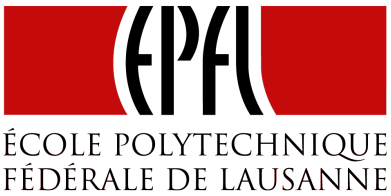


Twitter as an Indicator of Food Security

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

List of Figures

List of Tables

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. Especially global soaring food prices seem to have a negative effect by transmitting into rising food inflation rates in domestic markets. The recent food crisis in 2011 has driven 100 of millions into extreme poverty causing riots, falling markets and collapsing governments as experienced during the 2011 revolution that swept the middle east. Such suffering is likely going to increase in the future as our population grows and our richer diets call for more resources. However producing more resources is an extremely challenging task as we face rising energy prices and global warming. 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 consequence an increase of health-related diseases such as obesity or diabetes. 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. 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. Every second people generate large amounts of data. Partly voluntarily, partly involuntarily through their mobile phones and social media. As people use those, services they leave traces in the data and if their lives change for the better or the worse so do those traces. 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 [1]. Cleaned and aggregated this provides a valuable opportunity for *studying food security indicators on Twitter*. Understanding if and how price fluctuations 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.

[1] http://www.businessinsider.com/most-discussed-topics-on-social-media-2013-5

How do people talk about Food Security Indicators on Twitter? - Can we quantify the Semantics of indicative Words?

Food security Objects are well defined, but how do we capture Tweets that fall into these segments and how do we distinguish them from irrelevant discussions? This inherently comes down to finding contextually similar words to our given food security objectives. Most frequently a local co-occurece analysis is performed, however in this work we make a case that these approaches fail to capture a lot of indicative terms. Hence, *we quantitatively analyse the semantics of a word by using HAL.* Tweets have characteristics that are very different from classical text. As a result, *we extensively evaluate the performance of HAL and propose a set of metrics for analysing the semantics of social media*. Lastly, certain raw products are only very sparsely represented in Twitter conversations. To circumvent this problem *we introduce a methodology to increase the coverage of our lexicon.*

Can Social Media Data provide insight into rising Food Prices or Price Instability? - Are Food Security Objectives even discussed on Social Media?

To address these 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. We also apply a methodology introduced by [3] to automatically detect events 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.

Is it possible to use Social Media to model Commodity Prices? Can we predict the Price of a specific Commodity at some point in the Future?

It is widely assumed that the food crisis in 2008 was accelerated by speculative actions on the commodity future markets [11]. The prices of U.S. corn tripled from $94 to $281 during a period of only 3 years. Those basic goods constitute the diet and the currency of the poorest two billion people [32] strongly affecting their household income and purchasing power. With a Pearson correlation of 0.8436 there is a strong correlation with the Food Security Index making it an important indicator. We hence monitor volatile commodity prices by *introducing an Adaptive neuro-fuzzy inference system for time series modelling*. *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 with an RMSE as low as 0.0683 on normalised price data*.

The rest of this document is structured as follows: 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

Literature review goes here....

Chapter 3

Social Media Data Acquisition

In this chapter, we describe how we filtered for relevant Tweets using two different lexicons. The food lexicon contains keywords with food related terms (e.g. *rice, wheat, milk*) where the predictor lexicon contains terms with factors influencing the price and supply of the goods (e.g. *pricey*, *cheap*, *available*). We downloaded 2 TB of Tweets from the internet archive [1] over a span of October 2011 - September 2014. The filtering process resulted with 29 M food relevant Tweets.

Firstly we motivate and detail an algorithm Hyperspace Analogue to Language (HAL) [23] for candidate term selection. We then experimentally evaluate the different metrics influencing the performance of HAL, discuss different frameworks for selecting candidate terms for our food, respectively predictor lexicon and lastly explain the filtering process of the Tweets.

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 $N \times N$ 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 [7] 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.

[1] https://archive.org/details/archiveteam-json-Twitterstream
[2] https://code.google.com/p/airhead-research/wiki/HyperspaceAnalogueToLanguage

	Barn	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 *available*. Those two words seem unrelated, but given the context of food they express the same thing. Namely an abundance of food. Burgess and Lund [7] motivate that through the context, they possess elements of item's similarity but by themselves they would never be considered words with similar meaning. They further note that they 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 the seldom locally co-occur. They do however occur in the same context. Those observations motivated us to deviate from the commonly used co-occurrence analysis and take a step further to improve the precision of our filtering framework.

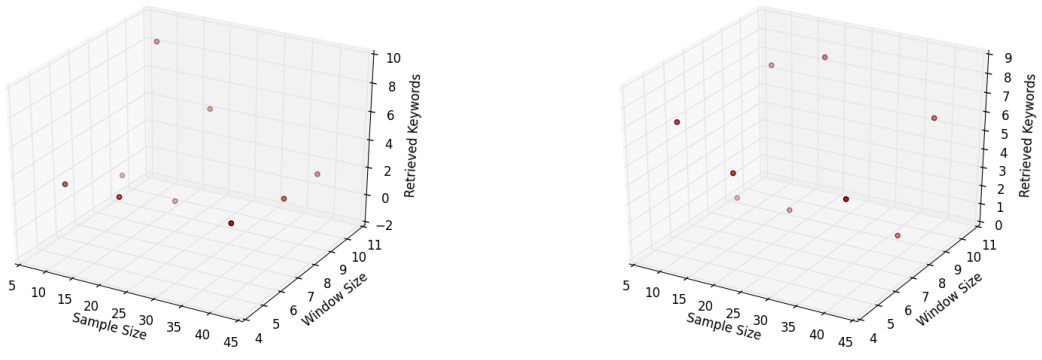
3.1.2 Experimental Evaluation

To increase the recall 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. Twitter is particularly suited for studying the meaning of words. Not only does it cover a wide vocabulary targeted around food but is further a close approximation of every day speech. This is very different from normal corpora which are usually based on specialised dictionaries [7].

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 quality of the results. 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 result. According to [23] 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.

3.1.3 HAL Performance Results

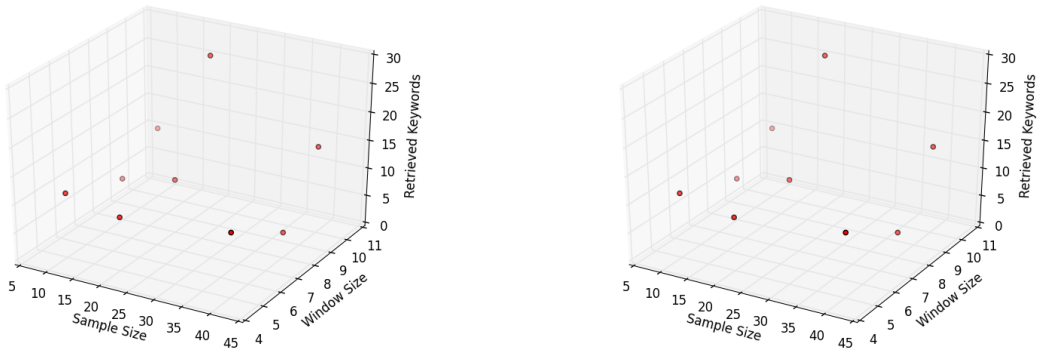
In Figure 3.1 and Figure 3.2 we observe that for all categories HAL performed best with a window size of 10 which contradicts the findings of [23]. Additionally, we see that the smaller sample sizes (10%, 20%) consistently produce more relevant keywords than the large sample size (40%). A large sample 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.



(a) HAL - Price

(b) HAL - Supply

Figure 3.1: HAL Evaluation for Price and Supply



(a) HAL - Needs

(b) HAL - Poverty

Figure 3.2: HAL Evaluation for Poverty and Needs

3.1.4 Discussion

We observed that HAL has a good precision given a high similarity threshold. For the top 20 keywords we evaluated a precision of 100 % for food relevant terms. In the top 20 we found other food items building the majority of the retrieved words. However, the precision varies with the window and sample size. These variables, as our evaluation has shown, are very

much dependent on the form of the corpus.

With decreasing similarity HAL highlighted some topics indirectly associated with food security. For example there was a high percentage of country names that showed a clear association with food. Where the majority of the retrieved countries such as Thailand, Bali [3] or the cities Singapore and Paris [4] are considered to be famous holiday destinations for food lovers other retrieved countries such as Pakistan, Syria, Jakarta India or the Philippines [5] are cities with a history of food insecurity and political unrest.

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 [6] along the ten most important staple foods that feed the world [7].

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 FAO [8].

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* and *other food of interest* build the top layer of our hierarchical representation as shown in Figure 3.3.

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 **1.)** can simply be the name of a category and subcategory (e.g. *meat, beef*) or **2.)** 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. In the production process of food items most factors that influence the price are static and predictable. One of the only fluctuating and unknown factors is the price of the raw product or in our case the commodity. 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. Lastly, 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.

[3] http://www.nomad4ever.com/2008/08/24/top-10-popular-foods-of-asia-explained/
[4] http://www.hellotravel.com/stories/best-food-cities-in-world
[5] http://foodsecurityindex.eiu.com/Country
[6] http://www.imf.org/external/np/res/commod/index.aspx
[7] http://knowledge.allianz.com/demography/health/?767/the-worlds-staple-foods
[8] http://www.fao.org/worldfoodsituation/foodpricesindex/en/

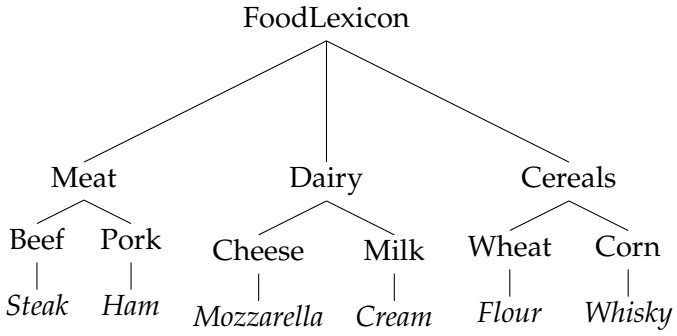


Figure 3.3: Food Lexicon - Hierarchy

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* $\rightarrow$ *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 $(c_1, c_2 ... c_6)$ 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 consider $k_i \in K_{final}$. For all keywords $k_i \notin K_{all}$ the condition of it being frequent is evaluated and if *True* added to the category *0ther food of interest* $c_6 \in K_{final}$. Food commodities that could not be assigned to any of the six categories were discarded (e.g. *orange, cocoa, onion*). Lastly, the set K_{final} was further enriched by using food products p_i that have been identified by [1] only if $p_i \in K_{all}$. 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 similar terms. Hand select those that are $\in K_{all}$ and add it to K_{final}.

The keyword set K_{final} was used to perform exact term matching on the Tweets collected from the internet archive. The resulting set of keywords in K_{final} forms our Food Lexicon.

Lexicon / Subset s	Keywords (i: from initial set, e: from [1] , h: from HAL space)
K_i Food	meal (i), meals (i) ,food (i), foods (i), wheat (i), rice v, 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_f 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_f 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_f 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 oi (i), sunflower oil (i), soybeans (i), soya (i), soy sauce (i), soja (i)
K_f Sugar	sugar (i), sugarcane (i), syrup (e), energy drink (e), cola (e), chocolate (e), nestle (e), cookies (h), cupcakes (h)
K_f 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 Food and Agriculture Organization of the United Nations (FAO) measures food security based on four dimensions namely *Access*, *Availability*, *Stability* and *Utilisation*. Where *Access* mostly captures the supply of food, *Availability* is concerned with the affordability of the basic goods. *Utilisation* captures the nutritional value of food and lastly *Stability* is a measure of the other three dimensions over time. "For food security objectives to be realised, all four dimensions must be fulfilled simultaneously" [27]. We mimic FAO's classification in our predictor lexicon.

