tweets to assess if the keyword captures relevant Tweets. In case of ambiguity we



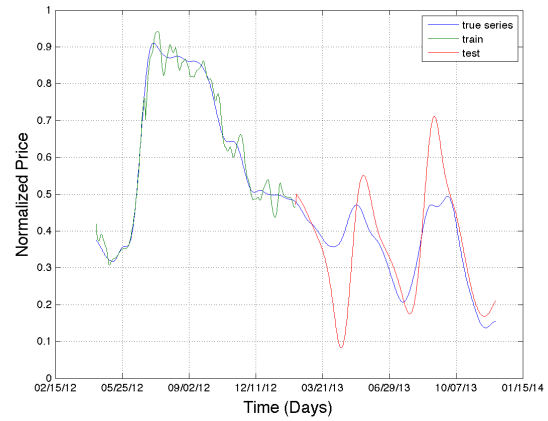
(a) 4 Day Horizon - RMSE 0.0015948



(b) 7 Day Horizon - RMSE 0.0030423



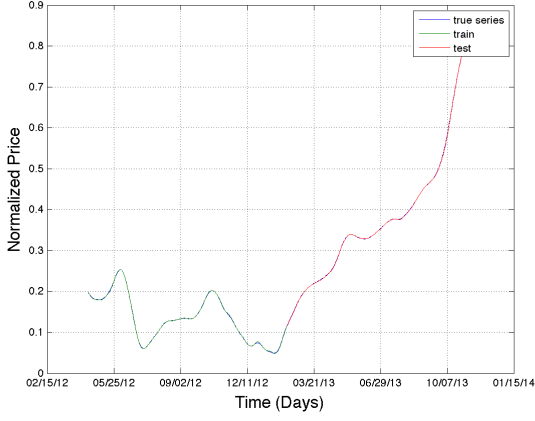
(c) 14 Day Horizon - RMSE 0.015192



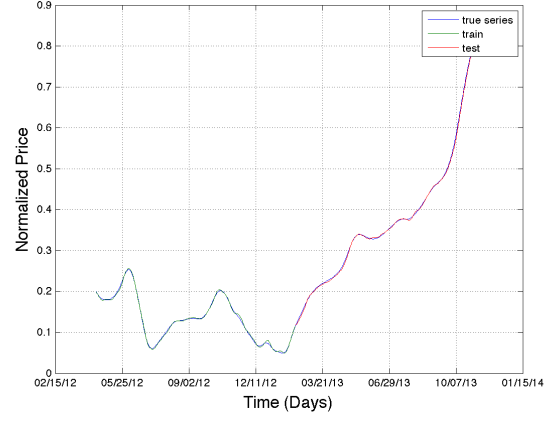
(d) 21 Day Horizon - RMSE 0.105

Figure 5.7: Benchmark Prediction Wheat

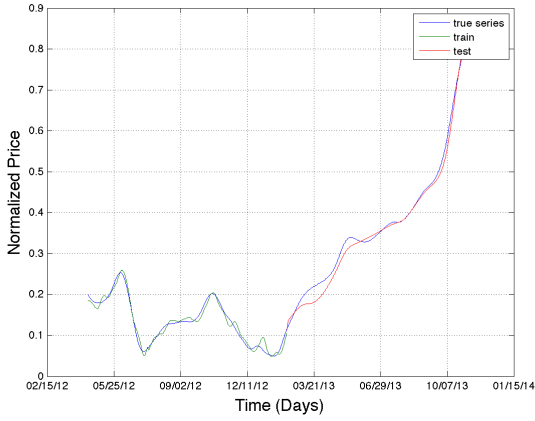
considered the bigrams to help us point to the right context. If the bigram pointed towards a relevant product we accepted the term as relevant else we subtracted the occurrence count of irrelevant bigram from the total count of the ambiguous term and accepted the remaining count as relevant. Furthermore we added the count of the irrelevant bigram to the total count of irrelevant terms. Consider the example of cream with a total keyword count of 1000. The bigram of ice cream occurs 600 times. We hence subtract the 600 from the 1000 and count 400 keywords as relevant and 600 as irrelevant. Following this approach we tried to quantify the entropy of our tweet. We did this analysis for the conversations centred around price, supply and poverty as the features crafted from these topics constitute the majority in our model. Notably, across all conversation topics, the discussions centred around wheat showed the highest relevance (i.e. the number of keywords retrieved deemed relevant was higher). Examples of accepted term counts were *bread*, *pasta*, *noodles* and *flour*. One of the only terms we rejected was *gluten*. The bigram pointed to the context of *gluten free* which is clearly not what we try to capture. Overall we approximate a value of 88 % of the tweets being relevant for the discussion centered around wheat. For beef frequently accepted terms were *beef*, *steak*, *cattle* *cow* and *meat*. A subset of words deemed irrelevant were *bitches* and *chicken*. Beef is used in urban language as *to have a grudge against somebody* which is why often the term *beef* and *bitch* were present in the same tweet. One of the bigrams pointed us to *chicken steak*. Such tweets off topic. In total we approximated 84 % of the tweets of being relevant. Lastly we



