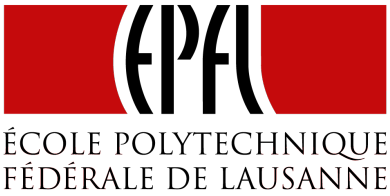


Title of Thessi

Master Thesis

S. Tudent

January 19, 2038

Advisors: Prof. Dr. A. D. Visor, Dr. P. Ostdoc

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

Chapter 2

Social Media Data Acquisition

In this section we describe the filtering process of the tweets and the creation of two 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 5.6 M food relevant tweets.

Firstly, we detail an algorithm Hyperspace Analogue to Language (HAL) [**?**] which was used to find relevant keywords for our two lexicons. We then describe our framework for retrieving food related keywords that form our food lexicon followed by a chapter describing the framework for creating the Predictor Lexicon. In the Chapter Experimental Evaluation we analyse the different metrics influencing the performance of HAL and present the results. Lastly, we describe the filtering algorithm used to create our set of food relevant tweets.

2.1 Hyperspace Analogue to Language

HAL creates a semantic space from word co-occurrences [**?**]. 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 an $N \times N$ square matrix of co-occurrences. Every time a specific word appears within 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 [**?**] gives an example of a simple sentence *"The horse raced past the barn fell."* in Table **??** with a sliding window of five. Let's consider the first row. *"The"* precedes *"Barn"* twice. Once within a distance of five and the other time it directly

[1] https://archive.org/details/archiveteam-json-twitterstream

precedes the word *"Barn"*. Hence, that cell receives a score of five for the proximate one and a score of one for the word further away resulting in a final score of six.

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

	Barn	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 2.1: Toy example of HAL

2.1.1 Motivating a Semantic Approach

HAL gives us a way to study the relationship between words. More specifically we aim to understand what words are represented in the context of *Food* and topics centered around *Food Security*. To achieve this we need a methodology for representing the meaning of a word. The reason that we analyse the context of a word is to identify new words that have a similar meaning or given the same context express the same thing. The later is concerned with identifying synonyms where as the former looks at contextual similarity. For example, let's look at the word *mold* and *available*. Those two words seem unrelate, but given the context of food they express the same thing. Namely an abundance of food. Through the role of the context they posses elements of items similarity but by themselves they would never be considered words with similar meaning. It's important to stress that they are not similar because they occur frequently locally, but because they occur frequently in similar sentential context. Burgess et al. [**?**] argues 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. This observation motivated us to deviate from the commonly used co-occurrence analysis an take a step further to improve the precision of our filtering framework.

2.2 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 keyword as $K_{initial}$. Words included are the most common traded food commodities as listed by IMF [2] along the ten most important staple foods that feed the world [3].

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

[2] http://www.imf.org/external/np/res/commod/index.aspx
[3] http://knowledge.allianz.com/demography/health/?767/the-worlds-staple-foods

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 [4].

FAO tries to measure the overall food fluctuation by five different food categories namely *meat, dairy products, cereals, vegetable oil* and *sugar*. The weighted average of those five categories as illustrated in [**?**] defines the international food price index which is an overall measure of the current food condition. We additionally created a further 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 frequently 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 **??**.

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. The only fluctuating and unknown factor 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 much more likely to capture the social attention then raw items due to their every day use.

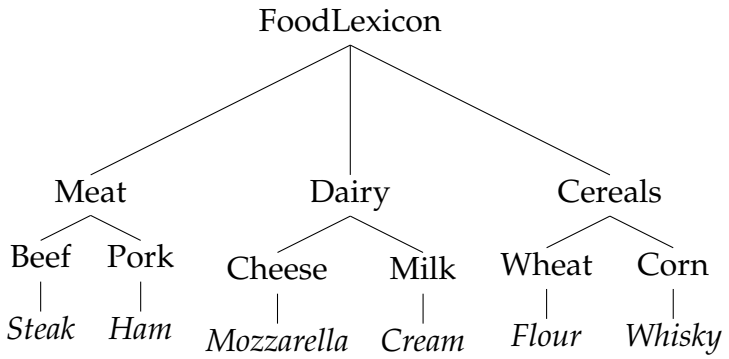


Figure 2.1: Food Lexicon - Hierarchie

Other than the sparsity of the data we further faced the problem of ambiguous key-

[4] http://www.fao.org/worldfoodsituation/foodpricesindex/en/

words. *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.

2.2.1 Candidate Food Term Selection

We initially assume an empty sett 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 keyword $k_i \in K_{initial}$ we evaluate if $k_i \in K_{all}$. If *True* we consider $k_i \in K_{final}$. For all keyword $k_i \notin K_{all}$ the condition of it being frequent is evaluated and if *True* added to the category *Other 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 [**?**] 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 summarise our framework as follows:

1.) We add all keywords $k \in K_{initial}$ to K_{final} only if $k \in K_{all}$ or k is frequent

2.) Further we add all p_i to K_{final} only if $p_i \in K_{all}$

3.) We create a HAL space using a random subsample of 10% from our initial collection with all keywords that occur > 100. $\forall c_i \in K_{final}$ we pick the keyword $k \in K_{final}$ that most frequently occurs over the entire sample and retrieve the top 500 similar terms. We 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_f forms our Food Lexicon.

2.3 Predictor Lexicon

From our basic food lexicon we proceeded to extract features that we can use to explain events around Food Security. The 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 the 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 [**?**].

To model food security we focus our work on those four dimension namely *Access, Availability, Utilisation* and *Stability*. Together those predictor categories build the set C_p. Attempts have been made to capture Availability by the UN [**?**].

Lexicon / Subset s	Keywords (i: from initial set, e: from [**?**] , 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 2.2: A Summary of the Evolution of our Food Lexicon

We define the predictor category *Access* by looking for tweets containing price as a keyword as in [**?**] but improve the recall by including synonyms of *price* that appear in the same context. *Availability* was defined in similar fashion by matching keywords that appear in the context of food availability as in [**?**], however a different set of keywords was selected as described in the following chapters. Unlike [**?**] we don't measure food Utilisation by observing the exact diet but capture the people's food needs. Lastly as a measure of *Stability* we focused our attention on economic stability. Keywords in the context of poverty were selected to match this predictor category similar to [**?**] [**?**].

2.3.1 Candidate Predictor Term Selection

Since HAL has not been extensively used in previous work for term selection we drafted two different frameworks which we evaluated. As a reminder K_f 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 [5].

Framework 1

1.) $\forall k \in K_f$ choose the keyword k with the highest occurrence form the entire sample. 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 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. The first step will give us the most prominent food term. This is most likely going to be something general such as the keyword *"Food"*. Step 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 step 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 will proceed with Framework 2 for three reasons. Firstly Framework 1 did not retrieve us the desired keywords for all categories. Secondly, between the results of Framework 1 and 2 there was a substantial overlap and lastly Framework 2 is more efficient to execute. This is particularly important since creating the HAL space is computationally very expensive. The final lexicon was further enriched by including synonyms from thesaurus [6] for supply, need, poverty, and price. The terms of the final predictor lexicon are presented in Table **??** and for future reference we will refer to it as K_p

2.3.2 Annotation and False Positive Removal of HAL Results

The workers were presented with four different tasks, one for each category. For every task we asked the workers to classify the term as A. Relevant, B. Likely, C. Unlikely and D. Not in English. Since Overlaps may occur, particularly for the category price

[5] http://www.crowdflower.com/
[6] http://www.thesaurus.com/

and supply as well as poverty and need we asked the workers to classify them as 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 3 independent annotations for every word and applied a majority 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.

2.4 Experimental Evaluation

In order 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 respectively. Our corpus of food related tweets has a number of appealing properties as it covers a large vocabulary centered around food. Unlike most corpora that represent formal business reports or specialised dictionaries our food corpus represents everyday speech. This gives us a closer approximation on how people would talk in the context of our predictor categories.

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 then 4 words which could impact the quality of the results. Hence, on the 40 % sample we further excluded tweets that contain less or equal to 4 words. Using the words $w \in F_c$ we produced a N by 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 [**?**] a window size of 8 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 also holds for microbologs. Since vector similarity measures are sensitive to the magnitude of the vectors we normalised all the vectors to a constant length. Once the HAL space was created we applied the above described Framework 2 to retrieve the desired keywords for our four categories.

2.4.1 Results

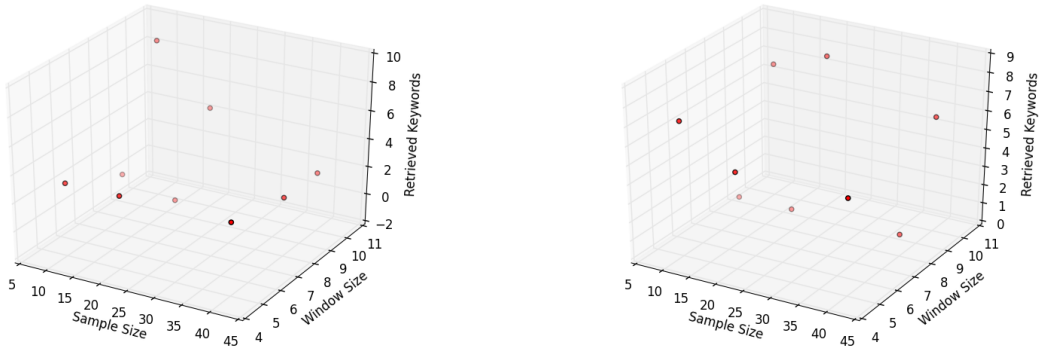
We manually assed the annotations produced by crowd flower to check for disagreements between the crowd workers an ourselves. For the category supply we rejected 26 from 69 (39%), for price 4 (12.5%) from 32, for needs 8 (7%) from 113 and for poverty 14 (%) from 106.