To capture Tweets associated with the category *Access* we filter for the term price as in [15] but improve the recall by including synonyms and contextually similar words (e.g. *expensive, bill, cost, affordable*). We filter *Availability* related Tweets in a similar fashion by matching keywords that are synonyms of the word supply (e.g. *available, amount, stock*) as in [4]. Unlike [1] we do not measure food *Utilisation* by observing the exact diet but by filtering for terms that capture the people's food needs (e.g. *love, want, yum*). As a measure of *Stability*, we focused our attention on economic stability. Keywords in the context of poverty were selected similar to [29] [4] (e.g. *starving, donation, help*).

3.3.1 Candidate Predictor Term Selection

HAL, to the best of our knowledge, has not been used in previous work for term selection. Hence, we drafted two different frameworks for our evaluation. As a reminder K_{final} refers to the set of terms in our Food Lexicon. F_c, on the other hand, refers to a corpus drafted from all food relevant Tweets. Finally, the manual selection of the keywords was done through crowd flower [9].

Framework 1

1.) $\forall k \in K_{final}$ choose the keyword k with the highest occurrence form the entire sample F_c. Let's call it k_{max}

2.) $\forall w \in F_c$ perform a similarity measure with k_{max}

3.) Retrieve the 500 most similar words and hand-select the words that occurs in the synonym lexicon thesaurus [10] for supply, price, poverty and needs.

4.) For each of those hand-selected words apply HAL

5.) For each predictor category retrieve the 500 most similar words and let crowd workers select the relevant terms.

The high-level intuition of this procedure is as follows. **1.)** will give us the most prominent food term. This is most likely going to be something general such as the keyword *food*. **2.)** and **3.)** will allow us to identify the most contextual similar keywords for each category. So the keyword is retrieved that is most likely used to describe supply in the context of food. In **4.)** and **5.)** we aim to retrieve similar words that could describe supply but maybe appear more frequently in different contexts. In other words, we aim to find synonyms here.

Framework 2

1.) $\forall w \in F_c$ perform a similarity measure with the keywords supply, price, needs and poverty

2.) Retrieve the 500 most similar words and let crowd workers select the relevant terms

Instead of finding a keyword that is a synonym of a predictor category as in Framework 1 we simply use our predefined category names as a base to retrieve contextually similar words.

For the discovery of predictor terms, we used Framework 2 for three reasons. **1.)** Framework 1 did not retrieve us the desired keywords for all categories. **2.)** between the results of Framework 1 and 2 there was a substantial overlap and **3.)** Framework 2 is more efficient to execute. This is particularly important since creating the HAL space is computationally expensive. The final lexicon was further enriched by including synonyms from thesaurus for supply, need, poverty, and price. The terms of the final predictor lexicon are presented in Table 3.3 along the source of the keyword.

Unlike the annotation of our food related terms, allocating a term to a specific category was a more challenging task due to the ambiguous meaning of certain terms. In the following section we explain in more detail how we assigned a keyword to a given category by using crowdflower.

3.3.2 Annotation and False Positive Removal of HAL Results

To annotate the term results of HAL 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.**

[9] http://www.crowdflower.com/
[10] http://www.thesaurus.com/

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.

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 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. 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

We manually assessed the annotations produced by crowd flower 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 (%) from 106.

The high disagreement for the supply category was 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*).

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 commodity.

Lexicon / Subset *s*	Keywords (h: from HAL space, t: from thesaurus)
Food Supply	*supply*, 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), expanse (t), extent (t), flock (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)
Food Price	*price*, 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)
Food Poverty	*poverty*, 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)
Food Needs	*need*, 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), looooove (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), oppluence (t), plentitude (t), prosperousness (t)

Table 3.3: 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.

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. Given the large size of the dataset efficiency was also a concern. We experimented with both string.split() and a tokenizer provided by the Natural Language Toolkit [22]. String.split() proved to be more *tweetable*. 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 [33] or Stanford CoreNLP [31] 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 [14] 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" [14]. 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.

Chapter 4

Analysis

This chapter will 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.

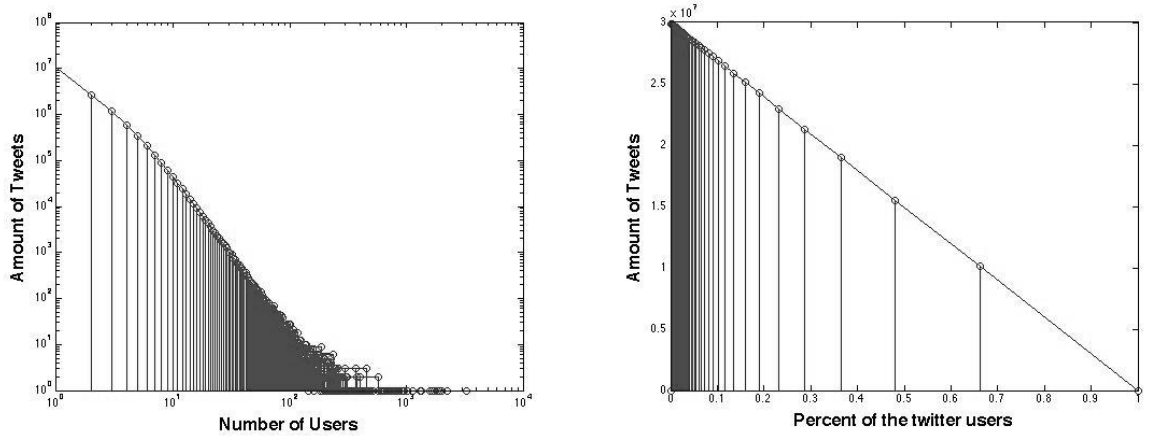
By drawing some basic statistics we want to emphasis the general popularity of food among the Twitter users and describe the term distribution of our different commodities. In the analysis we aim to show how food terms relate to each other and how they compare to indices which are intrinsic food security indicators (e.g. *affordability and availability of food*). Lastly, we investigate to what extent food security and market fundamentals are present in social media discussion.

4.1 User Distribution

Twitter is a social network and in general such networks follow a power law distribution [35]. We see in the bellow Figure **5.3a** and Figure **5.3b** 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 [5] suggest the contribution participation level of 80 %, 20 % does not seem to apply to Tweets about food. In Figure 4.1b we can see that the curve is almost linear. About 50 % of the Tweets are caused by 50 % of the users. This deviates highly form the normally observed 80 %, 20 % ratio. We assume that this is due to the wide spread interest of the topic.

4.2 Food Term Distribution

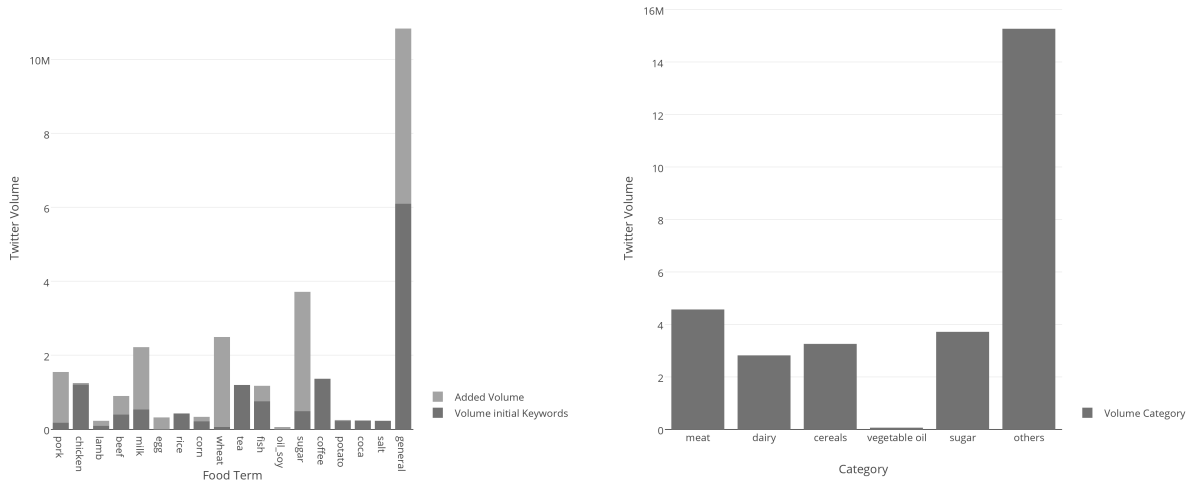
Our framework for the data acquisition successfully increased the total volume of food related Tweets. From an initial 13.7 M Tweets we raised the entire volume by 110% to a total of 29.9 M food related Tweets. The distribution of the volume per food term is displayed in Figure 4.2a. We illustrate in 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 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



(a) LogLog: Number of Tweets per User (b) Distribution: Number of Tweets per User

Figure 4.1: Volume of Tweets per Keyword and per Category

different categories with the only outlier being sugar [27]. We assume this is due to the highly popular products *coca cola* and chocolate which caused alone 70 % of the sugar related Tweets.



(a) Overall Distribution (b) Category Distribution

Figure 4.2: Volume of Tweets per Keyword and per Category

4.3 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 Argriculture Organisation of the United Nations

[1] and commodity quotes from candle [2]. 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 [27]. 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 months. We approximated those values by taking the average of the previous and the following month.

4.3.1 Results

Between the meat subcategories there is a positive linear relationship in the range of 0.7264 to 0.9361.This means if chicken increases in volume so does beef and pork. 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 categories and the three price indices. Incase of a slight correlation most of the categories are negatively correlated to 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. 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.

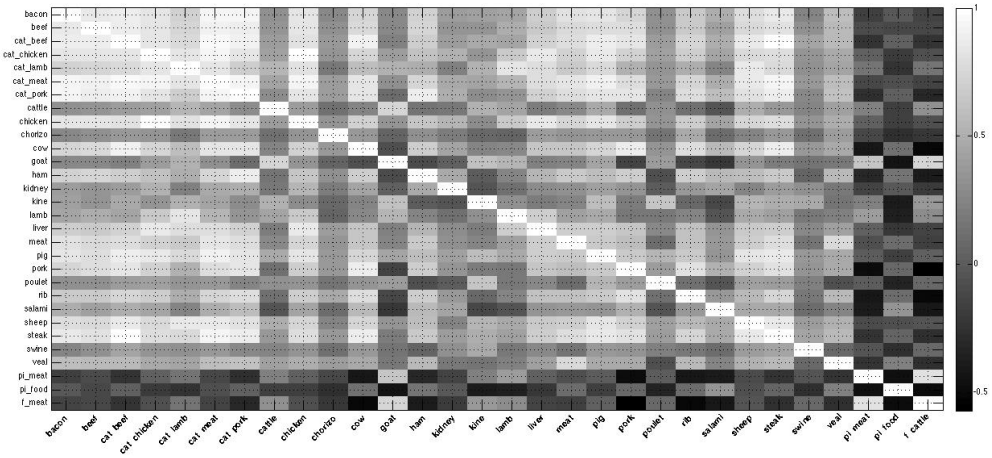


Figure 4.3: Heatplot Meat: Volume of Tweets per Keyword and per Category

[1] http://www.fao.org/worldfoodsituation/foodpricesindex/en/
[2] https://www.quandl.com/

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 then cereals do [27]. Furthermore the product pasta has a strong linear relationship with the commodity price of 0.7212.