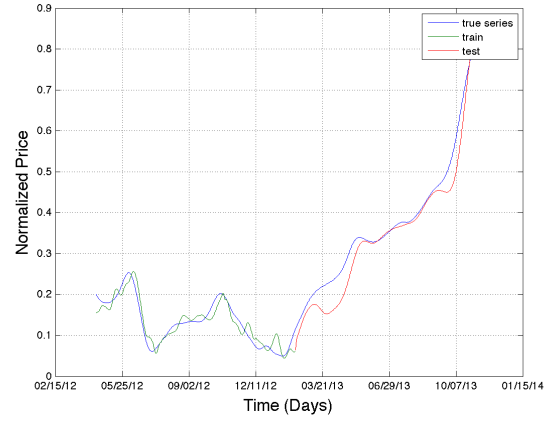
(a) 4 Day Horizon - RMSE 0.00086578



(b) 7 Day Horizon - RMSE 0.0059255



(c) 14 Day Horizon - RMSE 0.017041



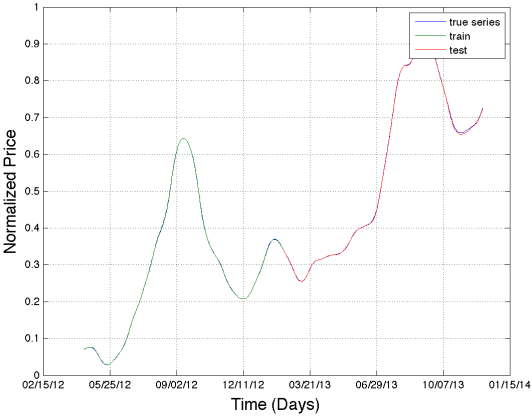
(d) 21 Day Horizon - RMSE 0.040397

Figure 5.8: Benchmark Prediction Beef

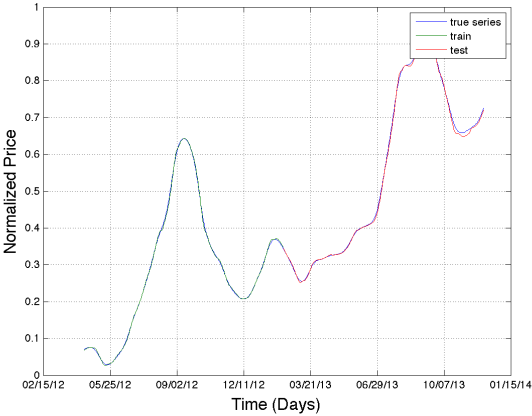
performed the experiment for the commodity milk. The only terms we marked as unambiguous were *milk* and *wheat*. The bigram containing the word cream showed that most of the conversation was centered around *ice cream*. Due to its distance relationship we consider this set of words as unrelated. Similarly for butter we observed that most of the tweets containing the word were referring to peanut butter. Again we consider the set of tweets containing peanut butter as irrelevant. Further terms that lead to irrelevant tweets were *cake*, *yogurt*. For milk we only observed a relevance of 60 %. On the basis of these findings, we conclude that the list of keywords used for milk retained a high number of Tweets that were not relevant, highly impacting the quality of the predictions.

We further investigate if we could determine a clear relationship between the features of wheat and the predicted outcome. The only variable that showed some describable relationship was the sentiment ratio of poverty. We calculate the sentiment ratio as $\#positive / \#negative$. It follows that with an increase of negative sentiment the price drops and rises as the ratio of positive sentiment increases.

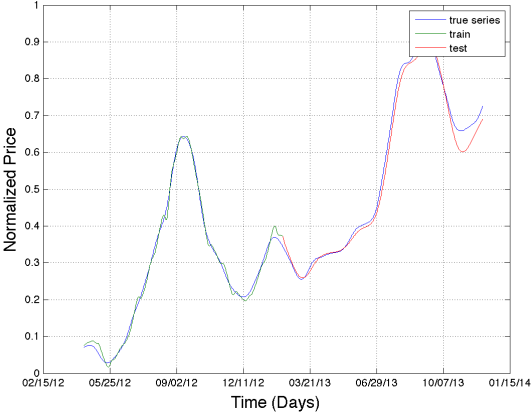
A visual analysis



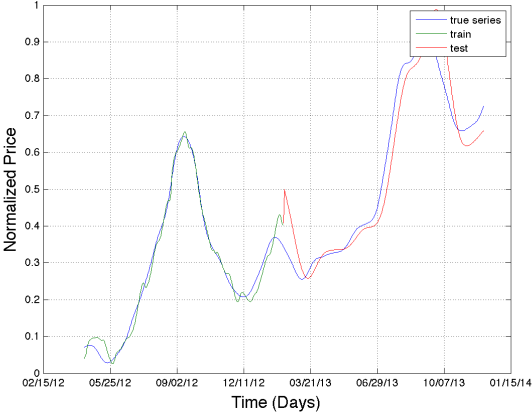
(a) 4 Day Horizon - RMSE 0.0015964



(b) 7 Day Horizon - RMSE 0.0049564

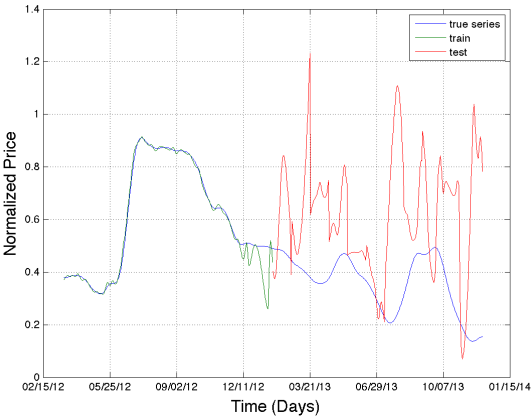


(c) 14 Day Horizon - RMSE 0.1185

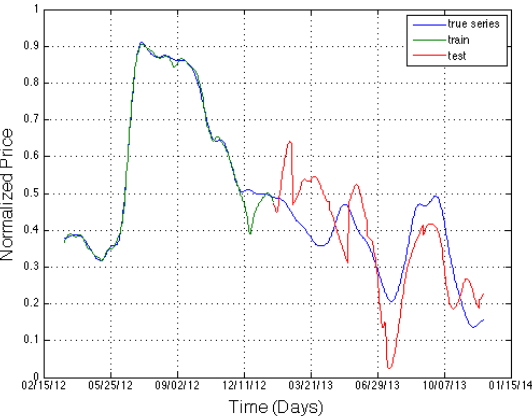


(d) 21 Day Horizon - RMSE 0.023702

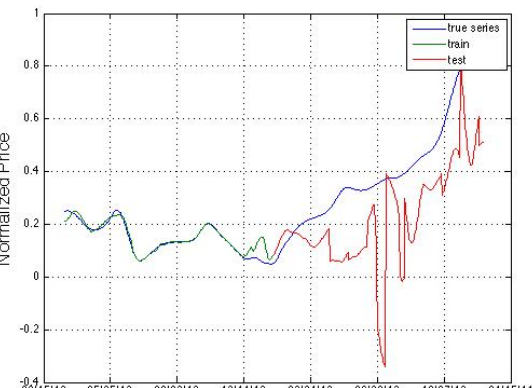
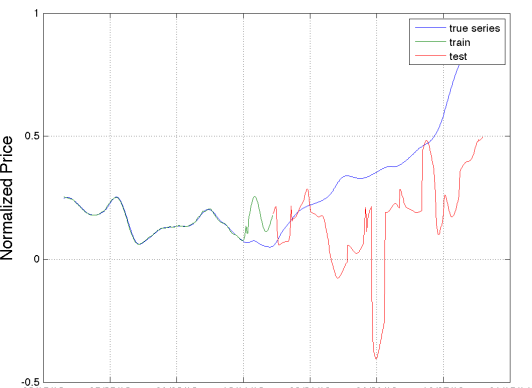
Figure 5.9: Benchmark Prediction Milk