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 [**?**] we observe that crowdsource annotators applied a more narrow definition of of the predictor categories overlooking some keywords associated with the cateogries. 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's commonly used to describe the price mood of a commodity.

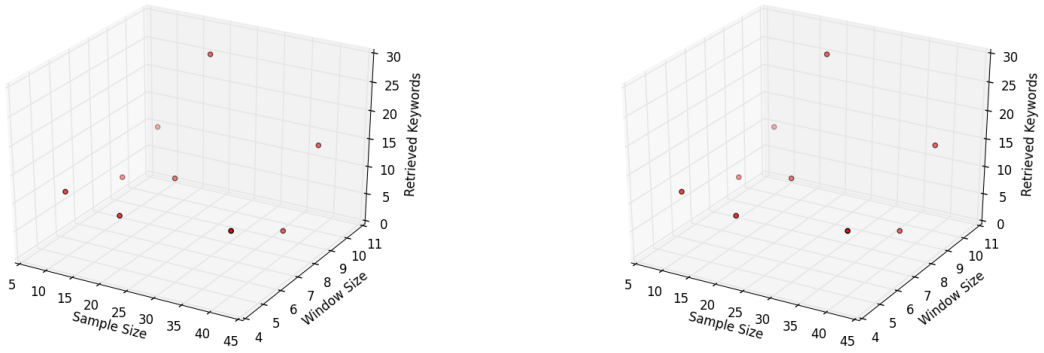
Looking at Figure **??** and Figure **??** we can observe that for all categories HAL performed best for a window size of 10 which contradicts the findings of [**?**]. Additionally we see that the smaller sample sizes consistently produce more relevant keywords then the large sample. A larger sample sizes increases the likelihood of a keywords occurrence. Since we set a fixed threshold of 100 occurrence across all samples we are more likely to include words with a smaller confidence , which might explain the poor performance.



(a) HAL - Price

(b) HAL - Supply

Figure 2.2: HAL Evaluation for Price and Supply



(a) HAL - Needs

(b) HAL - Poverty

Figure 2.3: HAL Evaluation for Poverty and Needs

2.4.2 Discussion

We observed that HAL has a very high 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 clear majority of the retrieved words. However the precision varies highly 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 similarityt HAL highlighted some topics indirectly associated with food security. For example there was a high percentage of country names in the retrieved results. Looking more closely at the retrieved countries we could see that most of them had a clear association to food. Where the majority of the retrieved countries such as Thailand, Bali [7] or the cities Singapore and Paris [8] are considered to be famous holiday destinations for food lovers other retrieved countries such as Pakistan, Syria, Jakarta India or the Philippines [9] are cities with a clear history of food insecurity and political unrest.

[7] http://www.nomad4ever.com/2008/08/24/top-10-popular-foods-of-asia-explained/
[8] http://www.hellotravel.com/stories/best-food-cities-in-world
[9] http://foodsecurityindex.eiu.com/Country

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 2.3: Keywords of Predictor Categories

2.5 Filtering

The filtering of the tweets was performed in three rounds. First we filtered for food relevant 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.

2.5.1 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 optimised 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 [**?**]. String.split() proved to be more tweekable. The result was a collection of 5.6 M tweets posted by 4.2 M users.

2.5.2 Predictor related Tweets

The first round drastically reduced our dataset to around 90 GB of tweets. This allowed us to perform a more involved filtering mechanism similar to [**?**].

For every word in a tweet and for every word in our predictor lexicon K_p the stem was computed. This was necessary to capture tweets that may contain a predictor term that is not in its base form. Fore example a tweet containing the word *pricey* would not match the term *price*. Furthermore the framework also accounts for misspelt words. To do this in a computationally efficient way the algorithm computes the edit distance between a given word and terms from the predictor set D. If the error is within a fixed threshold the predictor term with the minimal edit distance is returned.

2.5.3 Sentiment Extraction

Experiments in [**?**] showed that sentiment analysers such as SentiStrength [**?**] or Stanford CoreNLP [**?**] performed poorly on microblog content. Hence in [**?**] the decision was made to extract the sentiment by having specific terms for each sentiment (polarity). In addition 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 however 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 [**?**] with an F1 Classification Accuracy = 0.96 which outperformed human evaluators. 2.) Often keywords can not be manually assigned to a polarity without knowing it's 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. For example it accounts for emoticons which are commonly used to express a sentiment or even acronyms such as *LOL, WTF*. It's further worth mentioning that VADER is an unsupervised approach and is well suited for streaming data.

Chapter 3

Analysis

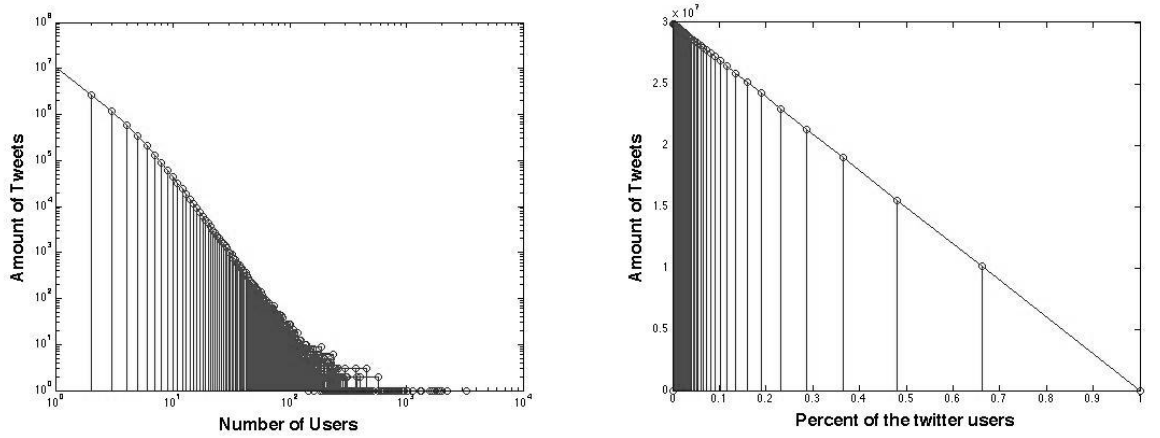
The following chapter will investigate if and to what extent Social Media data can be found to correlate with the international Food Price Index and the Commodity price quotes. This is accomplished by analysing 30 M tweets related to food. By drawing some basic statistics we want to emphasis the general popularity of food among the twitter users and describe the term distribution. In the analysis we aim to show how food terms related to each other and how they compare to Indices which are intrinsic food security indicators. Lastly we investigate to what extent Food Security and market fundamentals are present in social media discussion.

3.1 User Distribution

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

3.2 Food Term Distribution

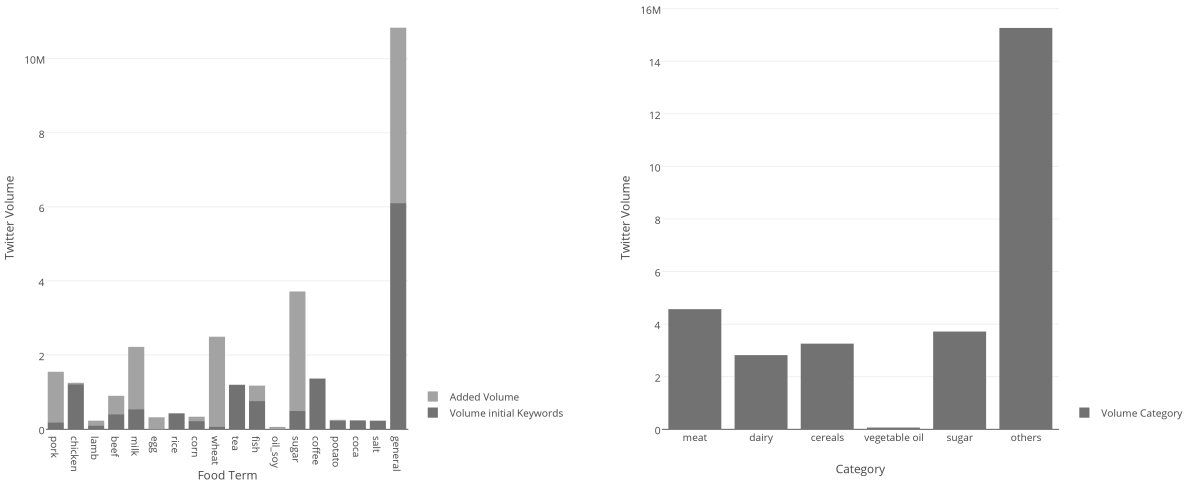
Our framework for the data acquisition successfully increased the total volume of food related tweets. From an initial 13.7 M tweets we raised the entire volume by 110% to a total of 29.9 M food related tweets. The distribution of the volume per food term is displayed in Figure **??**. We illustrate in orange the added volume alongside the initial size in blue. 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 **??**. By far the highest contribution has the category *others* due to general food related keywords such as *dinner* or *food*. It builds the absolute majority with 51 %. Meat



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

Figure 3.1: Volume of Tweets per Keyword and per Category

related keywords has the second highest contribution with around 15 % followed by 12% sugar, 11% cereals, 10 % dairy and lastly 0.2 % Vegetable Oils. Interestingly the volume roughly follows the economic importance of the different categories with the only outlier being sugar [**?**]. We assume this is due to the highly popular products *coca cola* and empty chocolate which caused alone 70 % of the sugar related tweets.