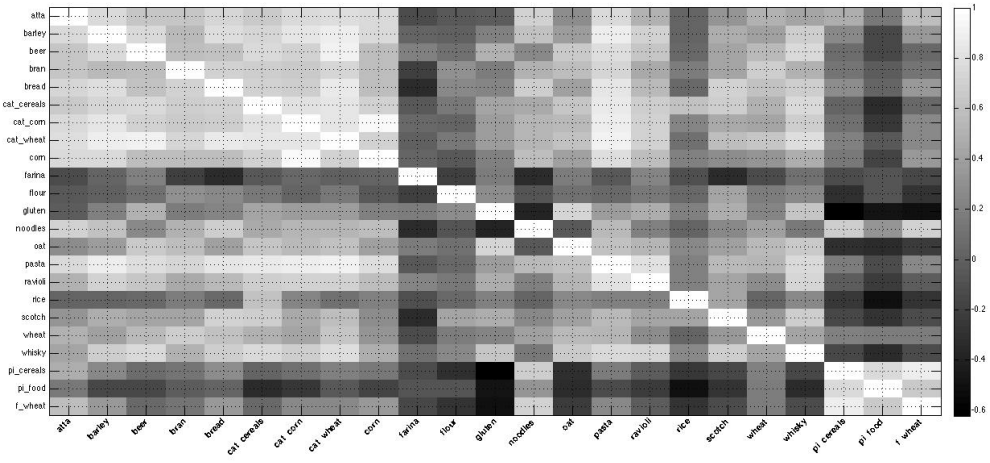


Figure 4.4: Heatplot Cereals: Volume of Tweets per Keyword and per Category

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.

4.3.2 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 5. According to [9] such models are better suited to utilise social media for predictions.

A smilier correlation analysis has been made by the UN [15]. 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 [15]. 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

4.4 Conversation Drivers

Following our correlation 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. More specifically we want to know if the conversations can be related to market fundamentals that cause soaring food prices. Following the two recent food crises in 2007 and 2010 a lot of research has been centered around defining causes of volatile food prices. In [32] they define a taxonomy for drivers of international food price spikes and differentiate 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. This taxonomy will serve us as a baseline in annotating our events.

4.4.1 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 events. 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. [21] defined three categories of temporal behaviours. “Continuous activity, periodic activity or activity concentrated around an isolated peak”. Continues activities are topics that are of daily interest such as weather. On the other hand periodic actives 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 [3]. 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 continuoes interest.

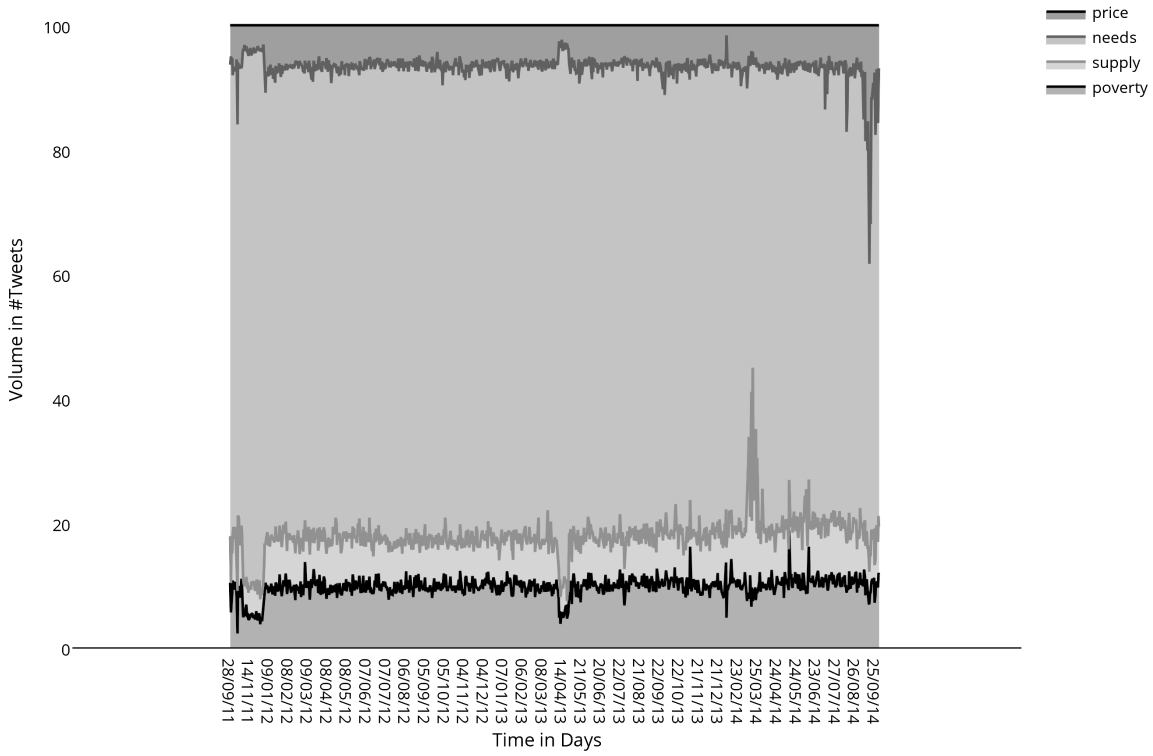


Figure 4.5: Topic Distribution - Food Security

4.4.2 Methodology

As described in the previous section the prominent peaks could not be attributed to any discussion around indicators that are relevant to food security. In this subsection we investigate in detail what topics cause the attention peaks and whether they can be attributed to market fundamentals or topics concerning food security. For this analysis we consider discussion around food supply, price, needs and poverty. We investigate 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 anomalies in our food topics we applied a similar approach as in [3] [21]. 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

[3] http://www.kpopstarz.com/articles/112632/20140916/t-ara-sugar-free.htm

and calculated the mean of the Twitter volume. From those values we calculated the Median absolute deviation (MAD) as in Equation 4.1:

$$\overline{MAD} = median_i(|X_i - median_j(X_j)|) \tag{4.1}$$

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 events. 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 used a similar approach as 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 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.

4.4.3 Event Annotation

We annotate each peak according to the definitions given bellow. Our classification mostly mimics the main dimensions of food security but also includes categories from the taxonomy of Tadesse et al[32]. There is a strong overlap between the two taxonomies where the later naturally 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 [17] 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 [27] 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 [32].

Stability
Poverty, political unrest and topics concerning extreme weather [27] 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.

4.4.4 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 then 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 [4].

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 [5]. As a consequence there was also a sharp increase in discussion around food banks. The UK observed a 51 % increase in food bank users [6].

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 [7]

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. [8]

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 [9]. 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.

[4] http://www.theguardian.com/science/2012/apr/25/mad-cow-disease-us-mutation
[5] http://www.cnbc.com/id/101588110
[6] http://www.bbc.com/news/business-27032642
[7] http://www.theguardian.com/society/2013/dec/17/government-under-fire-eu-funding-food-banks
[8] http://www.usatoday.com/story/news/world/2013/07/17/india-children-deaths/2523727/
[9] http://www.nytimes.com/2014/09/09/sports/football/ray-rice-video-shows-punch-and-raises-new-questions-for-nfl.html

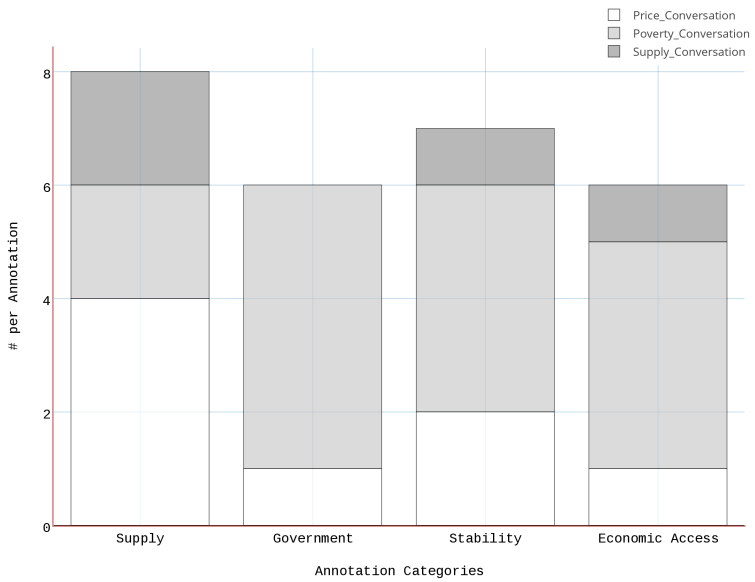


Figure 4.6: Annotation Distribution - Relevant to Food Security

4.4.5 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.

Chapter 5

Model Building

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 would like to lay out some basic concepts that will aid the reader in understanding the dynamics of the predictions and ultimately help to comprehend the results. We do not intend to give a thorough introduction as this is out to the scope of this dissertation. We further highlight details towards the data preprocessing, how we train our model, outline the methodology of our time series prediction and lastly discuss how we tackle the problem of high-dimensional data through clustering and feature selection. The later part of this chapter focuses on motivating the choice of features and lays out a performance comparison of a price data model, a social media model and lastly a mixture model.

5.1 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 [38] 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" [16]. In some areas such as time series prediction techniques such as ARMA and AR, have shown clear limitations [6]. Nonlinear approaches, such as Adaptive Neuro Fuzzy Inference System (ANFIS), have proven to be more successful [8]. Prediction accuracy is however not the only concern in forecasting models. Understanding the behaviour and gaining insight into the underlying dynamics is equally important [34]. This makes ANIFS especially appealing. Not only does it poses 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.

5.2 Fuzzy Logic

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 interference system (FIS).

5.2.1 Fuzzy Variables and Fuzzy Sets

Zadeh [38] 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. 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 5.1, as they are differentiable and desirable for optimisation purposes [36]. We will refer to a_i, b_i, c_i as the parameter set.

$$\mu_A(x) = \frac{1}{1 + |\frac{x - c_i}{a_i}|^{2b_i}} \tag{5.1}$$

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.

To illustrate the two concepts Fuzzy logic and Fuzzy sets let us consider Figure 5.1. 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 classify 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 map those two input variables to x_i, where x_i is our output value and a precise numerical definition of the colour object orange. 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.