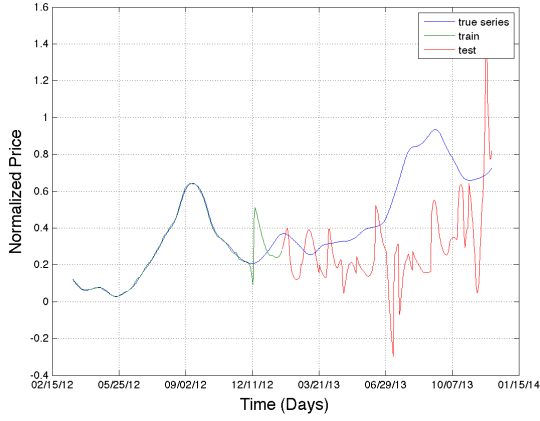


(a) 1 Day Horizon Wheat - RMSE 0.3580

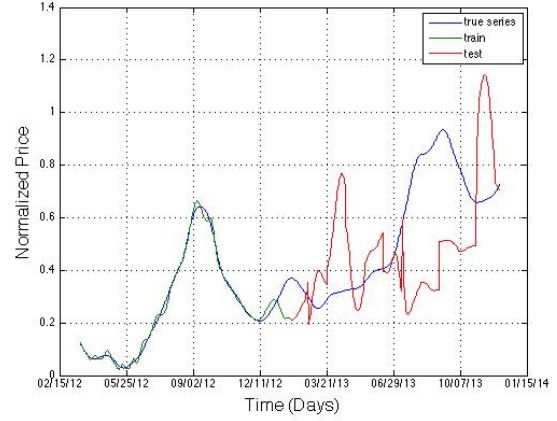


(b) 1 Day Horizon Wheat - RMSE 0.1083





(a) 1 Day Horizon Milk- RMSE 0.3782



(b) 1 Day Horizon Milk - RMSE 0.2960

Figure 5.11: Social Media Prediction

5.6.5 Mixture Model

As our mixture model exhibits similar behaviours to the price model, we show the exact predictions for each commodity in the Appendix C. Instead we focus our attention on the performance comparison of the three different models (e.g. price, social and mixture model). Figure 5.12 summarises our results and compares the three models. For all commodities, the social media model performs rather badly. We can hence conclude that a model solely based on social media feature will yield no accurate predictions. From the wheat forecasts we can see that both the mixture and price model have comparable results. Across all horizons, we recorded an average error of 0.0669 for the price model and an average error of 0.07315808 for the mixture model. It seems like the price models favour the intermediate future whereas the mixture model allows for more accurate predictions in the far foresight. To test this hypothesis we extended the analysis to y_{t+50} , however, rejected it as we could not record any clear performance improvements compared to the price model.

Similar patterns apply to the commodity beef. Since beef exhibits a strong and long lasting trend, the prediction accuracy is high for both models and a difference is hardly observable. At the end of the prediction horizon, the mixture model performs slightly better. For the price model, we observe an average error of 0.02816 whereas for the mixture model an error of 0.02814. For the commodity milk, the price predictions are clearly more accurate than the ones of the mixture model. This is attributed to the social media's high-level of noise and a lack of focus that we observed for the commodity milk.

5.6.6 General comments to the Results

We started our evaluation with defining three input models for a time series prediction based on historical price data, social media features and lastly a mixture model. Our general observation was that only our price model allowed to generalise well over over different commodities and price horizons. Social media features were shown to be more specific to a specific commodities. The model selection process clearly favoured sentiment measures over contextually sensitive tweets and raw volume. Similar for the mixture model the feature selection process resulted with a different set of features for each commodity and likewise favoured sentiment over twitter attention. Through our quantitative and qualitative analysis we clearly showed that an increase in twitter volume is not due to an increase in price.

A last observation we make regarding the input model is that prices that follow a clear trend

during a long period of time favour features crafted from historical price data where as commodities that follow a price trend that is nonlinear prefer social media features crafted from sentiment. The later can be explained by Robert Marx's theory that intermediate historical data features based on the intermediate past are predictive for prices characterised through long lasting motions.

For our concrete prediction we used the price model as a baseline. First experiments with the social media model failed to follow the price line which is why we experimented with different signal soothing techniques. Or conclusion was that a higher smoothing factor favoured future predictions which is not surprising as compared to the price data twitter volume and sentiment are characterised by a much higher day to day volatility. Twitter has been found to propagate news in some instances ahead of broadcasting channels such as CNN. We hence determined a lag factor between the twitter data and the commodity price data. We found that a lag of 3 days favoured the predictions.

Despite our efforts of strengthening the social media features predictions sole based on the social media input model were considerably less accurate then the ones achieved by the baseline. Our initial hypothesis was that daily twitter conversations fluctuates within a well defined baseline and that a sudden rise will be due to a strong increase in price. Although the increase in twitter volume could be attributed to events most of them have been found to be off topic. Nonetheless we found that for the commodity wheat a decrease in negative sentiment lead to an increase in price. in other words twitters users seem to be happy when the prices for commodity increase. However we were unable to genealise this claim among other commodities.

A stricter keyword selection procedure in Section ?? might migate this problem but could also lead to excluding informative tweets. It is a trade-off that requires further attention. The problem of irrelevant tweets is further fuelled by Twitter's gain in popularity. More often then not false event were caused by online celebrities. Such users are usually characterised by a high follower count causing a non-linear effect through a high number of retweets ?. Considering twitter meta data might help to distinguish spammers from trustworthy sources. Our experiments showed that the relevance or the entropy of the tweets highly affect the prediction accuracy. However for applying stricter filtering mechanisms we will need more relevant data. The key is relvence. We don't believe that more time series will help the model learn relevant patterns. We initially assumed that a larger training set would improve the accuracy. However our experiments showed that with growing training set the predictions did not increase in accuracy. We assume that older data loses its relevance to future predictions. Hence the above motivated optimisation steps will only be helpful if further datasources can be considered. We propose to specially daggregate the analysis to countries that are more sensitive towards price data. We are confident that this would allow for more robust models and help generalise claims.