(a) Overall Distribution (b) Category Distribution

Figure 3.2: Volume of Tweets per Keyword and per Category

3.3 Price Correlation

We observed a general popularity of food in our initial analysis and that certain food categories have a much stronger presence then others. There is however still a con-

cern 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 our 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 Index and a universal Food Price Index. The Category Index is specific to a food category (e.g. meat, cereals) so different among all categories, where as the Food Price Index 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 [**?**]. 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 then 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.

3.3.1 Results

Between the meat subcategories there is a strong positive linear relationship in the range of 0.7264 and 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. 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 Indices. 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 [**?**] theory that the commodity markets have a strong influence on the the rising food prices and are a strong factor for quantifying Food Security.

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

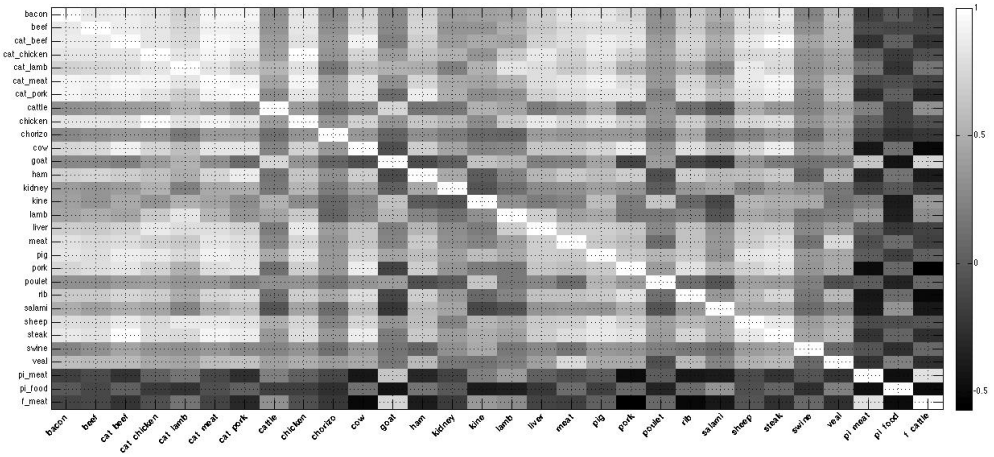


Figure 3.3: Heatplot Meat: Volume of Tweets per Keyword and per Category

For cereals similar to meat we likewise see a high correlation in volume of around 0.82 between the different cereal categories. The products beer, barley, bread, alta 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 Food Price Index. This is somewhat surprising as meat prices have a stronger influence on the universal Food Price Index then Cereals do [**?**]. Furthermore the product pasta has a strong linear relationship with the commodity price of 0.7212.

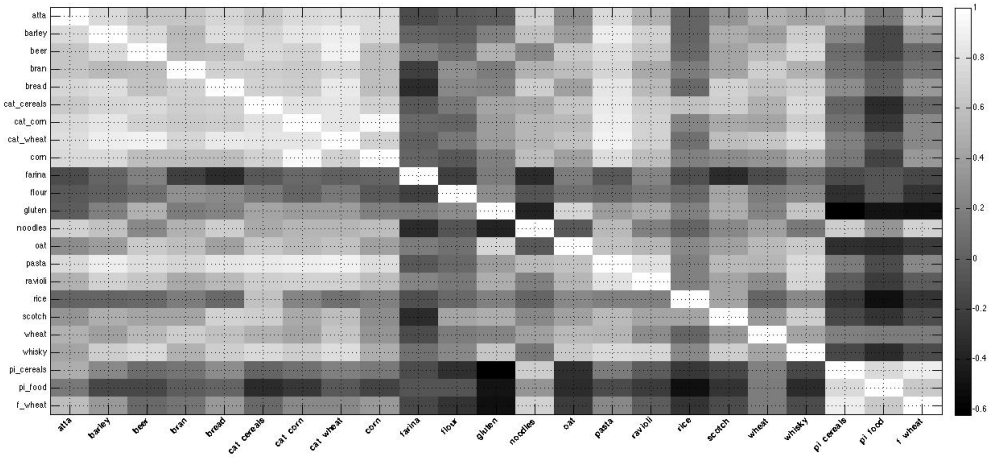


Figure 3.4: Heatplot Cereals: Volume of Tweets per Keyword and per Category

The heat plots of dairy, sugar and oil shows yet again no clear linear relationship between the twitter volume and the food price indices. More so then in other Food Categories the subcategories of dairy can be clearly distinguished between it's correlation. I.e. morzarela has a very 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.

3.3.2 Discussion

Our analysis did not show any 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 non-linear model to predict price quotes in further sections. According to [**?**] such models are better suited to utilise social media for predictions.

A smilier correlation analysis has been made by the UN [**?**]. 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 [**?**]. 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 3.1: Price Correlation

3.4 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 terms such as price) have a much stronger Pearson correlation then 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 entered around defining causes of volatile food prices. In [**?**] they define a taxonomy for drivers of international food prices spikes and differentiates

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.

3.4.1 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 **??**. 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. [**?**] 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 later 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 continues interest.

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

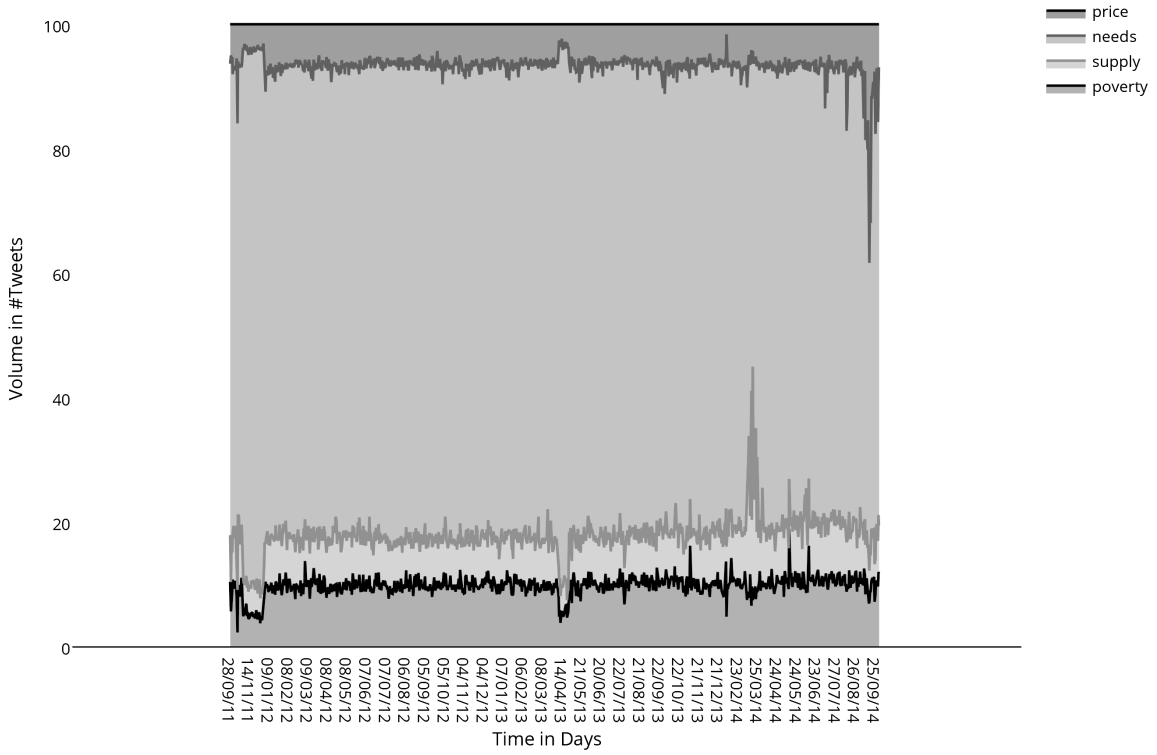


Figure 3.5: Topic Distribution - Food Security

3.4.2 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 [**?**] [**?**]. We used a fixed window size of $2m + 1$ where $m = 15$ giving us a month long window. Within the window we identified the median and calculated the mean of the twitter volume. From those values we calculated the Median absolute deviation (MAD) as follows:

$$\overline{MAD} = median_i(|X_i - median_j(X_j)|) \tag{3.1}$$

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 then 2 MAD from the mean. For this analysis we only consider positive peaks and ignore anomalies in form of a steep descent.

The discussion centred 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 [**?**]. 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 then two events caused the peak we marked it as ambiguous.

3.4.3 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[**?**]. 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 [**?**] definition. Food Waste refers to Food Loss that occurs at the retailers and consumers side where as 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 [**?**] 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 [**?**].

Stability
: Poverty, political unrest and topics concerning extreme weather [**?**] 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.