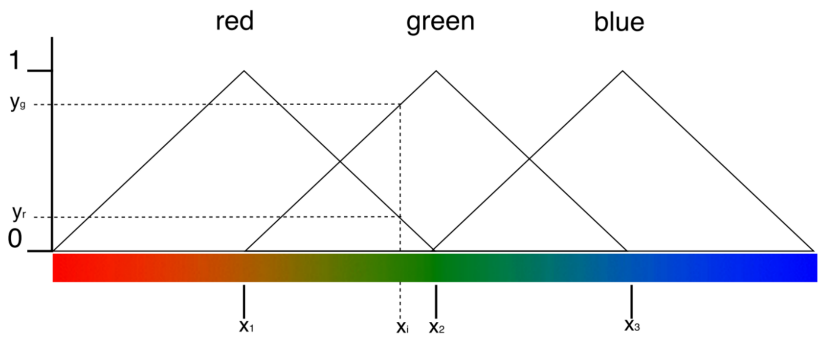


Figure 5.1: Fuzzy Variables and Fuzzy Sets - A Colour Object Example

5.2.2 Fuzzy Interference System

We first describe the components of the Fuzzy interference 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:

$$\textbf{IF} < A > \textbf{AND} < B > \textbf{THEN} < Conclusion >$$

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 5.2 as the *multiplication* of the membership function of set A $\mu_A(x)$ and B $\mu_B(x)$. The result is a weight w_i.

$$w_i = \mu_{A \cup B(x,y)} = \mu_A(x) \times \mu_B(y) \tag{5.2}$$

The intersection is defined in Equation 5.3 as the *probabilistic OR* of the membership function of set A $\mu_A(x)$ and B $\mu_B(x)$.

$$w_i = \mu_{A \cap B(x,y)} = \mu_A(x) + \mu_B(y) - \mu_A(x) \times \mu_B(y) \tag{5.3}$$

The output of one Fuzzy Rule is computed by Equation 5.4.

$$z = dx + ey + f \tag{5.4}$$

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 5.5.

$$Output = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i} \tag{5.5}$$

, where N is the number of rules.

5.2.3 Adaptive Neuro Fuzzy Inference System

We now model the above-described process as a Neural network illustrated in Figure **??**. In **Layer 1** every node maps the input variables x1 and x2 via the membership function in Equation 5.1 to a Fuzzy Set. **Layer 2** applies Equation 5.2 or Equation 5.3 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 5.6 in this Layer.

$$\overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1, 2 \tag{5.6}$$

Layer 4 applies Equation 5.7, which is similar to Equation 5.4 but multiplied by the factor obtained in **Layer 3** from Equation 5.6.

$$\overline{w_i} f_i = \overline{w_i}(dx + ey + f) \tag{5.7}$$

Finally, **Layer 5** computes the overall output as the summation of all incoming signals through Equation 5.5. The construction yields a network with 5 layers, 16 nodes and 24 parameters (12 in Layer 1, 12 in Layer 4).

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.

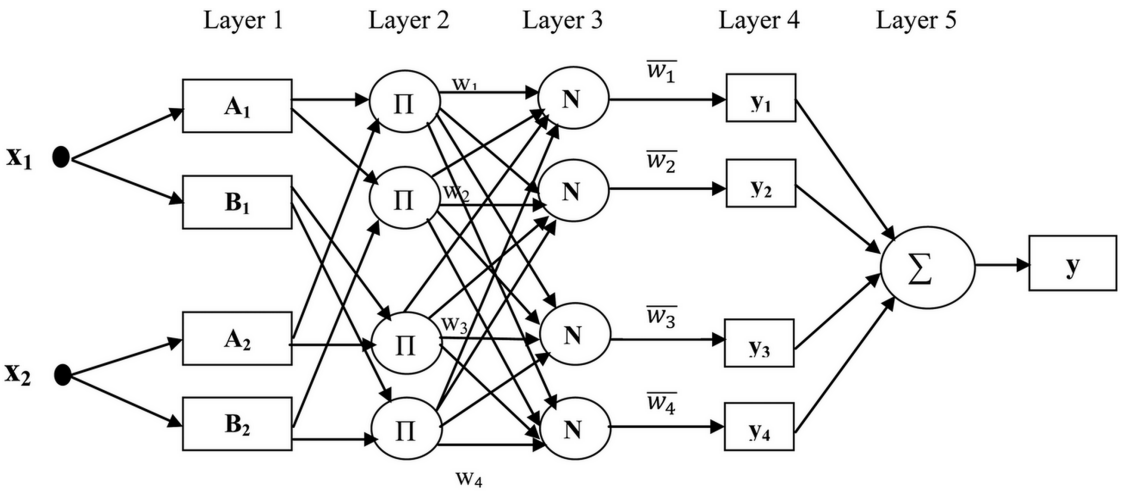


Figure 5.2: Real caption[1]

5.3 Data Preprocessing

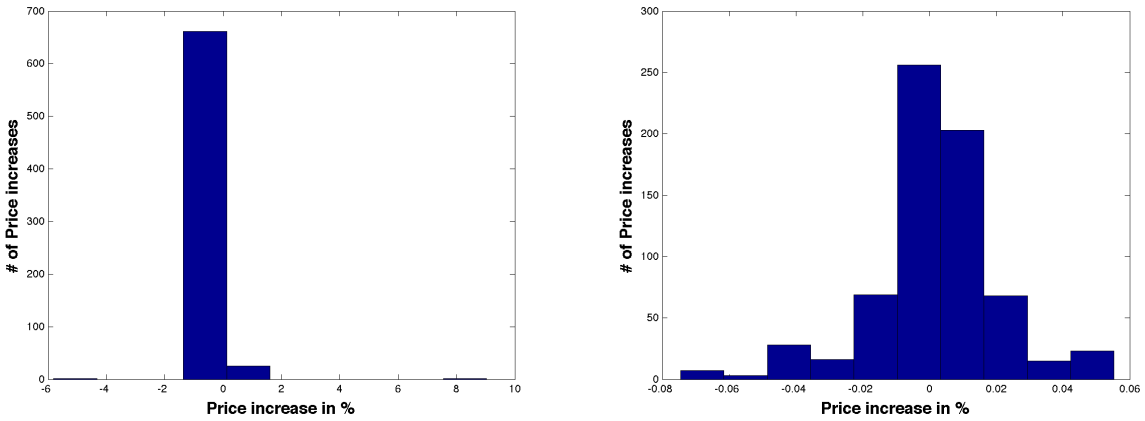
Before feeding the data to the Fuzzy Interfere System, we analysed and cleaned the data. The representation of the data is of great importance to assist ANFIS in learning the relevant patterns. In a first instance, we had to match both the time series of the Twitter data with the time series of the price data. During weekends and national holidays the markets are closed, hence we had to remove such instances from the Twitter data. Given the sparsity of the datasets available for commodities we were forced to hand selected quotes from different markets. We observed that some of them 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 [10]. 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 4.1. Min and Max are the lower and upper bounds of data set

[1] Image credit to: http://pubs.rsc.org/en/content/articlehtml/2014/ra/c4ra02392g

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.

$$\overline{y} = \frac{x - min}{max - min} + \alpha \tag{5.8}$$

Lastly, we performed a scaling of the data which is in agreement with our objective, namely to be more sensitive to long-term than to short-term fluctuations. As we can see in Figure **??** there are some extreme price increases and drops. By applying the Hodrick-Prescott decomposition [13] filter, we receive a distribution that is more normal by avoiding such outliers. Additionally the Figure 5.3 illustrates nicely the characteristics of commodities, namely that it is mostly driven by small price changes.



(a) Price Increase Distribution without Scailing

(b) Price Increase Distribution with Scailing

Figure 5.3: Volume of Tweets per Keyword and per Category

5.4 Training the Model

For training ant 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 [18]. 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. [37] 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. We decided to apply a variant of the sliding window approach and train or 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 Methodology for Forecasting with Fuzzy Logic

The goal of our research is to predict the price of a commodity in the future. The prediction model takes different features $y_t - y_{t-M+1}$ and an input. The problem of predicting the future value y_{t_1} can be formulated as:

$$y_{t+1} = fp(y_t, y_{t-1}, y_{t-M+1}) \tag{5.9}$$

where M is the number of features and fp is our fuzzy prediction model. 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.

Translating the Equation 5.9 into the fuzzy system a prediction would take the following form:

$$\textbf{IF} < y_t \in High > \textbf{AND} < y_{t-1} \in Medium > \textbf{THEN} < y_{t+1} \in Increase >$$

Where High, Medium and Increase are Fuzzy Sets. We measure the difference between the actual value and the predicted value by computing the Root Mean Square Error (RMSE) defined in Equation 5.10. It is an aggregation of all prediction errors for different time stamps. As RMSE is scaled dependent 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.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - \overline{y_i})^2} \tag{5.10}$$

For all experiments, we use the parameters in Table 5.1 unless otherwise noted.

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 5.1: Parameter Settings

5.6 Adapting to High Dimensional Data

ANFIS does not work well for high dimensional data as it generates rules by enumerating all possible combinations of membership functions. 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 5.11 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.

$$J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m ||x_i - v_j||^2 \tag{5.11}$$

μ_{ij} corresponds to the degree of membership of data point i to cluster centre j. This is captured by Equation 5.12. 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.

$$\mu_{ij} = \frac{\frac{1}{d_{ij}}^{\frac{1}{m-1}}}{\sum_{k=1}^{c} (\frac{1}{d_{ik}})^{\frac{1}{m-1}}} \tag{5.12}$$

The fuzzy center on the other hand is computed by Equation 5.13 $\forall_j = 1, 2....c$.

$$v_j = (\sum_{j=1}^{n} (\mu_{ij})^m x_i) / (\sum_{j=1}^{n} (\mu_{ij})^m) \tag{5.13}$$

Other then reducing the membership functions we can directly reduce the number of features used. For that purpose, we used a feature selection algorithm Relief that allows us to distinguish relevant from irrelevant features were relevance is measured on how much influence a feature has on the output value. Relief [19] [20] [28] find 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 5.14.

$$W_i = W_i - (x_i - nearestHit_i)^2 + (x_i - nearestMiss_i)^2 \tag{5.14}$$

A feature is considered irrelevant or redundant if $W_i <= 0$, that is if the difference between the same class is higher than the one of a different class.

5.7 Benchmark Model

5.7.1 Input Model

Yao et al. [ref] suggests that the market can be categorised using the major trends, the intermediate trends and the minor trends. Here 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. 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[**?**]. The bellow Table 5.2 illustrates the values assigned to each feature. x is the value of today and $x-1$ is a value one day in the past. $\mu(]x-y;x])$ 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	$\mu(]x-7;x])$
32	W2	$\mu(]x-14;x])$
33	M1	$\mu(]x-30;x])$

Table 5.2: Input Model: Benchmark Prediction

5.7.2 Feature Selection

The objective here is to find the most significant features among the 33 in our input model to decrease the complexity and lastly improve the prediction accuracy. To determine the most predictive features, we use the Relief algorithm described in Section 5.6. 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). Initially, we were hoping to cover all five food categories. Unfortunately, we were unable to get access to daily commodity prices of oil and sugar. We further took care to capture different price dynamics. The prices of milk and wheat move in all directions. Beef on the other hand is characterised by a long-lasting trend. The configuration *w-4* in the Table 5.3 refers to wheat with a prediction horizon of four days. The most 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 features [2].Those additional features ($\varphi_7 - phi_10$ 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

[2]We 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.

and outperform strategies based on features in the recent past. Robert Novy - Marx also observed this effect and investigated the hypothesis in [12]. 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

5.7.3 Results

The prediction results across different time horizons are highlighted in Figure 5.4. These results are obtained using the model described in Section 5.5 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. If the prediction task is to predict y_{t+3} and for illustration purposes we assume today is Monday, then the prediction goal translates to approximating the true value on the following Thursday. Assuming the feature selection process yields y_{t-1} and y_{t-1} this corresponds to using the price on the previous Sunday and Saturday. Finally our model with two membership functions High and Low gives us the following Fuzzy Model.

$$\textbf{IF} < Sunday \in High > \textbf{AND} < Saturday \in Low > \textbf{THEN} < Thursday \in 0.67 >$$

The output is a numerical value and not a class.