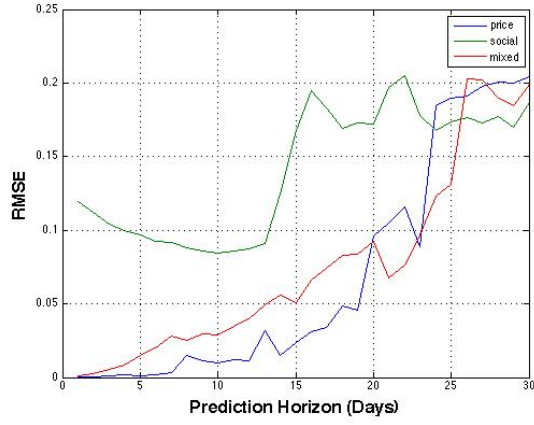
Although our social media features did not deliver the predictive power we were hoping for we still believe that coupled with price data it should be the preferred choice over a homogen price model.

We believe that the highest improvements in our thesis can be gained from identifying and removing spam

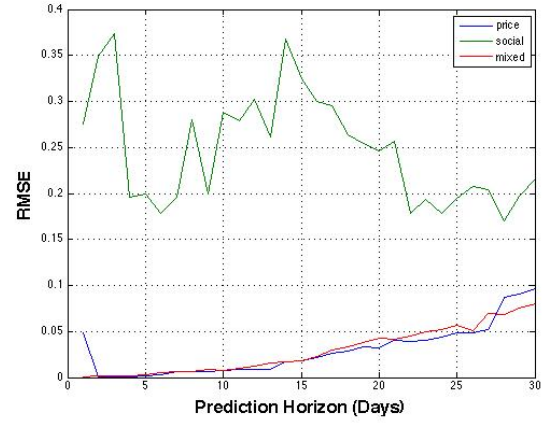
We believe excludng Identifying trusty users Are qualitative analysis showed that our sample is prune by noise.

made similar observations regarding prices with long lasting trends and conluded that

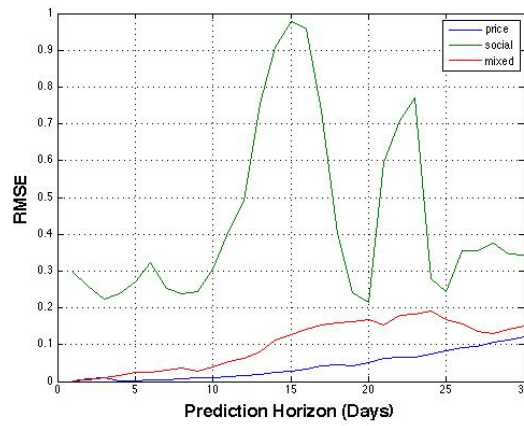
From a proposed set of 33 price features we found the variables in the intermediate past to be most descriptve of the price variance.



(a) Model Comparison Wheat- RMSE 0.016061



(b) 7 Day Horizon Milk - RMSE 0.031539



(c) 14 Day Horizon Milk - RMSE 0.11321

Figure 5.12: Social Media Prediction

To improve the overall predictions we experimented with different segmentation of the training data (50 % , 70 % , 85 %)

Conclusion

Food security has been shown to be a critical problem. A growing population and climate warming are making this a real threat not just for developing countries but developed countries alike. Although the topic receives a lot of attention from researchers, our observations have shown that monitoring attempts are mostly restricted to household surveys that fail to provide real-time information. Opinions, fears and expectations are increasingly represented in Social Media making it a valuable source of information for stronger policies that are inherently more evidence-based and provide accountability.

In this dissertation, we provide a semantic analysis of words indicative of food security. By extensively evaluating a word semantic analyser HAL we identified that a large window size of 10 and a small to middle sized corpus yielded the highest precision. The sparsity of certain commodities motivated us to structure our lexicon hierarchically. By considering categories (e.g. cereals), subcategories (e.g. wheat) and products (e.g. bread) we improved the recall by 110 %.

In our correlation analysis we show that on an aggregated level (e.g. meat) no real correlation exists however on a finer granularity (e.g. Sirloin steak) certain products exhibit a strong linear relationship of up to 0.7369 with the international Food price index. Our investigations of Twitter discussions showed that up to 13 % of the attention spikes can be attributed to food security objectives. The most discussed topic that we considered relevant to food security are concerns regarding the safety of food supply (e.g. mad cow disease).

In our time series analysis, we construct an Adaptive neuro fuzzy inference system to forecast a commodity price at some point in the future. Our results showed that the degree of attention and Twitter sentiment only explain a certain amount of the price variance. The success of a prediction is highly dependent on the quality of the keywords used to retrieve the Tweets. In other words, irrelevant Tweets can render predictions useless making it sensitive towards noise. On the other hand mixture models prove to be more reliable (e.g. historical price data and social media features) allowing us to accurately predict a trend four weeks into the future with a RMSE as low as 0.0683 on normalised price data.

To summarise our contribution:

-

6.1 Future Work

Our approach suffers from a bias towards English-speaking countries and countries that are highly developed. We would like to include different languages such as Spanish or Bahasa (Indonesia) as in [24] to capture a more diverse set of nations. Particularly Indonesia would

provide an interesting case as it is the third most active country on Twitter and food security issues are extremely prevalent with close to 20 million Indonesians being malnourished ¹. Language barriers would make it hard to select appropriate Tweets and further challenges are expected with gaining access to historical Twitter data. This extended analysis would shed some light on the different topics discussed between developed and developing countries and furthermore show if such Tweets are more indicative toward the global Food price index.

In Section 5.6.4 we identified that our Twitter data suffers from a lot of noise. Improvements could be achieved by using meta-data such as the number of followers to identify influential users or the field *statusescount* to identify potential spammers [35]. Further improvements are expected by training a classifier as in [1]. This would involve an additional crowd task to distinguish relevant Tweets from Tweets that are off topic.

¹<http://www.insideindonesia.org/food-security-in-indonesia-2>

Appendix A

Crowdfunder

For the categorisation of the keywords for our predictor lexicon, four crowd tasks were created. Here we detail the instructions given to the crowd workers for the four categorisation tasks in Section 3.3.2.

A.1 Categorise: Food Price

This is a categorization task centered around food security. Please categorize terms appearing in tweets about food to help us quantify users perception of Food Price. Overlaps may occur, i.e a term can potentially be indicative of both food price and food supply. Such keywords should always be classified as B. Likely .

Is the word or pair of words likely to be indicative of a user perception of food price?

A. YES, the term is indicative of food cost and/or can be used as a synonym of price

- pricy
- expensive
- cheap
- affordable
- bill
- receipt
- cost

B. LIKELY, the term might be indicative of food supply or food cost

- low
- high
- increasing

C. NO, the term is unlikely to be indicative of food cost

- when
- chair
- boy

D. Not in English, not understandable, other issues.

A.2 Categorise: Food Supply

This is a categorization task centered around food security. Please categorize terms appearing in tweets about food in order to help us quantify users perception of Food Supply. Overlaps may occur, i.e. a term can potentially be indicative of both food supply and food cost. Such keywords should always be classified as B. Likely.