3.4.4 Results

The distribution of the annotations is visualised in Figure **??**. 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

Discussion around stability were usually headlined by extreme poverty causing riots. A food program that provided free lunch to underprivileged school kids used poisened 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 celeberaties 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 footballplayer, 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

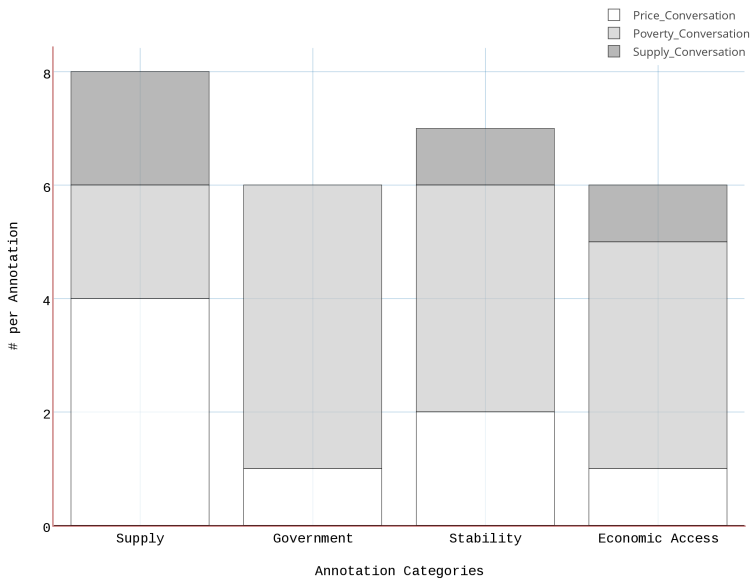


Figure 3.6: Annotation Distribution - Relevant to Food Security

3.5 Discussion

Wrap up the Analysis part.....

Chapter 4

Model Building

Intro goes here .....

4.1 A fuzzy approach for Time Series Modeling

Compared to other approaches, Fuzzy Logic has seen only a few applications in forcasting despite it's 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 [**?**] introduced the concept to describe real world phenomenas that do not have 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 mathmatical models [**?**]. In some areas such as Time Series prediction techniques such as ARMA and AR, have shown clear limitations [**?**]. Nonlinear approaches, such as ANFIS, have proven to be more successful [**?**]. Prediction accuracy is however not the only concern in forecasting models. Understanding the behaviour and gaining insight into the underlying dynamics is equally important [**?**]. This make 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. This is particularly important as the results might help us better understand the determinants of food security risks in social media.

4.2 Fuzzy Logic

In this section we present the fundamental terminologies and concepts around Fuzzy Logic. More specifically, we explain the basics of Fuzzy Variables and Fuzzy Sets. Additionally, we explain how Fuzzy Interference is performed before lastly deriving a neural network with an underlining Fuzzy Interference System.

4.2.1 Fuzzy Variables

Zadeh 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 is defined by its wavelength, which is a precise numerical definition. In natural language we tend to classify colours not by its numerical value, but by colour objects, which fall into a specific wavelength, scope or how it's commonly labelled a defined fuzzy set. Red or blue describe the object's colour, but it's by no means a precise definition. By applying a membership function of the corresponding fuzzy set (red or blue) we can precisely define the colour. With regard to our Fuzzy Logic system we will distinguish between Input variable and Output variable. The Output variable or Crisp Value depends on the Input value's corresponding membership function and a decision matrix, which we will both describe in more detail in the following sections.

4.2.2 Fuzzy Sets

Fuzzy Logic is based on fuzzy sets where the requirements for membership are not precisely defined. By introducing the concept of partial membership, the fuzzy set definition accounts for imprecise requirement for the membership of an element in a set. This means the element can take any value between 0 and 1. The most common membership functions are Gauss function, Trapezium function and the Triangle function. We choose to use a Gaussian membership function defined by Equation **??**, as they are differentiable and for our second part desirable for optimisation purposes [**?**]. We will refer to a_i, b_i, c_i as the parameter set.

$$\mu_A(x) = \frac{1}{1 + |\frac{x-c_i}{a_i}|^{2b_i}} \tag{4.1}$$

4.2.3 Fuzzy Interference System

We first describe the components of the Fuzzy Interference System (FIS) and then give intuition on how the system can be modelled as a Generalisde 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 carefully consider the input variables in order to minimize 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 **??** 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

$$w_i = \mu_{A \cup B(x,y)} = \mu_A(x) \times \mu_B(y) \tag{4.2}$$

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

$$w_i = \mu_{A\cap B(x,y)} = \mu_A(x) + \mu_B(y) - \mu_A(x) \times \mu_B(y) \tag{4.3}$$

The output of one Fuzzy Rule is computed by Equation **??**.

$$z = dx + ey + f \tag{4.4}$$

The final output of the system is the weighted average of all rules computed by Equation **??**.

$$Output = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i} \tag{4.5}$$

where N is the number of rules.

4.2.4 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 **??** to a Fuzzy Set. **Layer 2** applies Equation **??** or Equation **??** 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 **??** in this Layer.

$$\overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1, 2 \tag{4.6}$$

Layer 4 applies Equation **??**, which is similar to Equation **??** but multiplied by the factor obtained in **Layer 3** from Equation **??**. Finally, **Layer 5** computes the overall output as the summation of all incoming signals through Equation **??**. The construction yields a network with 5 layers, 16 nodes and 24 parameters (12 in Layer 1, 12 in Layer 4). The parameters are optimised through a hybrid algorithm. In a forward pass the system uses least-squares to find the parameters in **Layer 4**. In the backward pass the errors are propogated backwards and the parameters in **Layer 1** are optimised by gradient descent.

$$\overline{w_i} f_i = \overline{w_i}(dx + ey + f) \tag{4.7}$$

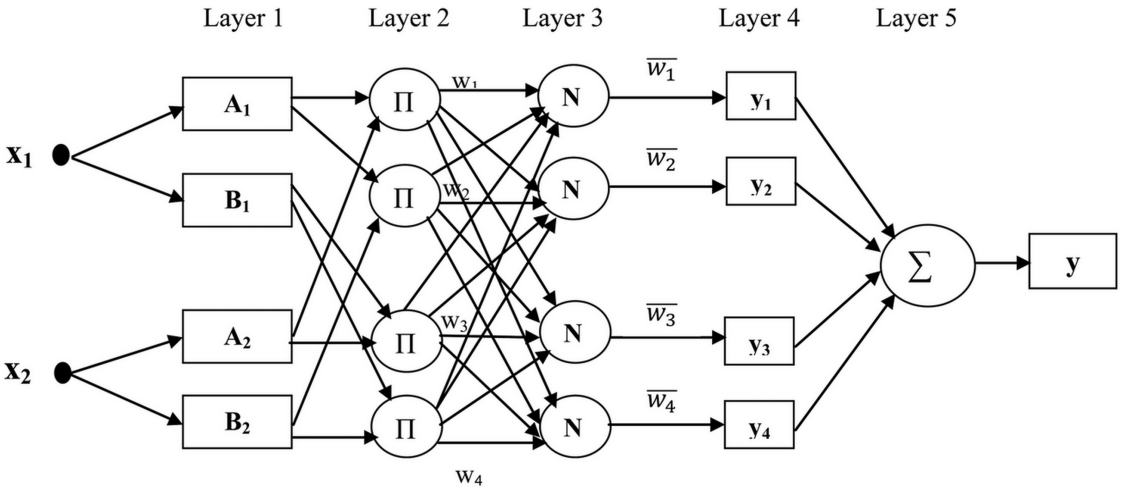


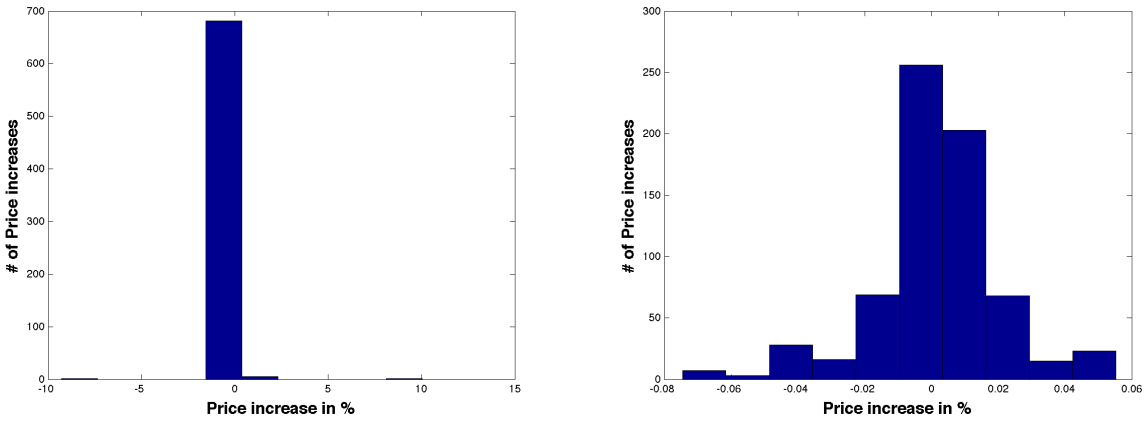
Figure 4.1: Annotation Distribution - Relevant to Food Security

4.3 Data Source and 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 in order to assist ANFIS in learning the relevant patterns. In a first instance we had to mach both the time series of the twitter data with the time series of the price data. The markets are closed during weekends and national holidays, 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 in order 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 which 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 [**?**]. 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 where α is a small constant we introduced to avoid zero divisions. It is is generally advised to normalise the data else the training algorithm might loose its sensitivity towards smaller scaled features.

$$\overline{y} = \frac{x - min}{max - min} + \alpha \tag{4.8}$$

Lastly we performed a scaling of the data which is in accordance of our objective, namely to be more sensitive to long-term than to short-term fluctuations. As we can see in Figure 4.1 a) there are some extreme price increases and drops. By applying the Hodrick-Prescott decomposition [**?**] filter we receive a distribution which is more normal by avoiding such outliers. Additionally the Figure 4.1 illustrates nicely the characteristics of commodities, namely that it's mostly driven by small price changes.