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 5.5 5.6 5.7. For all three commodities, the predictions within one week are extremely 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 5.7.2 our input variables from the intermediate past perform exceptionally well 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 captures the underlining trend. The overestimation seen in Figure 5.5d and Figure 5.7d 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. Better predictions can be expected by considering different training samples.

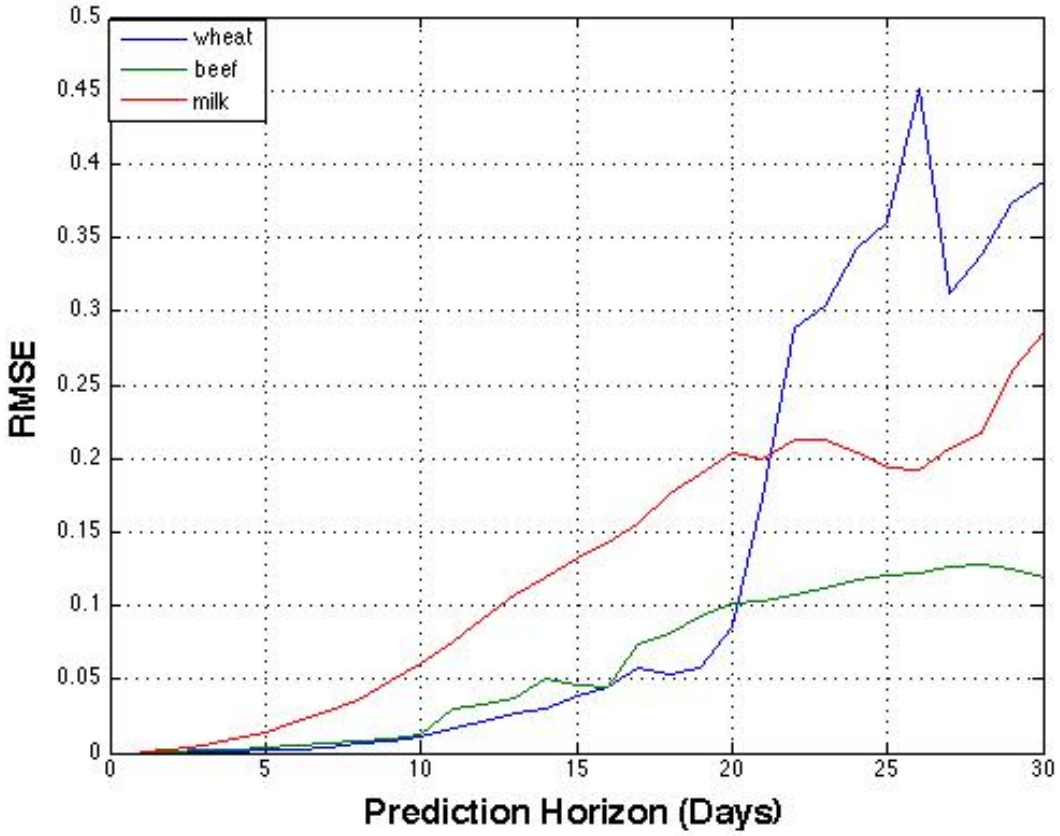
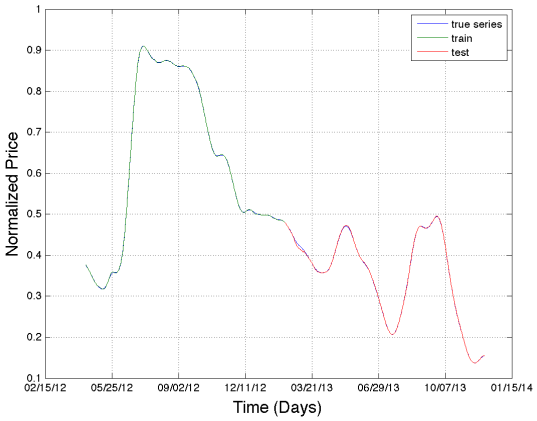
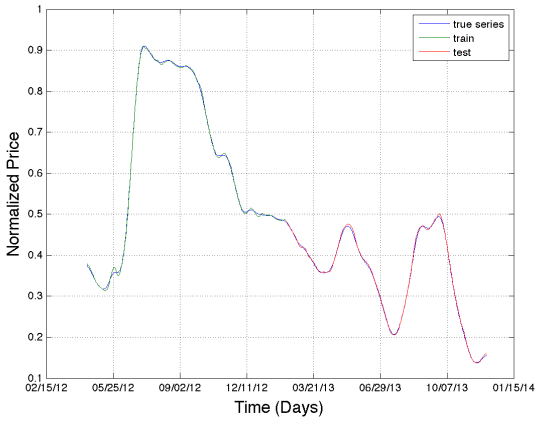


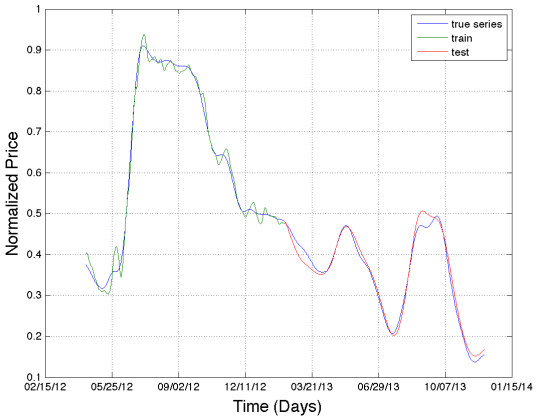
Figure 5.4: Prediction Accuracy - Benchmark



