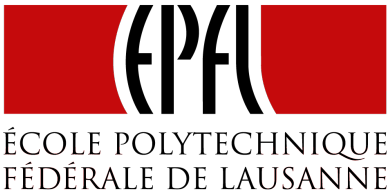


Title of Thessi

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

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 keywords 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) [13] 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 [13]. 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 [6] gives an example of a simple sentence *"The horse raced past the barn fell."* in Table 2.1 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. [6] 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

Initially we started with a simple list of food related keywords. To avoid ambiguities we will refer to the initial keyword list as K_i. As a first source for our set K_i we used the most common traded food commodities as it would easily allow us to verify our results using the price dataset made available by IMF [2]. We further decided to include the ten most important staple foods that feed the world as defined by Allianz[3].

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

We filtered the archive dataset using exact string matching on K_i. The distribution of the food related tweets showed that we need to categorise our lexicon in order to have sufficient data for further analysis. Where global keywords such as *food* are highly represented, more specific keywords such as *beef* occur infrequently. Other than the sparsity of the data we also have 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*).

To create categories we chose to mimic the categorisation 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 [7] defines the international food price index. 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.

The six subsets s are $\in K_f$ where s is one of the six categories mentioned above. C is an imaginary set that contains the five categories *meat, dairy products, cereals, vegetable oil, sugar* each being a subset containing all possible food items belonging to a specific category (e.g. the subset dairy would contain all possible dairy products). If the following relationship holds $k \in C$, where k is a keyword, for any keyword $k \in K_i$, we consider $k \in K_f$. For all keyword $k \notin C$ the condition of it being frequent is evaluated and if true added to K_f. Food commodities that could not be assigned to any of the six categories were discarded (e.g. *orange, cocoa, onion*). Upon manual examination of the dataset we realised that people are much more likely to talk about a specific food product rather than the raw material. *Cereals* are not a public interest. However products such as *bread* or *flower* occur much more frequently. The set K_f was further enriched by using food products that have been identified by [1] in set K_f only $\forall\ k \in K_f$ that are also $\in C$. To further improve our coverage of the six food categories we filtered for synonyms and contextual similar words using HAL.

2.2.1 Candidate Food Term Selection

We took several steps in order to improve our detection of the desired food commodities. K_f was created as follows:

1.) We add all keywords $k \in K_i$ to K_f only if $k \in C$ or k is frequent

2.) Further we add all keywords $k \in K_f$ to K_f only if $k \in C$

3.) We create a HAL space using a random subsample of 10% from our initial collection with all keywords that occur > 100. $\forall c \in C$ we pick the keyword $k \in K_f$ that most frequently occurs over the entire sample and retrieve the top 500 similar terms. We hand select those that are $\in C$.

The keyword set K_f 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.

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

Lexicon / Subset s	Keywords (i: from initial set, e: from K_f , 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

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 [14].

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 [9].

We define the predictor category *Access* by looking for tweets containing price as a keyword as in [9] 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 [4], however a different set of keywords was selected as described in the following chapters. Unlike [1] 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 [15] [4].

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

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

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 2.3 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 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 [13] 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

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

applied the above described Framework 2 to retrieve the desired keywords for our four categories.

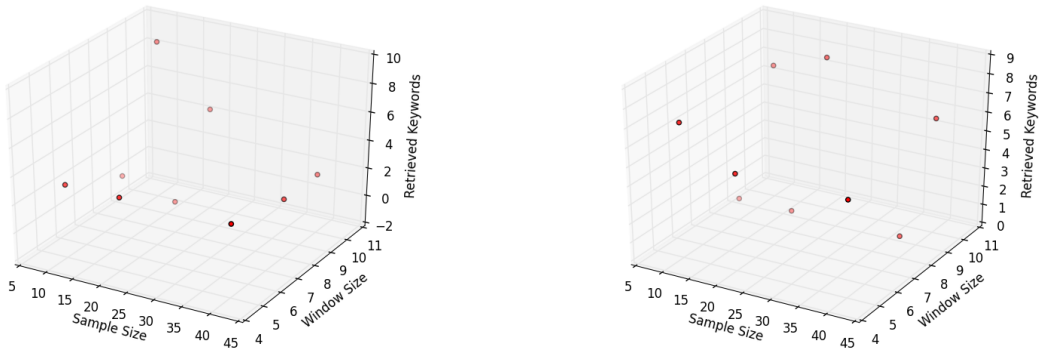
2.4.1 Results

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 [3] 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.