(a) Price Increase Distribution without Scailing (b) Price Increase Distribution with Scailing

Figure 4.2: Volume of Tweets per Keyword and per Category

4.4 Training the Model

For training ant testing our model we choose the time 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 [**?**]. 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 for 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. [**?**] proposed to use statistical methods to investigate the best time period for training the network. However to make any statistically significant claims we need more then 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 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 form 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 time line. To achieve a good generalisation the proportion of train and validation set always remain the same as we add additional data. We used the first 50 % of the data to train the model. From that set we excluded 15 % for validation purposes. These values were obtained empirically.

4.5 Methodology for Forcasting 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:

$$y_{t+1} = fp(y_t, y_{t-1}, y_{t-M+1}) \tag{4.9}$$

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 4.9 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 **??**. 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 compare the results across the different products.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - \overline{y_i})^2} \tag{4.10}$$

4.6 Parameters

All experiments have been conducted with the following parameters

- 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
- Step size increase rate: 1.1
- Retrain rate: 28 days
- Initial training window size: 50 %
- Segmentation: 85 % Training , 15 % Validation

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 4.1: Input Model: Benchmark Prediction

4.7 Benchmark Model

4.7.1 Input Model

Yao et al. [ref] suggests that the market can be categorised using the major trend the intermediate trend and the minor trend. Where the major trend lasts more then a year, intermediate trends are anything between 3 weeks to three months. We capture such trends by taking the moving average of 1 week, 2 weeks and 1 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 known to remove the day to day instability and extract the underlying trend.

4.7.2 Feature Selection

The objective here is to find the most significant features among the 33 in our input model in order to decrease the complexity and lastly improve the prediction accuracy of our model. To investigate the relevance we used the Relieff algorithm []. We measure the most significant features with respect to the horizon of 4 days, 7 days and 14 days. Furthermore we consider three different commodities, wheat (w), beef (b) and milk (m). The configuration in the Table *w-4* refers to wheat with a prediction horizon of 4 days. The most significant 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 refer to the relative importance of the feature for a specific configuration. It is interesting to observe that the Relieff 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. This value was empirically evaluated and is a trade off between computational complexity and accuracy. Better prediction accuracy is expected for values greater the 10 however the search will become too exhaustive. Those additional feature ($\varphi_7 - phi_10$ not displayed in Table

?? 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 **??** which are all placed in the intermediate past. This observation suggests that models based on the intermediate past poses more predictive power and outperform strategies based on features in the recent past. Robert Novy - Marx also observed this effect and investigated the hypothesis in [**?**]. Indeed 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.

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 4.2: Feature Selection: Benchmark Prediction

4.7.3 Results

The time series prediction results over a range of different horizons are illustrated in Figure **??**. To give an example of the interpretation of the model the results are obtained by applying the model describe in Section 4.5. 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 eassiest prediction task and further analysis will clearly show why. We observe that RMSE linearly increases over the time 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 **?? ?? ??**. 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 Feature Selection our input variables from the intermediate past perform exceptionally well if there is a clear underlying motion.

For the model y_{t+14} and y_{t+20} ANFIS clearly overestimates a decrease and increase in price but still manages to captures the underlining trend. On the other hand Figure **?? ??** strongly approximates the observed pattern in the training sample explaining the strong deviation from the actual prediction. The only commodity excluded from this behaviour is beef due to it's strong underlying motion. Better predictions can be expected by considering different training samples.

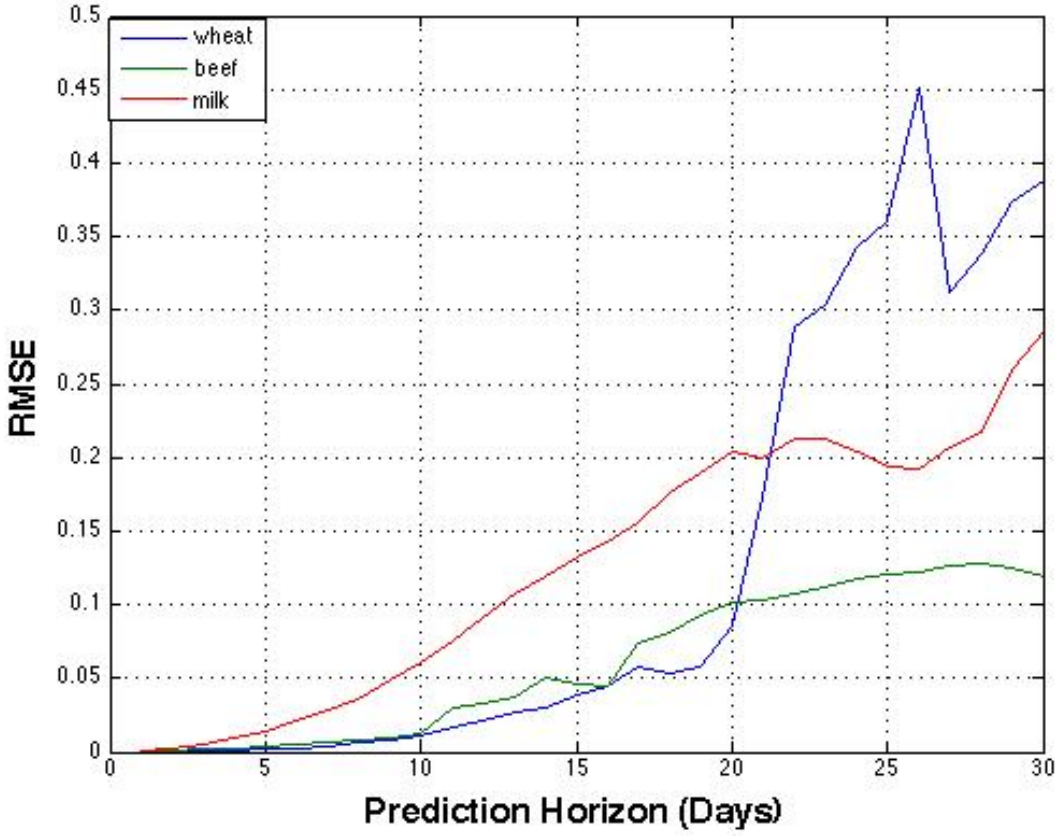
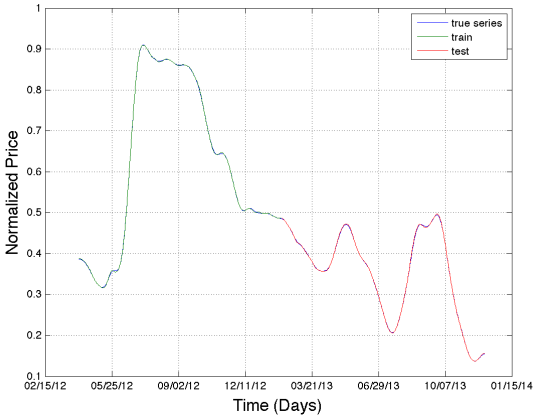
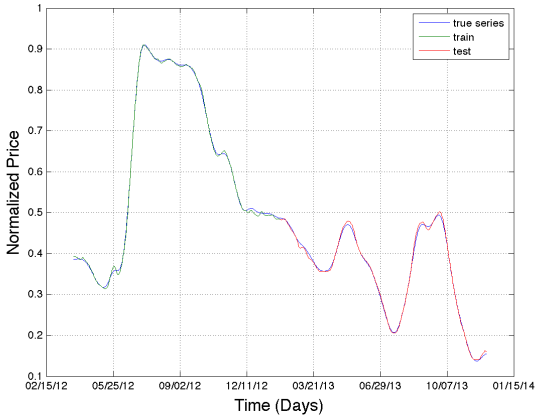


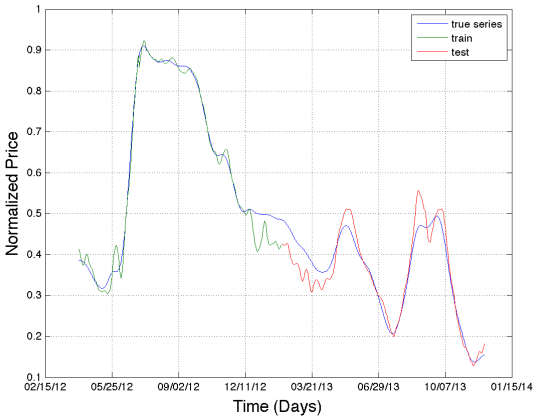
Figure 4.3: Prediction Accuracy - Benchmark