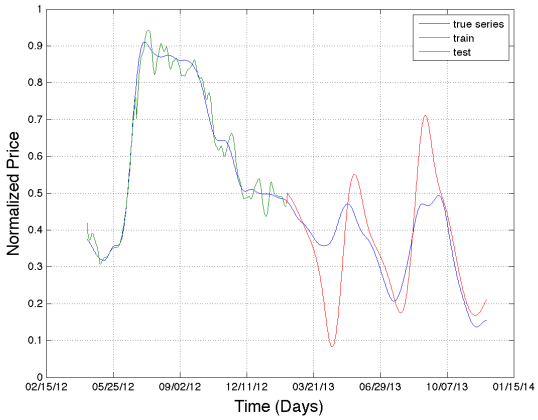
(a) 4 Day Horizon - RMSE 0.0015948



(b) 7 Day Horizon - RMSE 0.0030423

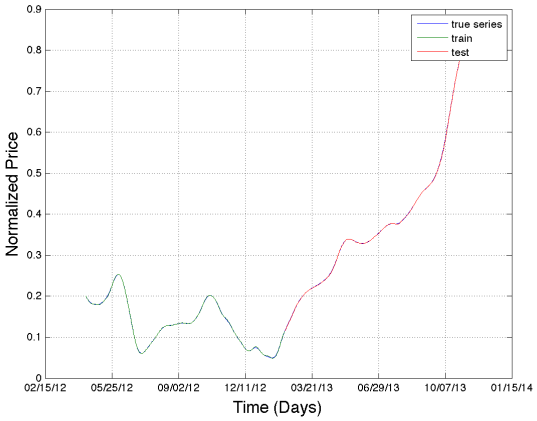


(c) 14 Day Horizon - RMSE 0.015192

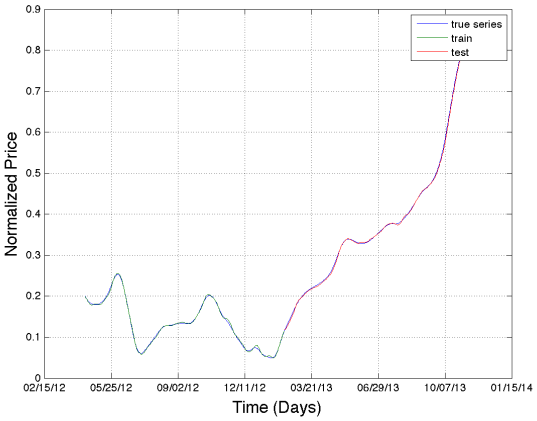


(d) 21 Day Horizon - RMSE 0.105

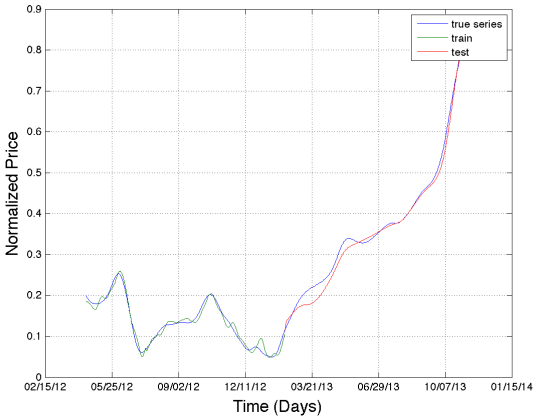
Figure 5.5: Benchmark Prediction Wheat



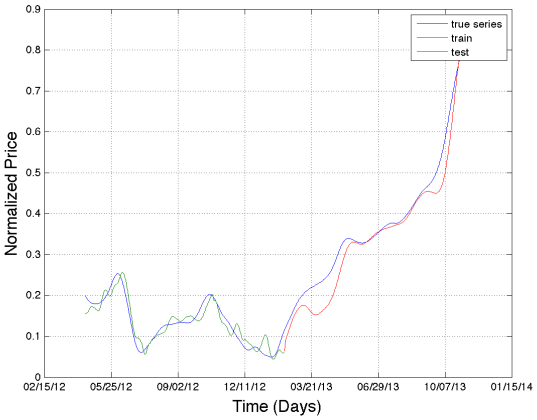
(a) 4 Day Horizon - RMSE 0.00086578



(b) 7 Day Horizon - RMSE 0.0059255

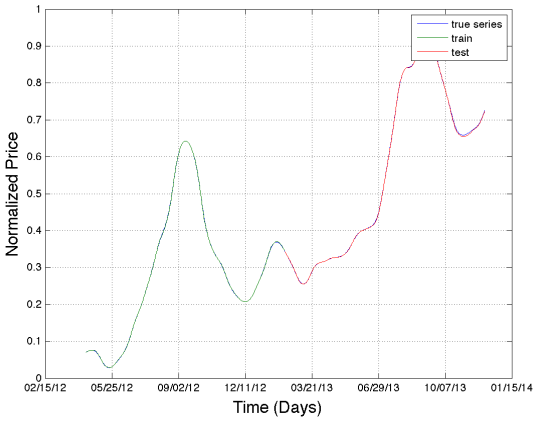


(c) 14 Day Horizon - RMSE 0.017041

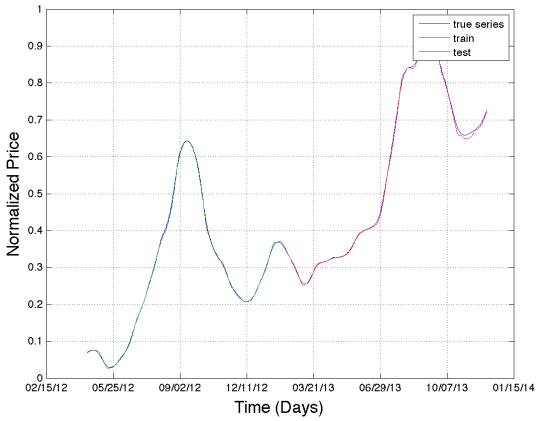


(d) 21 Day Horizon - RMSE 0.040397

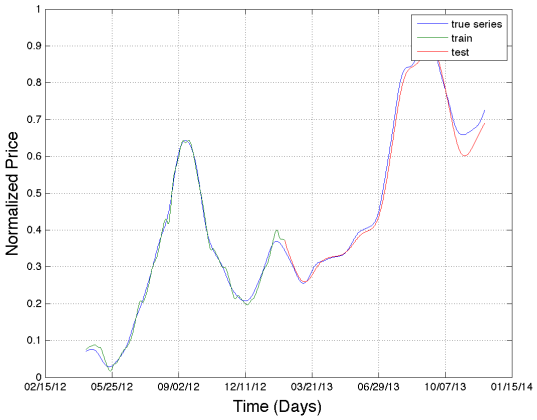
Figure 5.6: Benchmark Prediction Beef



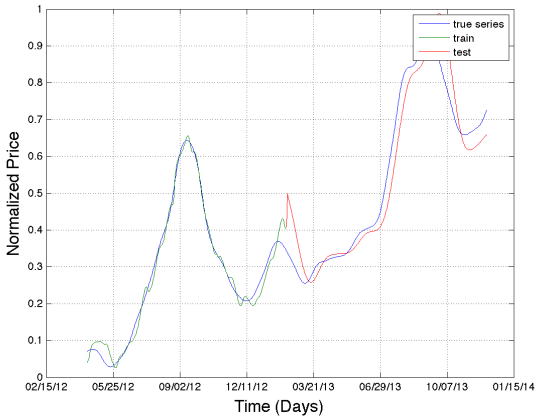
(a) 4 Day Horizon - RMSE 0.0015964



(b) 7 Day Horizon - RMSE 0.0049564



(c) 14 Day Horizon - RMSE 0.1185



(d) 21 Day Horizon - RMSE 0.023702

Figure 5.7: Benchmark Prediction Milk

5.8 Social Media Model

5.8.1 Input Model

Xue Zhanga et al. [?] found that Twitter conversation correlates and is even predictive of financial market movements. They measured the attention of a given subject by aggregating the daily volume. In Section 4.3.1 we investigated if such a correlation is present and concluded that for most commodities no such linear relationship exists. Nonetheless we include the social attention as a possible feature as it might prove to be useful in the form of a lagged variable (i.e. that the volume three days ago correlates to the commodity price of today). 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). 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 4.4 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 [26] 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 [25], the proportion of neutral and total Tweets, the ratio between the numbers of non-neutral and total posts [39], the ratio between the number of positive and negative discussions and lastly the ratio between the numbers of neutral and non-neutral messages.

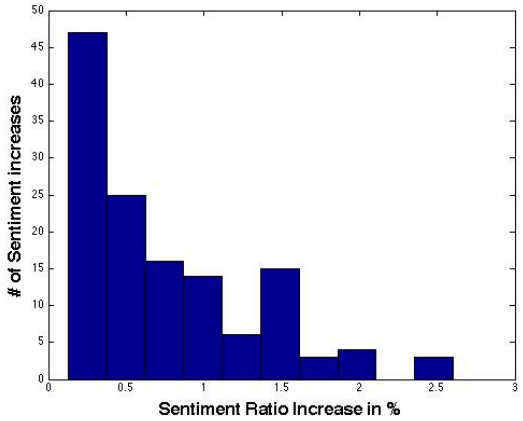
Table 5.4 concludes our input model. We measure the sentiment for both Twitter buzz and contextually sensitive Tweets. The result is an input model with 51 features.

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

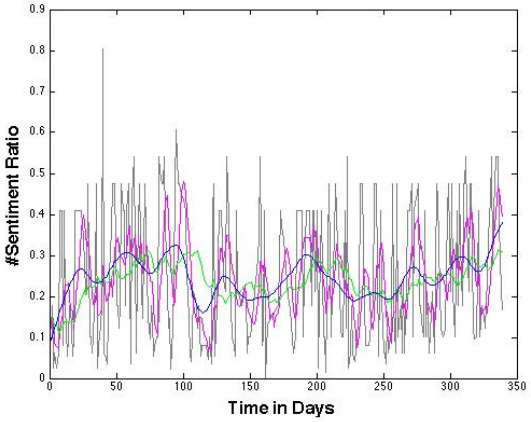
Table 5.4: Input Model: Social Media Prediction

5.8.2 Strengthening Social Media Features

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 5.8a. 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 5.8b. 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 5.5. 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) Smoothing functions: Past Week (Magenta), Past Month (Green), HPD (Blue)

Figure 5.8: Volume of Tweets per Keyword and per Category

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 5.5: Sentiment Correlation

We further investigated to what extent Twitter lags as an indicator. If a casual 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 [30]. We choose a window of three days because none of the attention spikes showed to last longer than three days (Section

4.4). 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.

5.8.3 Feature Selection

For the purpose of this analysis, we repeat the exercise performed in 5.7.2. Again we try to find features that generalise well among different commodities and horizons. Unlike the price data, the 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. Among the three were **1.)** the attention count of the selected products that showed a high correlation in our analysis (AC50), **2.)** the first sentiment ratio of poverty (SR32) and **3.)** the first sentiment of the subcategory (SR15). Other features that generalised well for all prediction tasks of a given commodity were 1.) the second sentiment of the subcategory (SR16), 2.) the 2nd sentiment of poverty (SR33), the fourth sentiment of poverty (SR35), the 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 for each given commodity are illustrate in Table 5.6.

Sentiment ratios are clearly the favoured features where only the highly selected volume counts seem to have a considerable effect. The sentiment of our contextually sensitive Tweets appears to be a strong indicator of the price variance, especially the sentiment for poverty that we capture in the form of three different ratios. In 4.4.4 we showed the peaks had the highest relevance for the discussions centred around poverty. Admittedly, it is not a decisive indicator of the quality of the feature. However, it might explain why poverty features rank so highly. Similarly the three food security indicators, which have been selected by Relief all showed a degree of relevance in our analysis. Needs had no relevance and hence it does not come as a surprise that it was not selected as a potential feature. It appears that Relief successfully removes redundant features. We would, for example, expect needs and poverty to be similar likewise supply and price. None of them overlaps in the ratios used. Lastly, the sentiment ratios 1 and 2 seem to be more indicative than others.

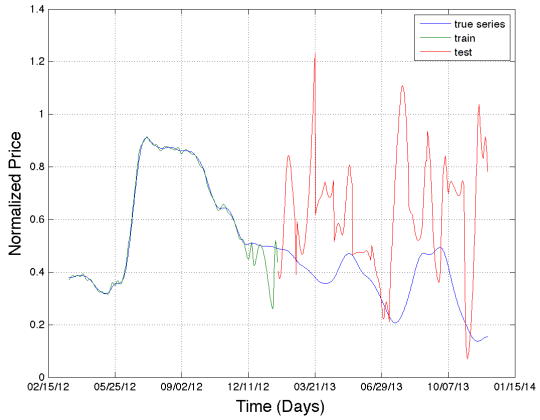
Feature	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
wheat	AC50	SR32	AC51	SR35	SR16	SR15
beef	AC50	SR45	SR32	SR33	SR38	SR44
milk	AC50	AC51	SR45	SR38	SR35	SR49

Table 5.6: Feature Selection: Benchmark Prediction

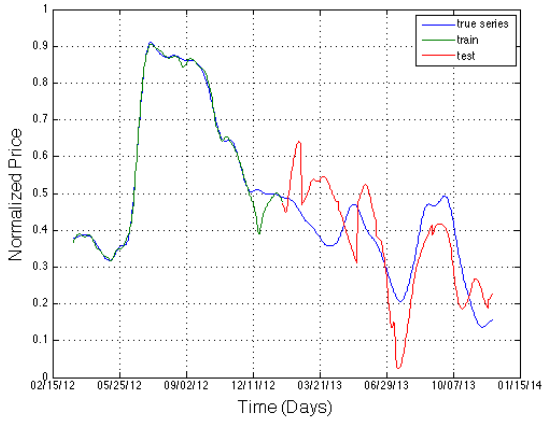
5.8.4 Results

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 5.10 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 2nd 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 specifi-

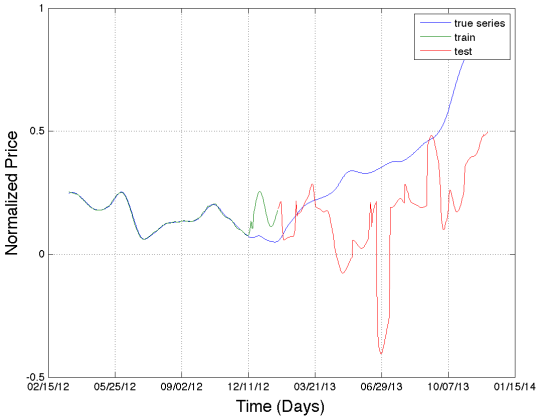
cally we found that certain features showed a negative correlation in the train set, however, a positive relationship in the test set. We experimented with various sizes of train sets but none of them notably improved the prediction results. Lastly, we were motivated to examine the content of the Tweets. The set of wheat Tweets might be more relevant to our given topic then the set of milk Tweets. For the purpose of this analysis, we computed the top 50 keywords, excluding stop words, of the Tweets containing at least one term of the subcategory wheat or milk. 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 a higher relevance (i.e. the number of keywords retrieved deemed relevant was higher). For milk the most frequent keywords were ice, butter, milk, cream, chocolate and peanut. At a first glance, butter and cream seem relevant. The bigrams showed that ice and cream are related (ice cream) as well as peanut and butter (peanut butter). Both bigrams are considered noise as they only have a very distant relationship with milk. On the other hand for wheat, the most common terms retrieved were beer, gluten, bread, pasta, wheat, barley and noodles. Those terms are clearly highly relevant to the subcategory wheat. On the basis of these finding, we conclude that the list of keywords we used retained a high number of Tweets that were not relevant enough.