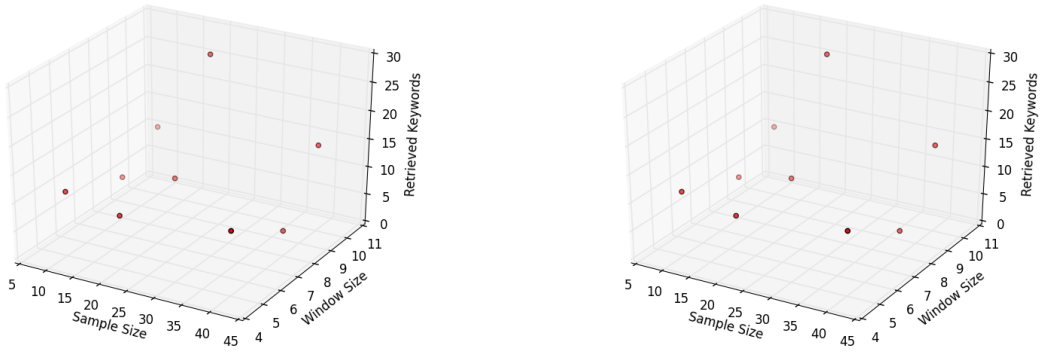
Looking at Figure 2.1 and Figure 2.2 we can observe that for all categories HAL performed best for a window size of 10 which contradicts the findings of [13]. Additionally we see that the smaller sample sizes consistently produce more relevant keywords then the large sample. A larger sample sizes increases 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 2.1: HAL Evaluation for Price and Supply



(a) HAL - Needs

(b) HAL - Poverty

Figure 2.2: 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 [12]. 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 [4].

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 [4] showed that sentiment analysers such as SentiStrength [18] or Stanford CoreNLP [16] performed poorly on microblog content. Hence in [4] 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 [8] 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

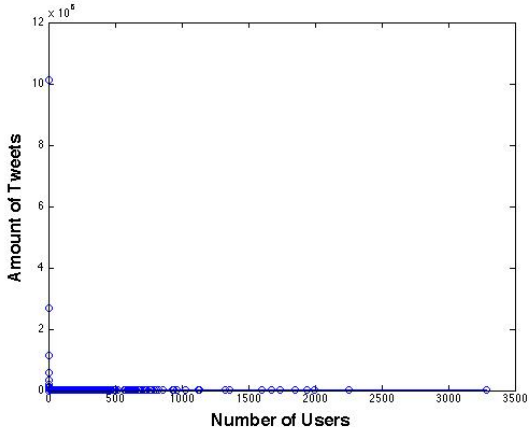
3.1 Data Analysis

3.1.1 User Distribution

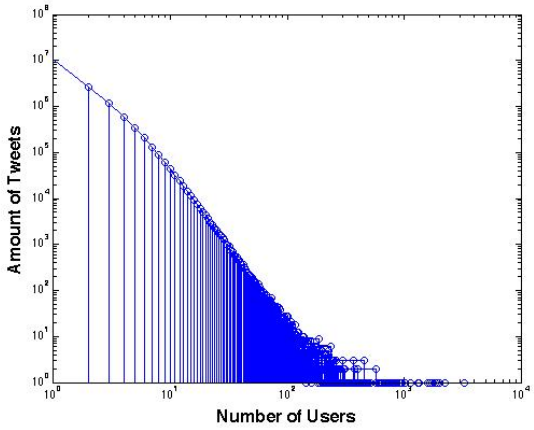
Twitter is a social network and in general such networks follow a power law distribution [19]. We see in the bellow Figure **3.1a** and Figure **3.1b** that the distribution of the number of tweets per user deviates from a normal power law. A lot of individuals send only a few tweets about the subject and only a small number of users transmit a large amount of tweets. Unlike [5] suggest the contribution participation level of 80 %, 20 % does not seem to apply to tweets about food. In Figure 3.1c 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.