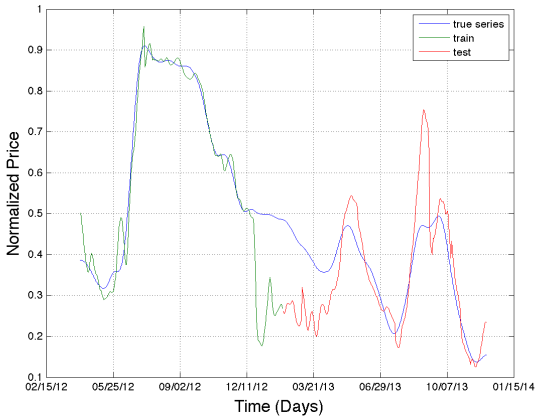
(a) 4 Day Horizon - RMSE 0.0012



(b) 7 Day Horizon - RMSE 0.0047

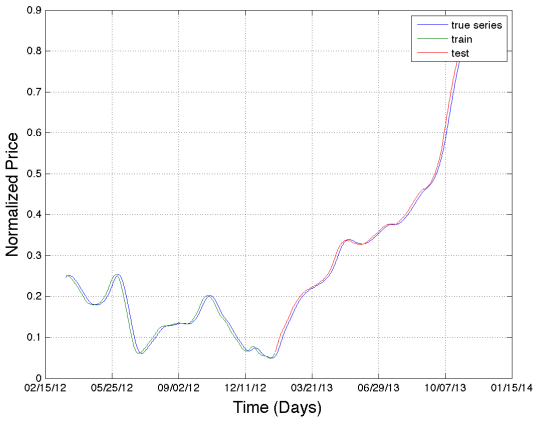


(c) 14 Day Horizon - RMSE 0.0353

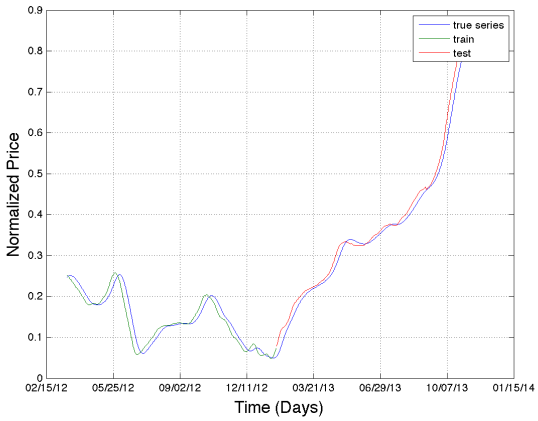


(d) 20 Day Horizon - RMSE 0.1031

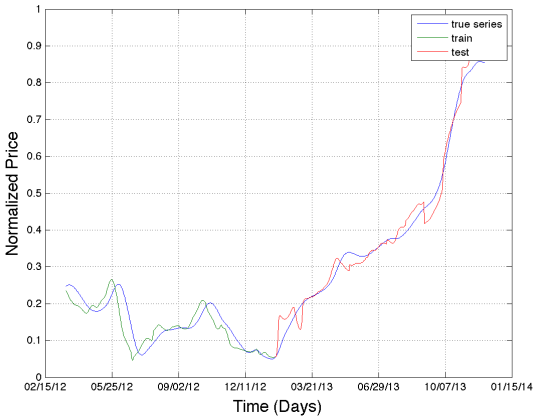
Figure 4.4: Benchmark Prediction Wheat



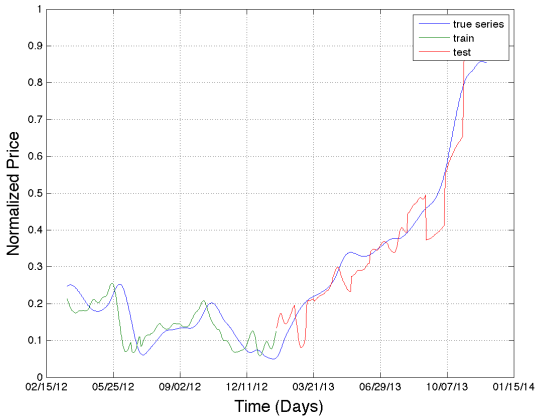
(a) 4 Day Horizon - RMSE 0.0027



(b) 7 Day Horizon - RMSE 0.0066

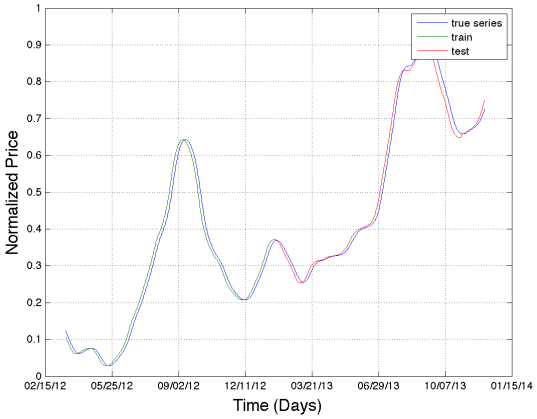


(c) 14 Day Horizon - RMSE 0.0508

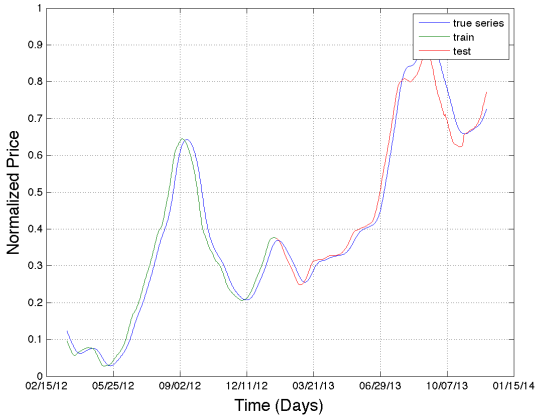


(d) 20 Day Horizon - RMSE 0.1013

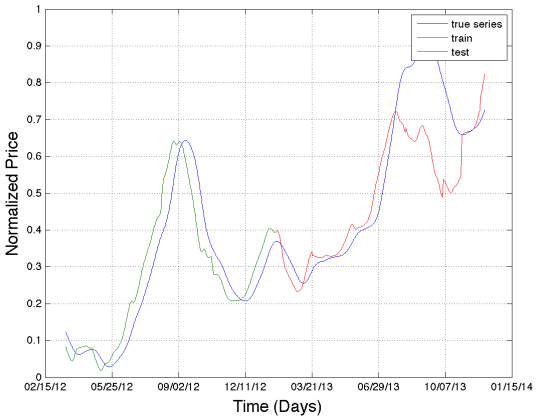
Figure 4.5: Benchmark Prediction Beef



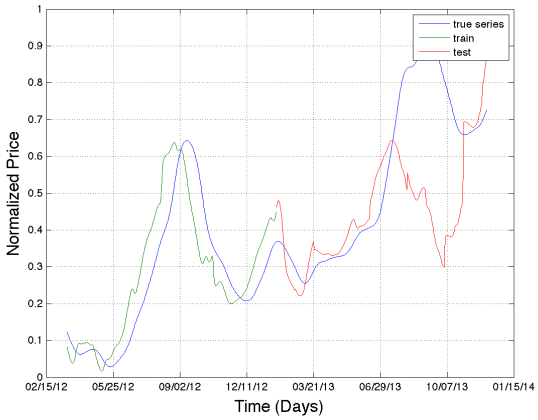
(a) 4 Day Horizon - RMSE 0.0088



(b) 7 Day Horizon - RMSE 0.0286



(c) 14 Day Horizon - RMSE 0.1185



(d) 20 Day Horizon - RMSE 0.2039

Figure 4.6: Benchmark Prediction Milk

4.8 Social Media Model

4.8.1 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 **??** we investigated if such a correlation is present and concluded that for most commodities no such linear relationship exists. We however still include the social attention as a possible feature since it might prove to be useful in form of a lagged variable (i.e. that the volume 4 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 **??**. We further include products which are not part of a subcategory but show a strong linear relationship with the predicted commodity (i.e. we consider goat as a feature for beef despite it not being a related product to beef).

Generally speaking microblogs capture one topic due to its 140 keyword limitation. The topics can usually be inferred by capturing 1 or 2 keywords. However certain terms are more intrinsic then others i.e. the keyword *IBM* can unambiguously be related to the company where as 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 **??**.

Despite having identified a significant correlation for contextually sensitive tweets our analysis in Section **??** concluded that the attention volatility is mostly driven by unrelated topics. Public mood states or sentiment might hence be a more valuable indicator. This intuition is supported by [] namely that financial decisions are not uncommonly driven by emotions and mood. For the purpose of this analysis we consider different ratios of sentiment, namely the ratio between the numbers of positive and negative tweets [**?**], the proportion of neutral and total tweets, the ratio between the numbers of non-neutral and total posts [**?**], the ratio between the number of positive and negative discussions and lastly the ratio between the numbers of neutral and non neutral messages.

Table **??** conculdes our input model. We measure the sentiment for both twitter buzz and contextually sensitive tweets. The result is an input model with 51 features.