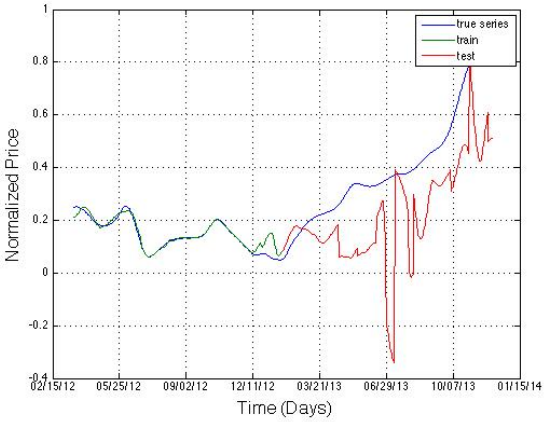
(a) 1 Day Horizon Wheat - RMSE 0.3580



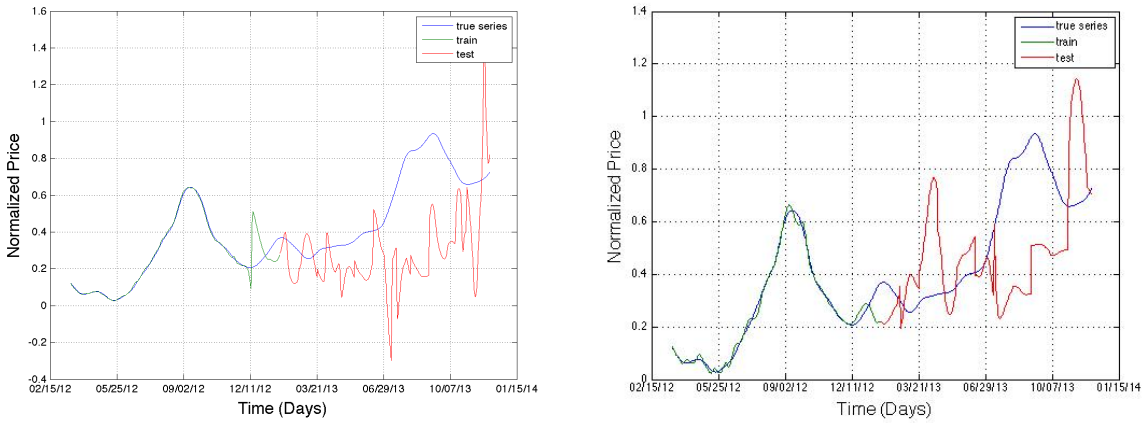
(b) 1 Day Horizon Wheat - RMSE 0.1083



(c) 1 Day Horizon Beef - RMSE 0.3209



(d) 1 Day Horizon Beef - RMSE 0.2383



(a) 1 Day Horizon Milk- RMSE 0.3782

(b) 1 Day Horizon Milk - RMSE 0.2960

Figure 5.10: Social Media Prediction

5.9 Combined Model

5.9.1 Feature Selection

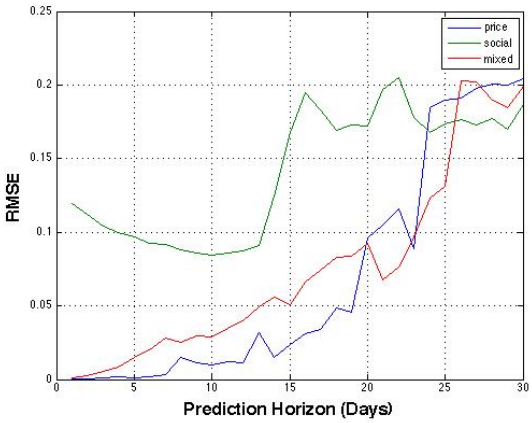
Combining the price and social media features we want to investigate if Relief proposes as similar set as for the separate models or if the combined model favours a different group of input variables. Only a few features generalised well over all commodities which is why we focus our search for finding input variables that are specific to beef, milk and wheat but generalise well over different prediction models. 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).

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

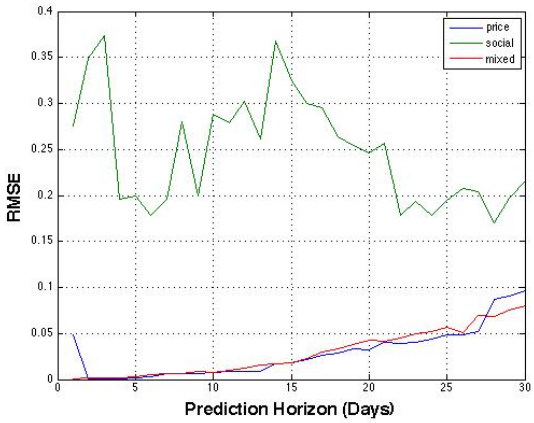
Table 5.7: Feature Selection: Benchmark Prediction

5.9.2 Performance Comparison

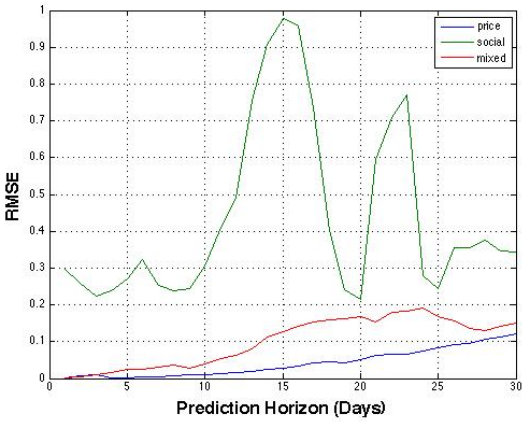
As our mixture model exhibits similar behaviours to the price model, we show the exact predictions for each commodity in the Appendix C. Instead we focus our attention on the performance comparison of the three different models (e.g. price, social and mixture model). Figure 5.11 summarises our results 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



(a) Model Comparison Wheat- RMSE 0.016061



(b) 7 Day Horizon Milk - RMSE 0.031539



(c) 14 Day Horizon Milk - RMSE 0.11321

Figure 5.11: Social Media Prediction

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.

Chapter 6

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. Or 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, irreverent 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.

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 [15] to capture a more diverse set of nations. Particularly Indonesia would provide and 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 5.8.4 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 [24]. 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

Data

A.1 Crowd Flower

For the categorisation of the keywords for our predictor lexicon four crowd tasks were created. This section details the instructions given to the crowd workers for the four categorisation tasks.

A.1.1 Categorise: Food Price

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 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.1.2 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
- ressource

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.1.3 Categorise: Food Poverty

This is a categorisation task centered around food security. Please categorise terms appearing in tweets about food in order 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.1.4 Categorise: Food Needs

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 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

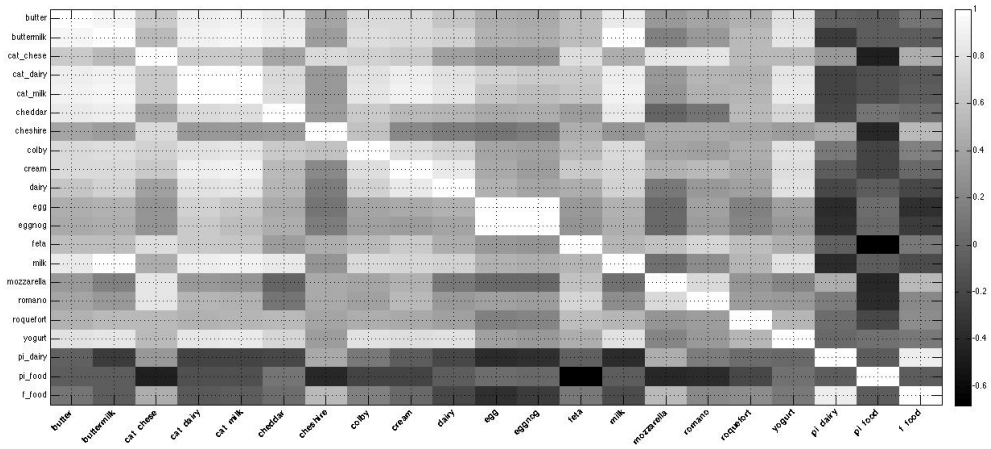


Figure B.1: Heatplot Dairy: Volume of Tweets per Keyword and per Category

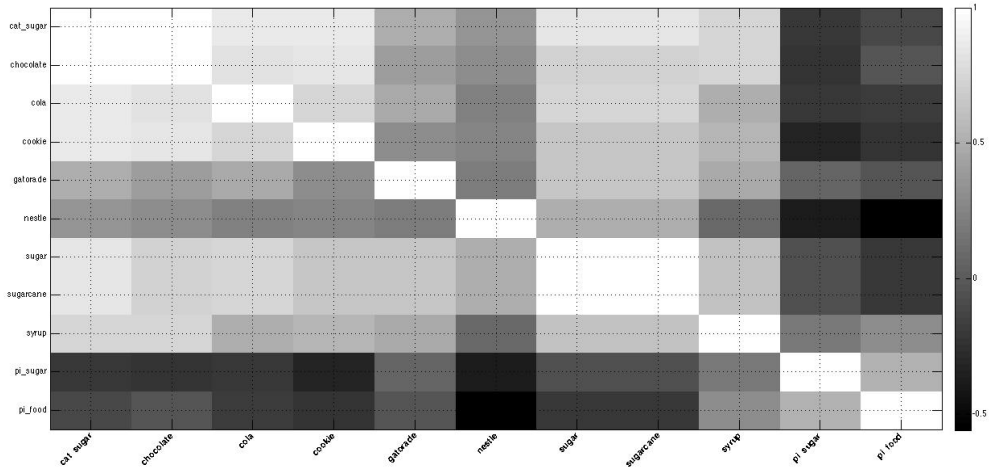


Figure B.2: Heatplot Sugar: Volume of Tweets per Keyword and per Category

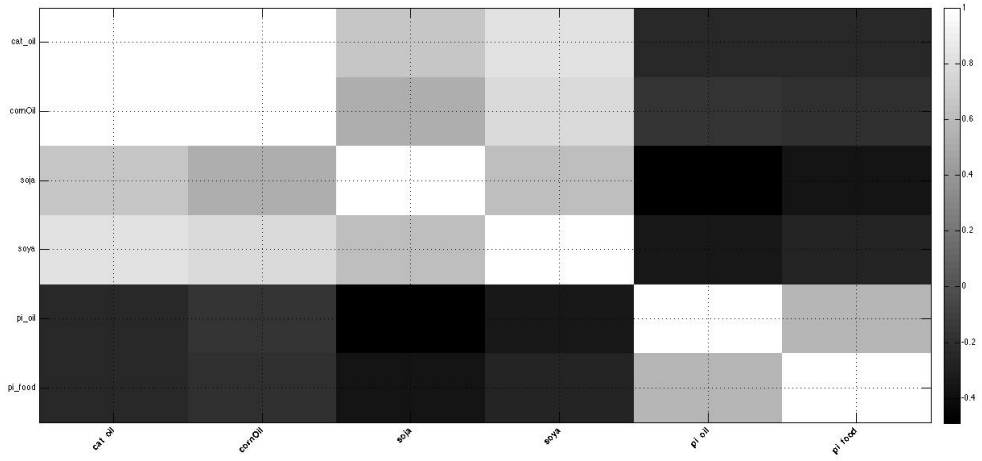
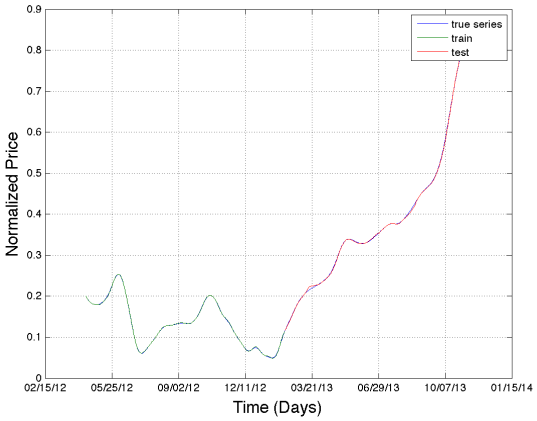