Is the word or pair of words likely to be indicative of a user perception of food supply?

A. YES, the term is indicative of food supply

- available
- accessible
- lack
- amount
- number
- stock
- resource

B. LIKELY, the term might be indicative of food supply or food cost

- low
- high
- increasing

C. NO, the term is unlikely to be indicative of food supply

- when
- chair
- boy

D. Not in English, not understandable, other issues.

A.3 Categorise: Food Poverty

This is a categorization task centered around food security. Please categorize terms appearing in tweets about food to help us quantify users perception of Food Poverty. Overlaps may occur, i.e. a term can potentially be indicative of both food poverty and food needs. Such keywords should always be classified as B. Likely.

Is the word or pair of words likely to be indicative of a user perception of food poverty or the user perception of wealth?

A. YES, the term is indicative of food poverty or wealth

- starving
- donation
- wealth
- luxury
- profit
- help
- diabetes
- obesity
- healthy

B. LIKELY, the term might be indicative of food poverty and wealth or might be an indicator for food needs

- crave
- urgent
- must
- need

C. NO, the term is unlikely to be indicative of food poverty or wealth

- when
- chair
- boy

D. Not in English, not understandable, other issues.

A.4 Categorise: Food Needs

This is a categorization task centered around food security. Please categorize terms appearing in tweets about food to help us quantify users perception of Food Needs. Overlaps may occur, i.e. a term can potentially be indicative of both food needs and food poverty. Such keywords should always be classified as B. Likely .

Is the word or pair of words likely to be indicative of a user perception of food needs?

A. YES, the term is indicative of food needs

- love
- want
- hate
- favorite
- satisfied
- foodporn
- yum

B. LIKELY, the term might be indicative of food needs or food poverty

- crave
- urgent
- must
- need

C. NO, the term is unlikely to be indicative of food needs

- when
- chair
- boy

D. Not in English, not understandable, other issues.

Appendix B

Price Correlation

In Section 4.1.3 we performed a detailed analysis of the correlation between different food tweets and the international food price index. We show the remaining heat plots for the category dairy, sugar and oil.