Features	
Feature Type	Feature Example
Attention	#Commodity
Context	#Price f. Commodity #Supply f. Commodity #Poverty f. Commodity #Need f. Commodity
Sentiment Ratio	#positive : #negative tweets #positive : #(positive + negative) tweets #negative: #(positive + negative) tweets #neutral : #(positive + negative) tweets #(positive+negative) : #all tweets #neutral : #all tweets

Table 4.3: Input Model: Social Media Prediction

4.8.2 Horizon

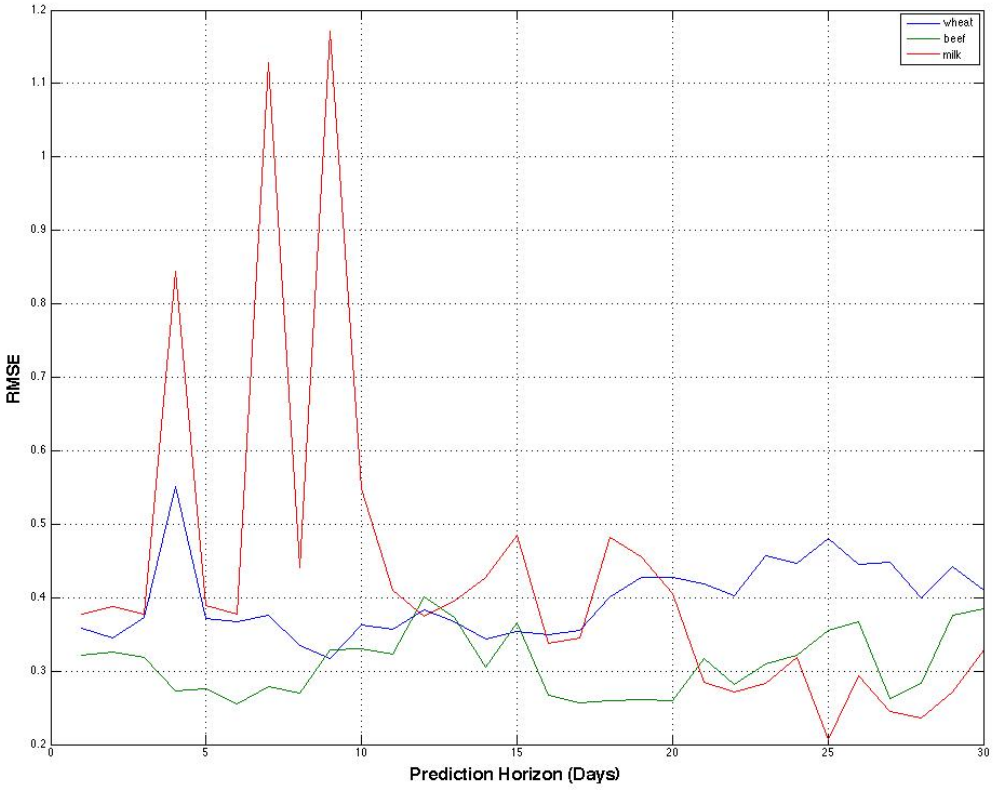
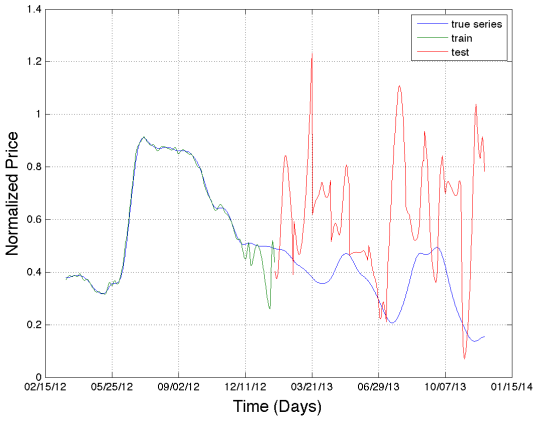
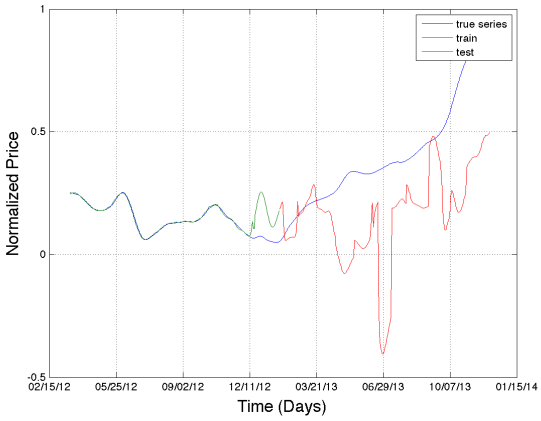


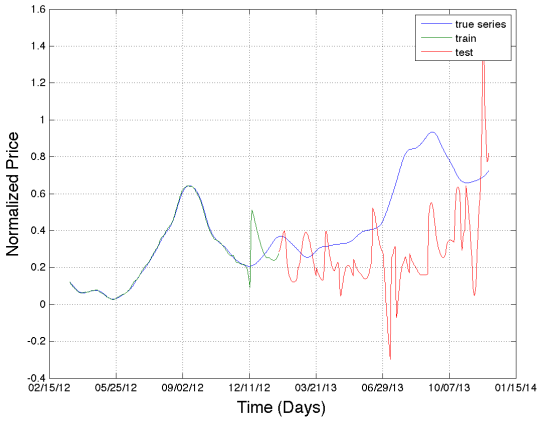
Figure 4.7: Prediction Accuracy - Social Media Features



(a) 1 Day Horizon Wheat - RMSE 0.3580



(b) 1 Day Horizon Beef - RMSE 0.3209



(c) 1 Day Horizon Milk- RMSE 0.3782

Figure 4.8: Social Media Prediction

4.8.3 Number of Features

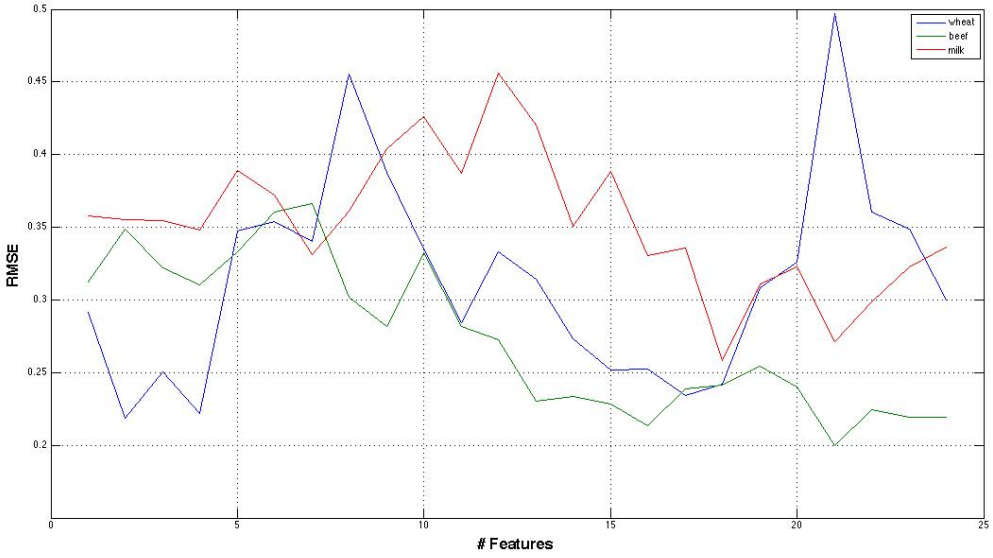


Figure 4.9: Prediction Accuracy - Social Media Features

4.8.4 Curse of of Dimensionality

For Fuzzy interference systems there are generally three types of input space partitionings. We can classify them as grid, true and scattered portionin. We first applied grid pertaining which generates rules by enumerating all possible combinations of membership functions. However for three membership functions and x features this leads to 3^x possible combinations. Instead of enumerating all possible rules we used a sub clustering method to provide a fast, one pas method to take input-output training data and generate a Fuzzy Interference System.

Data Source:

Cattle [1] Milk [2]

Corn [3] Wheat [4]

Fuzzy logic based modeling techniques are appealing because of their interpretability and potential to address a broad spectrum of problems. In particular, fuzzy inference systems exhibit a combined description and prediction capability a s a consequence of their rule based structure [27, 49]

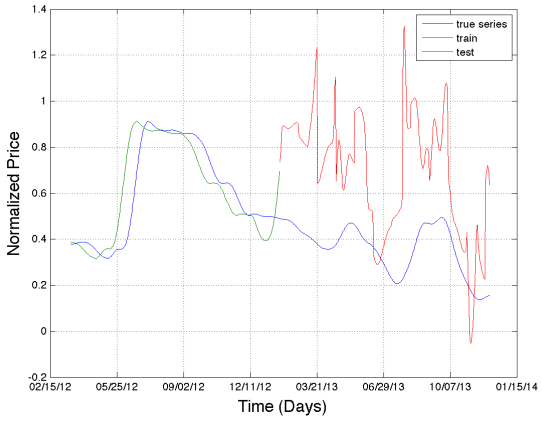
Non linear system better results in time series predictions.

[1] $https://www.quandl.com/data/OFDP/FUTURE_DA1-CME-Class-III-Milk-Futures-Continuous-Contract-1-DA1-Front-Month$

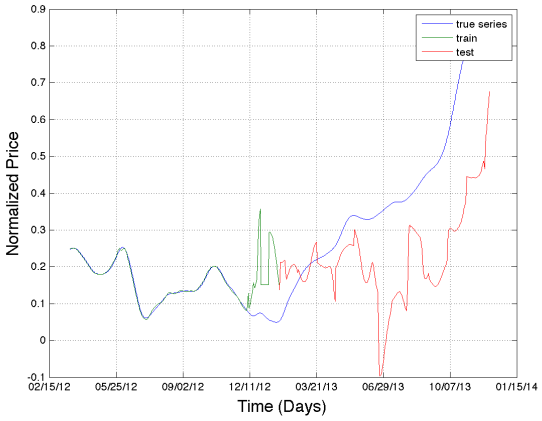
[2] $https://www.quandl.com/data/WSJ/MILK-Milk-Non-Fat-Dry-Chicago$

[3] $https://www.quandl.com/data/OFDP/FUTURE_C1-CBOT-Corn-Futures-Continuous-Contract-1-C1-Front-Month$

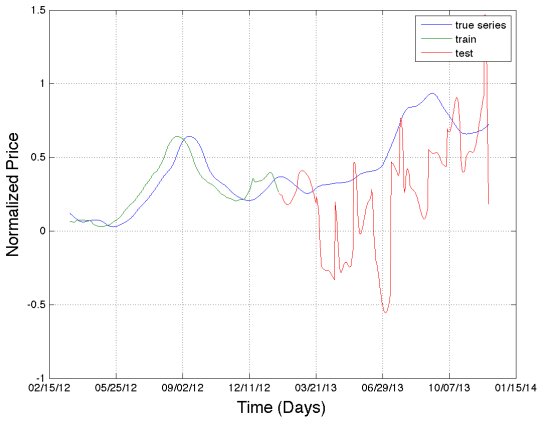
[4] $https://www.quandl.com/data/OFDP/FUTURE_W1-CBOT-Wheat-Futures-Continuous-Contract-1-W1-Front-Month$



(a) 1 Day Horizon Wheat 18 Features - RMSE 0.2422



(b) 1 Day Horizon Beef 18 Features - RMSE 0.2412



(c) 1 Day Horizon Milk 18 Features - RMSE 0.2581

Figure 4.10: Social Media Prediction

M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, ?Can blog communication 20 dynamics be correlated with stock market activity?,? in Proceedings of the nineteenth ACM

$http://www.cemla.org/red/papers2002/RED_VII_CANADA-Lalonde-Zhu-Demers.pdf$
This reference showd prediction for commodities in rang of 0.06 and 0.08 for 4 days. we are within acceptable range.