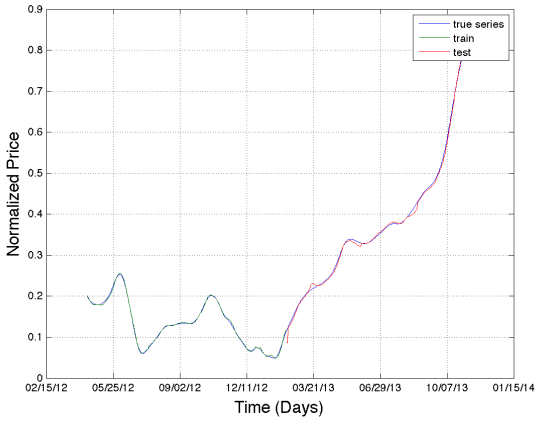
Figure B.3: Heatplot Oil: Volume of Tweets per Keyword and per Category

Appendix C

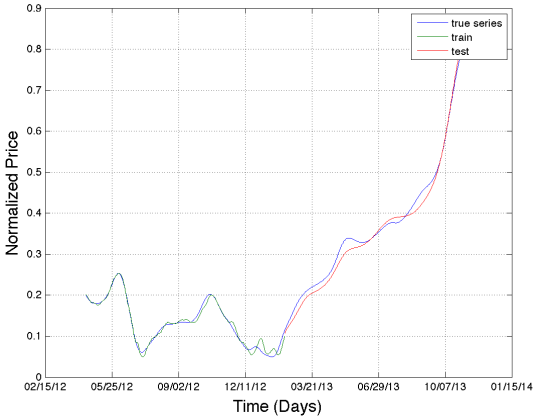
Time Series Modeling



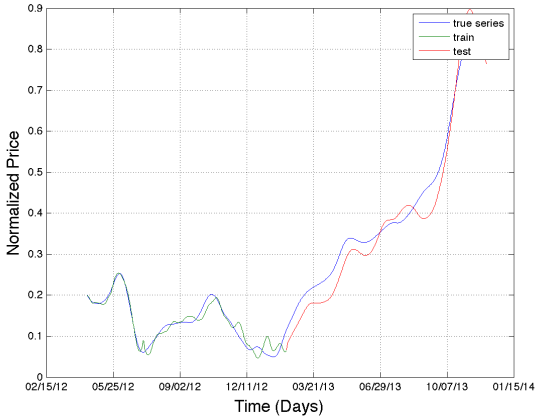
(a) 4 Day Horizon Beef - RMSE 0.0021991



(b) 7 Day Horizon Beef - RMSE 0.0065262

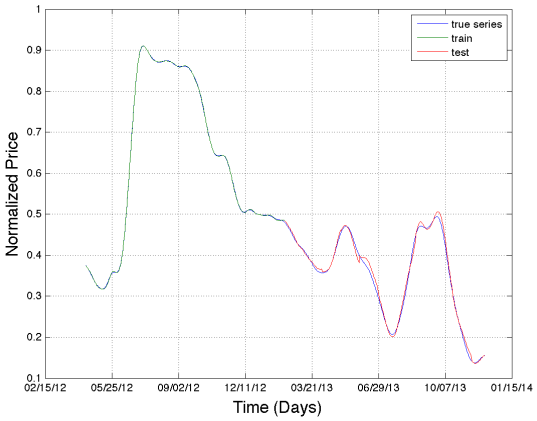


(c) 14 Day Horizon Beef - RMSE 0.016448

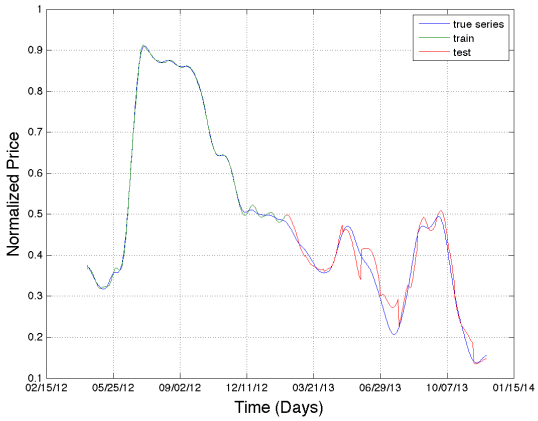


(d) 21 Day Horizon Beef - RMSE 0.042037

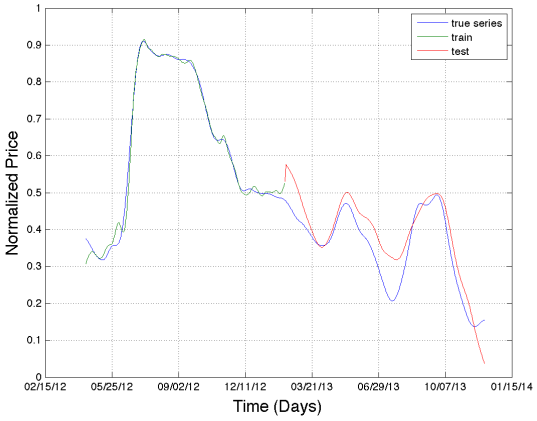
Figure C.2: Social Media Prediction



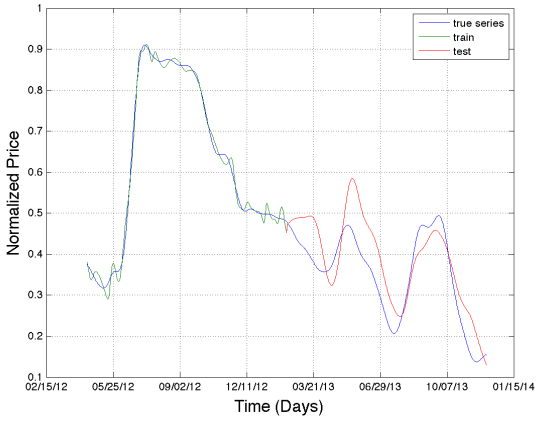
(a) 4 Day Horizon Wheat - RMSE 0.0084



(b) 7 Day Horizon Wheat - RMSE 0.0281

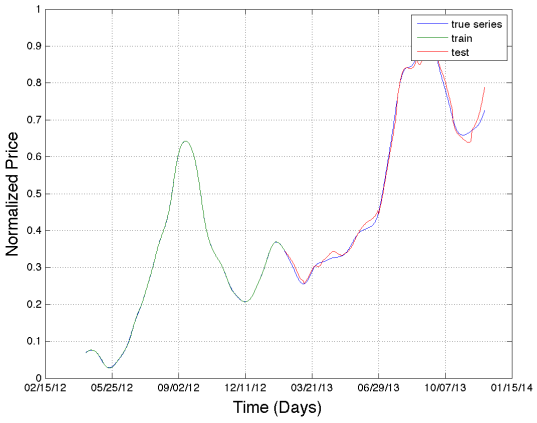


(c) 14 Day Horizon Wheat - RMSE 0.0557

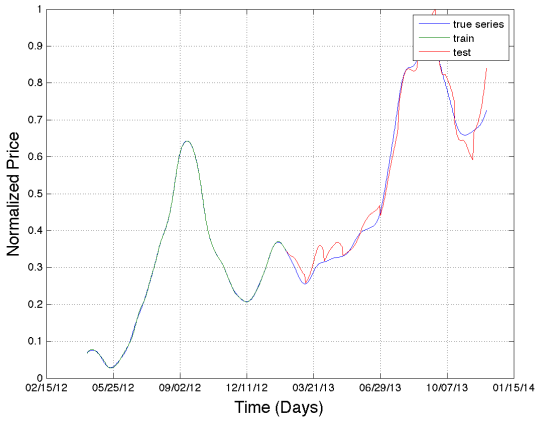


(d) 21 Day Horizon Wheat - RMSE 0.0673

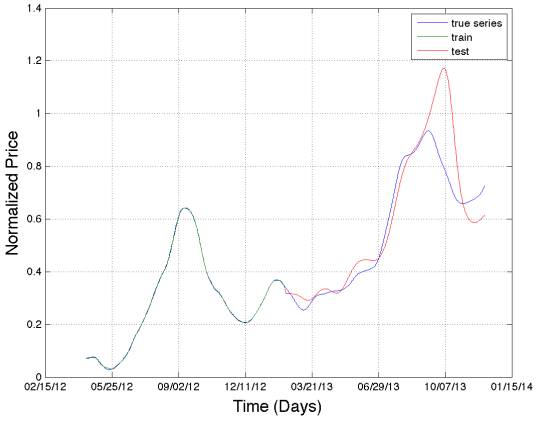
Figure C.1: Social Media Prediction



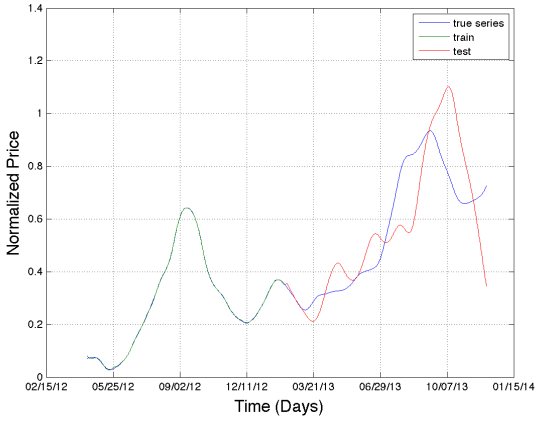
(a) 4 Day Horizon Milk - RMSE 0.016061



(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] 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.

[6] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

[7] Curt Burgess and Kevin Lund. The dynamics of meaning in memory, 1998.

[8] Christopher Chatfield. *The analysis of time series : an introduction*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton (Fl.), 2004.

[9] 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.

[10] John D'Errico. Interpolates and extrapolates nan elements in a 2d array, 2012.

[11] Food and Water watch. Casino of hunger, 2009.

[12] Qiang Gong, Ming Liu, and Qianqiu Liu. Is momentum really momentum? international evidence, 2011.

[13] 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.

[14] C. J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. 2014.

[15] Pulse Lab Jakarta. Mining indonesian tweets to understand food price crises. *Food and Agriculture*, 2013.

[16] 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.

[17] Sarah Macnaughton Julian Parfitt, Mark Barthel. Food waste within food supply chains: quantification and potential for change to 2050.

[18] Iebeling Kaastra and Milton Boyd. Designing a neural network for forecasting financial and economic time series, 1996.

[19] 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.

[20] Igor Kononenko. Estimating attributes: Analysis and extensions of relief. pages 171–182. Springer Verlag, 1994.

[21] 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.

[22] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[23] K. LUND and C. BURGESS. PRODUCING HIGH-DIMENSIONAL SEMANTIC SPACES FROM LEXICAL CO-OCCURRENCE. *Behavior research methods, instruments & computers*, 28(2):203–208, 1996.

[24] Miranda Mowbray. The twittering machine. In *WEBIST (2)'10*, pages 299–304, 2010.

[25] 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.

[26] J. Nofsinger. Social mood and financial economics. In *The Journal of Behavioral Finance*, 144–160, 2005.

[27] EC FAO Food Security Programme. An introduction to the basic concepts of food security, 2008.

[28] Marko Robnik-Sikonja and Igor Kononenko. An adaptation of relief for attribute estimation in regression, 1997.

[29] 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.

[30] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction.

[31] Stanford. Corenlp. 2011.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Jingtao Yao and Chew Lim Tan. A case study on using neural networks to perform technical forecasting of forex, 2000.

[38] L.A. Zadeh. Fuzzy sets. *Information Control*, 8:338–353, 1965.

[39] 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.