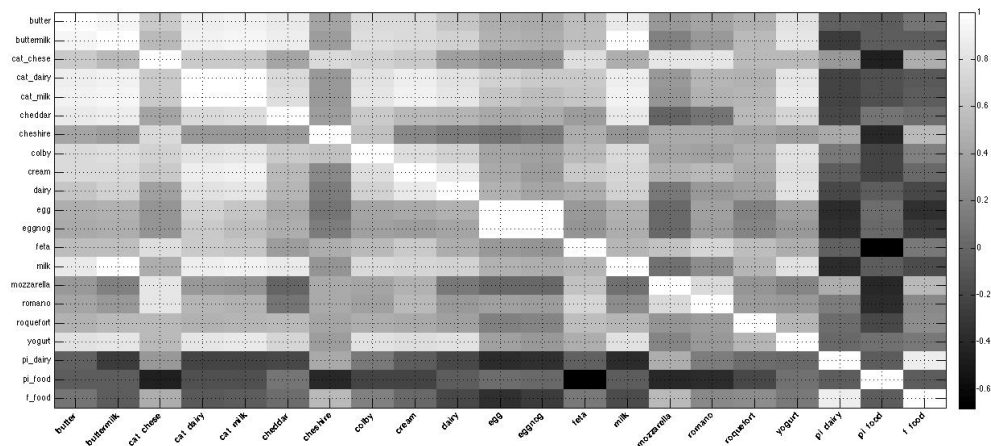


Figure B.1: Heatplot Dairy: Volume of Tweets per Keyword and per Category

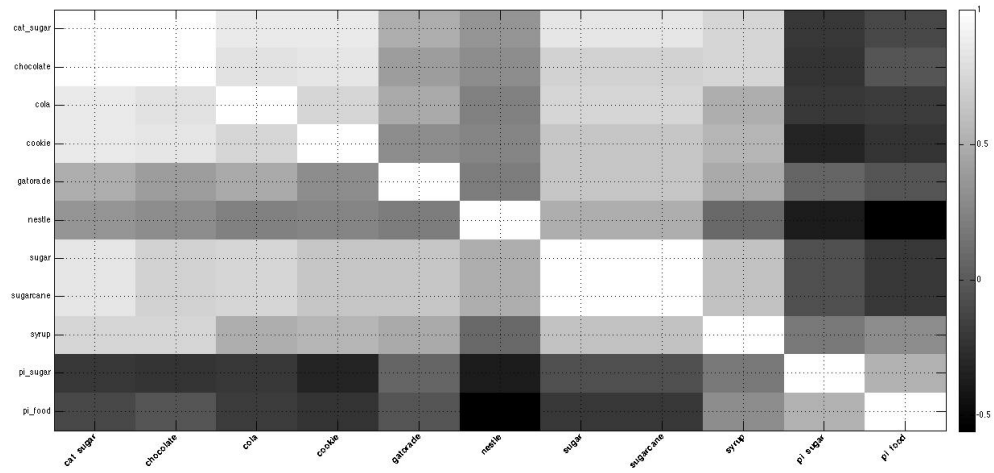


Figure B.2: Heatplot Sugar: Volume of Tweets per Keyword and per Category

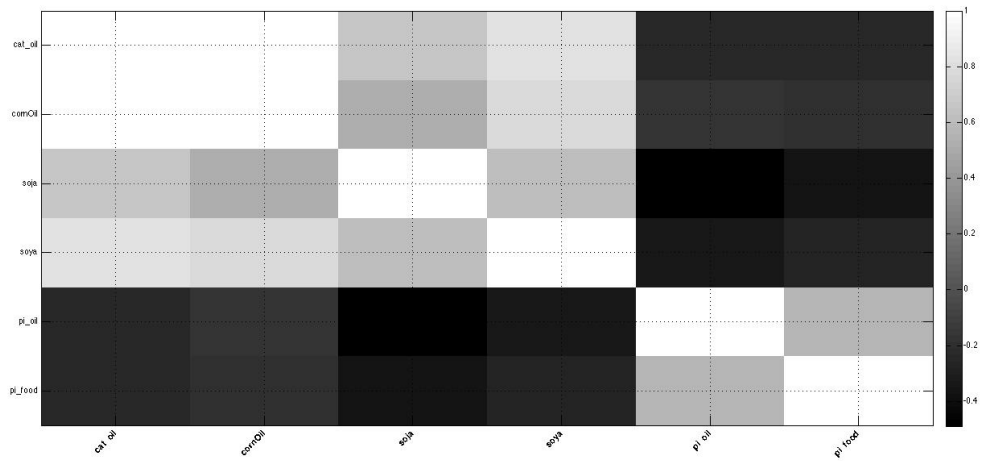
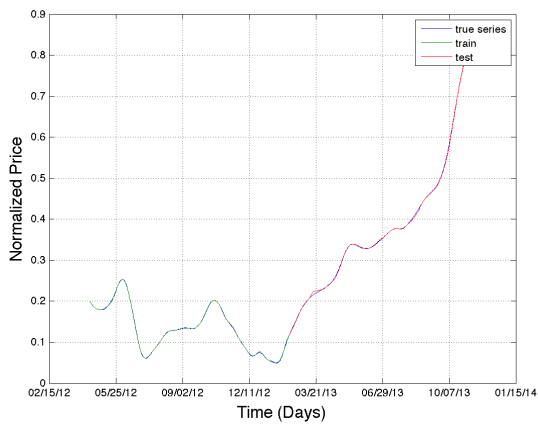


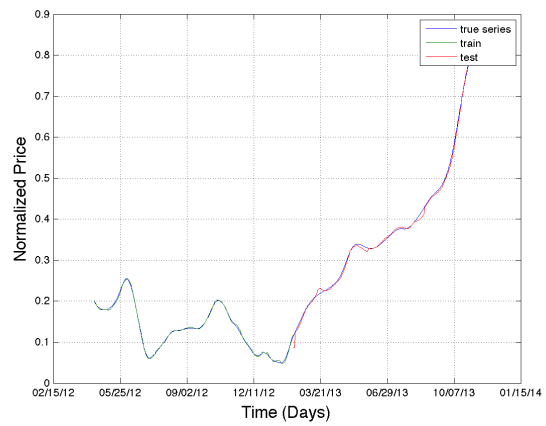
Figure B.3: Heatplot Oil: Volume of Tweets per Keyword and per Category