Appendix A

Data

A.1 Processing and Storage

To facilitate the storage and processing of this large amount of data we used an AMD supercomputer with 64 cores. Inspired by the map reduce paradigm we split the dataset into 64 parts and assigned each to a single core. To efficiently use the hardware resources we manually controlled for the memory assignment using numactl. As illustrated in **??** eight cores directly access one out of eight memory blocks. Each dataset was filtered in parallel reducing the 64 dataset to two lexicons.

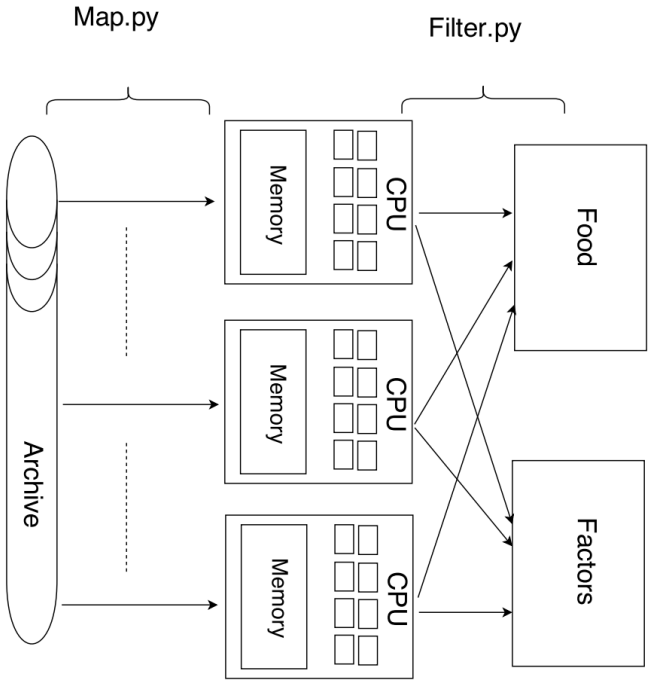


Figure A.1: Data Processing

A.2 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.2.1 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.2.2 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.2.3 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.2.4 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

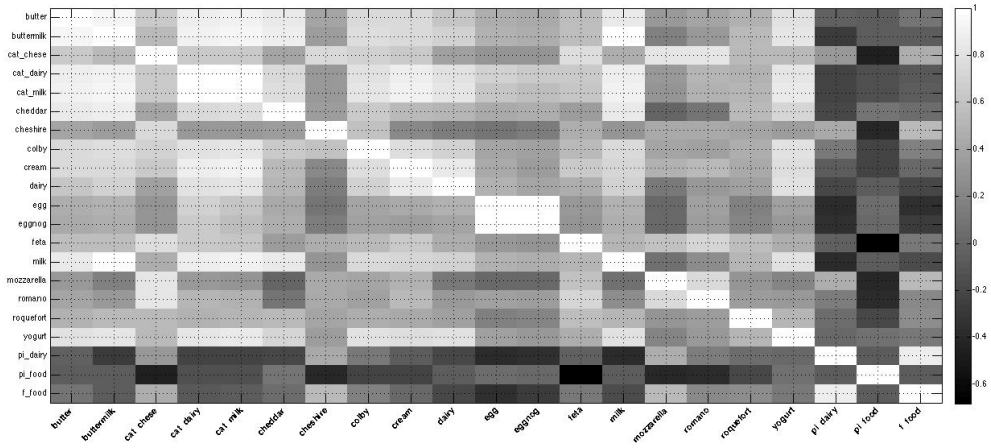


Figure B.1: Heatplot Dairy: Volume of Tweets per Keyword and per Category

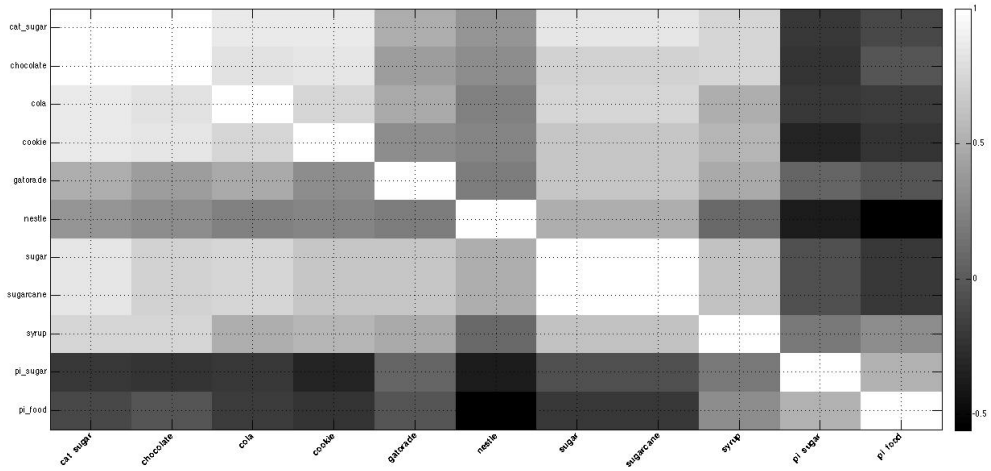


Figure B.2: Heatplot Sugar: Volume of Tweets per Keyword and per Category

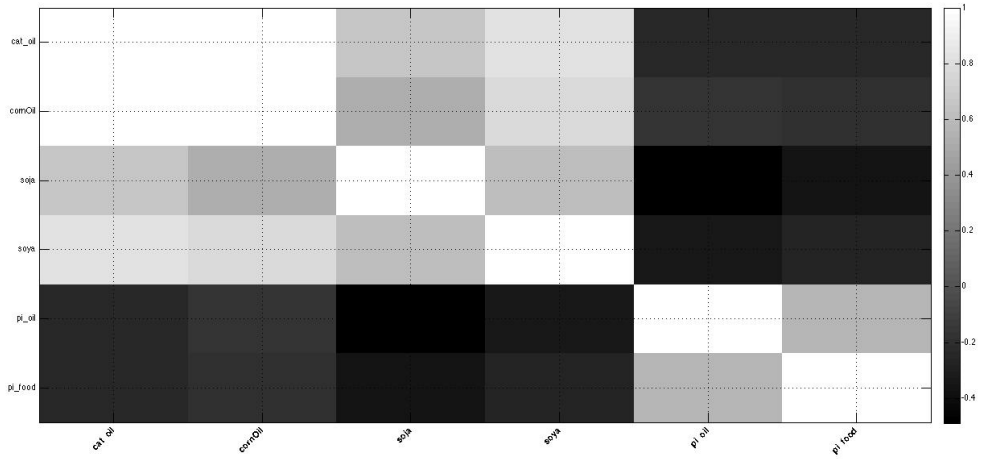


Figure B.3: Heatplot Oil: Volume of Tweets per Keyword and per Category

Benchmark Prediction - $RMSE_{Test}$ $(RMSE_{Train})$				
Horizon	7 Days	14 Days	30 Days	45 Days
Wheat				
4 Days	0.0576 (0.0495)	0.0485 (0.0455)	0.0456 (0.0434)	0.1929 (0.0381)
7 Days	0.0639 (0.0633)	0.0689 (0.0593)	0.0704 (0.0573)	0.3387 (0.0464)
14 Days	0.1185 (0.0911)	0.1147 (0.0877)	0.1116 (0.0824)	1.0089 (0.0486)
Beef				
4 Days	0.0514 (0.0240)	0.0385 (0.0241)	0.0477 (0.0221)	$1.03x10^5$ (2.77)
7 Days	0.0577 (0.0315)	0.0538 (0.0307)	0.0638 (0.0282)	0.0992 (0.0097)
14 Days	0.0816 (0.0418)	0.0979 (0.0391)	0.0666 (0.0366)	0.0664 (0.0144)
Milk				
4 Days	0.0587 (0.0439)	0.0553 (0.0445)	0.0448 (0.0423)	0.0667 (0.0344)
7 Days	0.0824 (0.0583)	0.0645 (0.0574)	0.0599 (0.0568)	0.1128 (0.0446)
14 Days	0.1136 (0.0808)	0.1053 (0.0779)	0.0989 (0.0717)	0.2628 (0.0654)