Appendix C

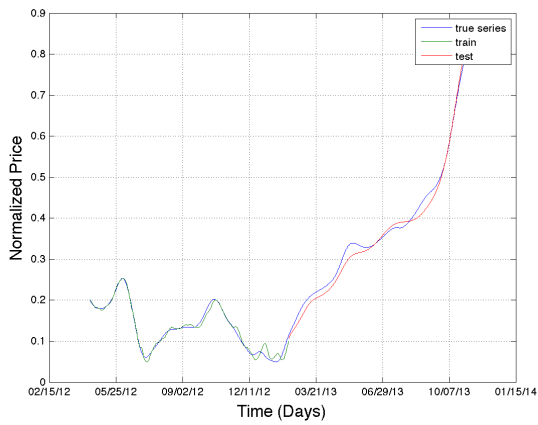
Time Series Modeling



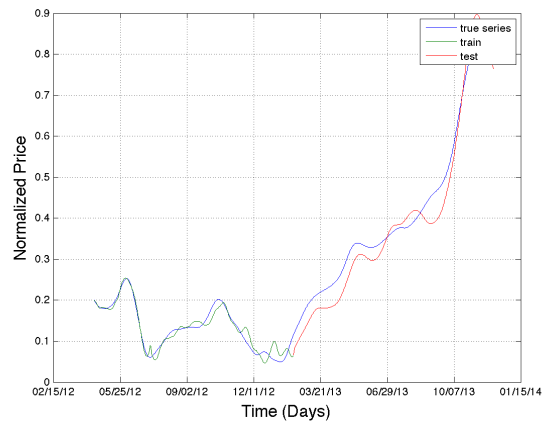
(a) 4 Day Horizon Beef - RMSE 0.0021991



(b) 7 Day Horizon Beef - RMSE 0.0065262

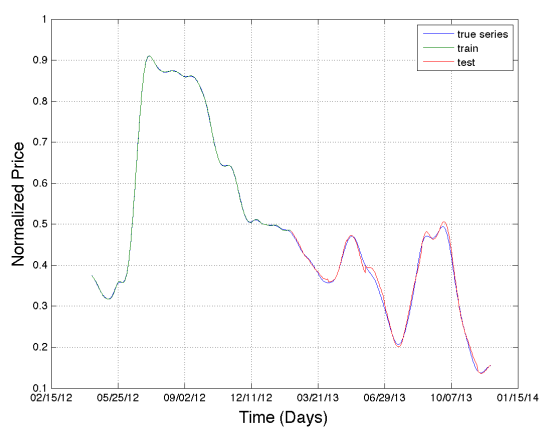


(c) 14 Day Horizon Beef - RMSE 0.016448

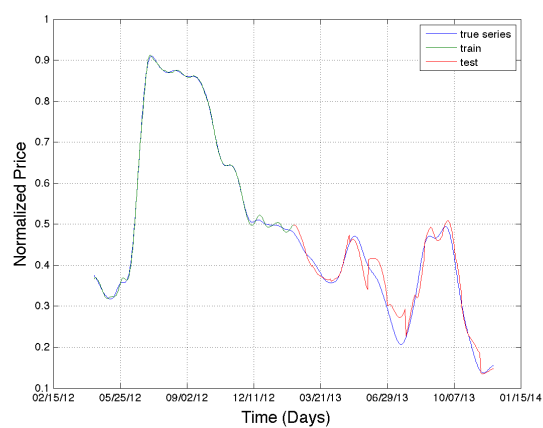


(d) 21 Day Horizon Beef - RMSE 0.042037

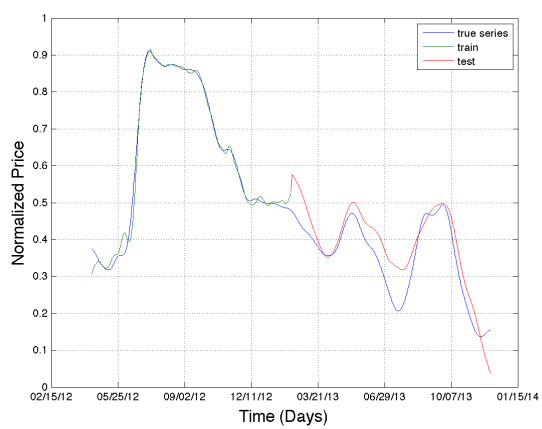
Figure C.2: Social Media Prediction



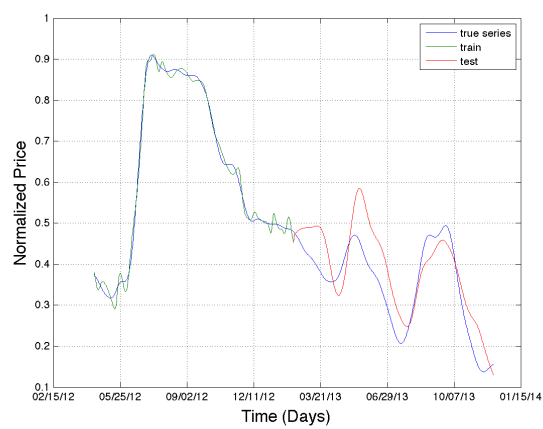
(a) 4 Day Horizon Wheat - RMSE 0.0084



(b) 7 Day Horizon Wheat - RMSE 0.0281

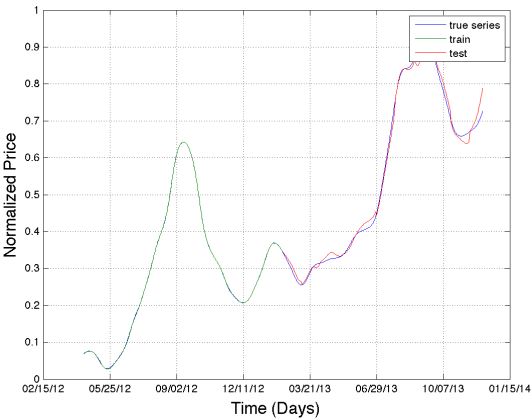


(c) 14 Day Horizon Wheat - RMSE 0.0557

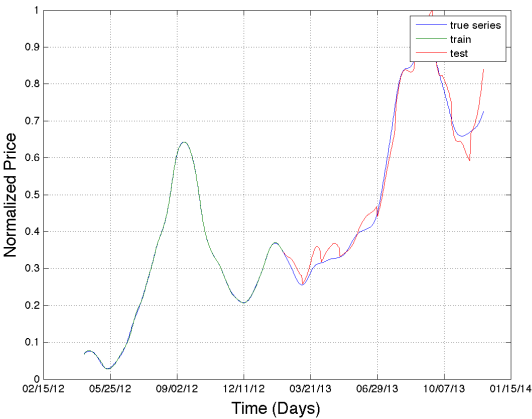


(d) 21 Day Horizon Wheat - RMSE 0.0673

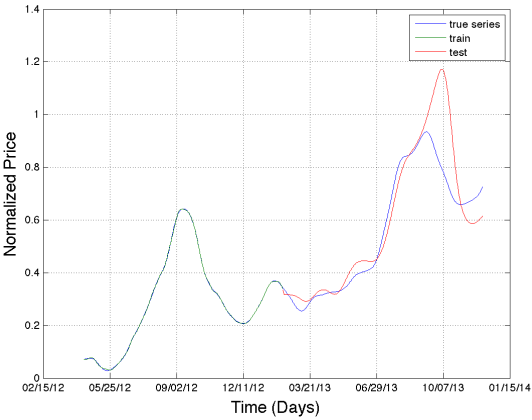
Figure C.1: Social Media Prediction



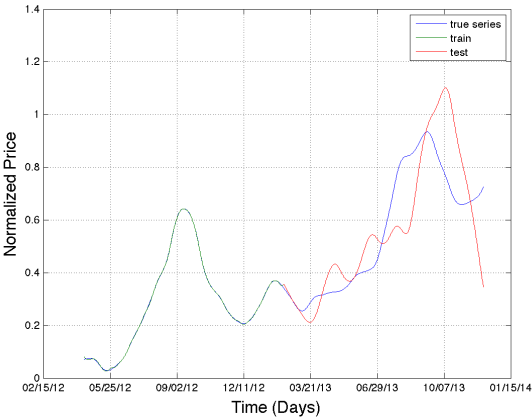
(a) 4 Day Horizon Milk - RMSE 0.016061



(b) 7 Day Horizon Milk - RMSE 0.031539



(c) 14 Day Horizon Milk - RMSE 0.11321



(d) 21 Day Horizon Milk - RMSE 0.15269

Figure C.3: Social Media Prediction

Bibliography

- [1] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. You tweet what you eat: Studying food consumption through twitter. *CoRR*, abs/1412.4361, 2014.
- [2] Philip C. Abbott, Christopher Hurt, and Wallace E. Tyner. What's Driving Food Prices? March 2009 Update. Number 48495, March 2009.
- [3] A. Olteanu C. Castillo N. Diakopoulos K. Aberer. Comparing events coverage in online news and social media: The case of climate change. In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM'15)*, 2015.
- [4] Gabriel Grill Joseph Boyd Stefan Mihaila Alexander Buesser Anton Ovchinnikov Ching-Chia Wang Duy Nguyen Fabian Brix. A monitoring and prediction toolset for volatile commodity prices in developing countries, 2014.
- [5] Jon Scott Armstrong. *Principles of forecasting: a handbook for researchers and practitioners*, volume 30. Springer Science & Business Media, 2001.
- [6] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '10*, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.
- [7] C. Benson and K. Kosonen. *Language Issues in Comparative Education: Inclusive Teaching and Learning in Non-Dominant Languages and Cultures*. Comparative and International Education: A Diversity of Voices. SensePublishers, 2013.
- [8] David R. Bild, Yue Liu, Robert P. Dick, Z. Morley Mao, and Dan S. Wallach. Aggregate characterization of user behavior in twitter and analysis of the retweet graph. *ACM Trans. Internet Technol.*, 15(1):4:1–4:24, March 2015.
- [9] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *J. Comput. Science*, 2(1):1–8, 2011.
- [10] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.
- [11] Curt Burgess and Kevin Lund. The dynamics of meaning in memory, 1998.
- [12] Christopher Chatfield. *The analysis of time series : an introduction*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton (Fl.), 2004.

- [13] John M. de Castro. Seasonal rhythms of human nutrient intake and meal pattern, 1991.
- [14] Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. Can blog communication dynamics be correlated with stock market activity? In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, HT '08, pages 55–60, New York, NY, USA, 2008. ACM.
- [15] Adeline Decuyper, Alex Rutherford, Amit Wadhwa, Jean-Martin Bauer, Gautier Krings, Thoralf Gutierrez, Vincent D. Blondel, and Miguel A. Luengo-Oroz. Estimating food consumption and poverty indices with mobile phone data. *CoRR*, abs/1412.2595, 2014.
- [16] Natalie Dixon, Ekaterina Yudin, Bruno Jakic, and Mark Mooij. Food Mood: Measuring Global Food Sentiment One Tweet at a Time. In *Proceedings of the Designing Interactive Systems Conference*, DIS '12, pages 793–794, New York, NY, USA, 2012. ACM.
- [17] John D'Errico. Interpolates and extrapolates nan elements in a 2d array, 2012.
- [18] Food and Water watch. Casino of hunger, 2009.
- [19] Rui Gaspar, Sara Gorjão, Beate Seibt, Luisa Lima, Julie Barnett, Adrian Moss, and Josephine Wills. Tweeting during food crises: A psychosocial analysis of threat coping expressions in spain, during the 2011 european ehac outbreak. *Int. J. Hum.-Comput. Stud.*, 72(2):239–254, February 2014.
- [20] Qiang Gong, Ming Liu, and Qianqiu Liu. Is momentum really momentum? international evidence, 2011.
- [21] Robert Hodrick and Edward Prescott. Post-war u.s. business cycles: An empirical investigation. Discussion Papers 451, Northwestern University, Center for Mathematical Studies in Economics and Management Science, 1981.
- [22] C. J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. 2014.
- [23] Scott H. Irwin, Dwight R. Sanders, and Robert P. Merrin. Devil or Angel? The Role of Speculation in the Recent Commodity Price Boom (and Bust). *Journal of Agricultural and Applied Economics*, 41(02), August 2009.
- [24] Pulse Lab Jakarta. Mining indonesian tweets to understand food price crises. *Food and Agriculture*, 2013.
- [25] Jyh-Shing Roger Jang. Fuzzy modeling using generalized neural networks and kalman filter algorithm. In Thomas L. Dean and Kathleen McKeown, editors, *AAAI*, pages 762–767. AAAI Press / The MIT Press, 1991.
- [26] Sarah Macnaughton Julian Parfitt, Mark Barthel. Food waste within food supply chains: quantification and potential for change to 2050.
- [27] Ieabeling Kaastra and Milton Boyd. Designing a neural network for forecasting financial and economic time series, 1996.
- [28] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

-
- [29] Igor Kononenko. Estimating attributes: Analysis and extensions of relief. pages 171–182. Springer Verlag, 1994.
 - [30] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.
 - [31] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 251–260, New York, NY, USA, 2012. ACM.
 - [32] Abby Levenberg, Stephen Pulman, Karo Moilanen, Edwin Simpson, and Stephen Roberts. Predicting economic indicators from web text using sentiment composition. In *International Journal of Computer and Communication Engineering*, Barcelona, Spain, February 2014. IACSIT Press.
 - [33] K. LUND and C. BURGESS. PRODUCING HIGH-DIMENSIONAL SEMANTIC SPACES FROM LEXICAL CO-OCCURRENCE. *Behavior research methods, instruments & computers*, 28(2):203–208, 1996.
 - [34] M.R.M.T. Manuel A. Hernandez. *Beyond the numbers: How urban households in Central America responded to the recent global crises*. Intl Food Policy Res Inst.
 - [35] Miranda Mowbray. The twittering machine. In *WEBIST (2)'10*, pages 299–304, 2010.
 - [36] Le T. Nguyen, Pang Wu, William Chan, Wei Peng, and Ying Zhang. Predicting collective sentiment dynamics from time-series social media. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '12, pages 6:1–6:8, New York, NY, USA, 2012. ACM.
 - [37] J. Nofsinger. Social mood and financial economics. In *The Journal of Behavioral Finance*, 144–160, 2005.
 - [38] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to Expect When the Unexpected Happens: Social Media Communications Across Crises, 2015.
 - [39] EC FAO Food Security Programme. An introduction to the basic concepts of food security, 2008.
 - [40] Marko Robnik-Sikonja and Igor Kononenko. An adaptation of relief for attribute estimation in regression, 1997.
 - [41] P. Rutsaert, Z. Pieniak, A. Regan, A. McConnon, Dr. M. Kuttschreuter, M. Lores, N. Lozano, A. Guzzon, D Santare, and W. Verbeke. Social media as a useful tool in food risk and benefit communication? a strategic orientation approach, 2014.
 - [42] Pavel Savor and Mungo Wilson. How Much Do Investors Care About Macroeconomic Risk? Evidence from Scheduled Economic Announcements. *Journal of Financial and Quantitative Analysis*, 48(02):343–375, April 2013.
 - [43] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction.
 - [44] Stanford. Corenlp. 2011.

- [45] Getaw Tadesse, Bernardina Algieri, Matthias Kalkuhl, and Joachim von Braun. Drivers and triggers of international food price spikes and volatility. Number 0, pages 117 – 128, 2014.
- [46] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December 2010.
- [47] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Weppe. Predicting elections with twitter: What 140 characters reveal about political sentiment, 2010.
- [48] Twitter, 2013.
- [49] S. Van Belleghem. *The Conversation Company: Boost Your Business Through Culture, People and Social Media*. Kogan Page, 2012.
- [50] Andreas S. Weigend and Neil A. Gershenfeld, editors. *Time series prediction : forecasting the future and understanding the past : proceedings of the NATO advanced research workshop on comparative Time series analysis held in Santa Fe, New Mexico, May, 1992*, Santa Fe Institute studies in the sciences of complexity. Proceedings volumes, 1992, Santa Fe, 1993. Addison-Wesley.
- [51] Robert West, Ryen W. White, and Eric Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs, 2013.
- [52] Steve Whittaker, Loren Terveen, Will Hill, and Lynn Cherny. The dynamics of mass interaction. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, CSCW '98*, pages 257–264, New York, NY, USA, 1998. ACM.
- [53] Dongrui Wu. Twelve considerations in choosing between gaussian and trapezoidal membership functions in interval type-2 fuzzy logic controllers. In *FUZZ-IEEE*, pages 1–8. IEEE, 2012.
- [54] Jingtao Yao and Chew Lim Tan. A case study on using neural networks to perform technical forecasting of forex, 2000.
- [55] L.A. Zadeh. Fuzzy sets. *Information Control*, 8:338–353, 1965.
- [56] Wenbin Zhang and Steven Skiena. Improving movie gross prediction through news analysis. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '09*, pages 301–304, Washington, DC, USA, 2009. IEEE Computer Society.
- [57] Xue Zhang, Hauke Fuehres, and PeterA. Gloor. Predicting asset value through twitter buzz, 2012.