3.1.2 Food Term Distribution

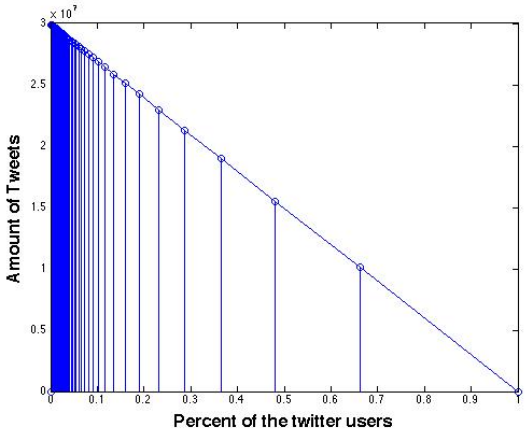
Our framework for the data acquisition successfully increased the total volume of food related tweets. From an initial 13.7 M tweets we raised the entire volume by 110% to a total of 29.9 M food related tweets. The distribution of the volume per food term is displayed in Figure 3.2a. We illustrate in 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 3.2b. 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 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 [14]. 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) Linear: Number of Tweets per User

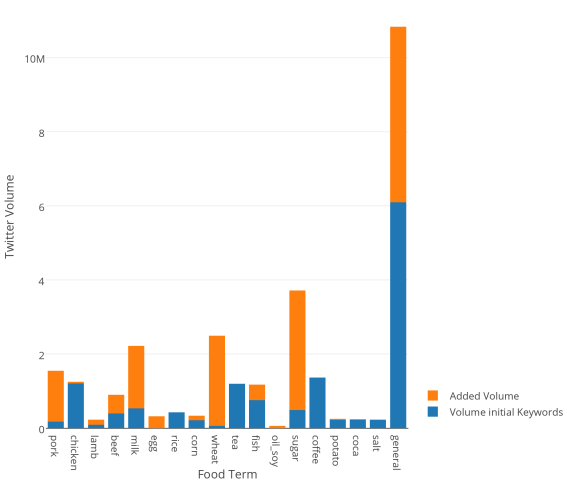


(b) LogLog: Number of Tweets per User

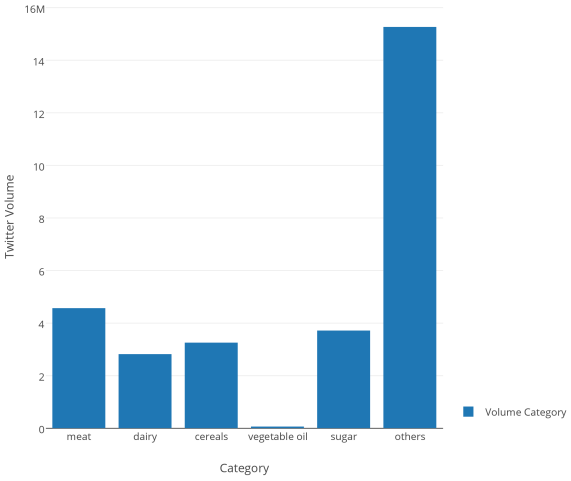


(c) Distribution: Number of Tweets per User

Figure 3.1: Volume of Tweets per Keyword and per Category



(a) Overall Distribution



(b) Category Distribution

Figure 3.2: Volume of Tweets per Keyword and per Category

3.2 Price Correlation

We observed the 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 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 our correlation analysis we used the price quotations of the Food and Argriculture Organisation of the United Nations [1]. 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 definition of the FAO [14] Since the price quotes of the FAO are based on a monthly average we aggregated the daily tweet volumes over a month and took the average volume per food term. We only included food terms that had an average of greater then 90 tweets per day.

Between the meat categories there is a strong positive linear relationship in the range of 0.9914 and 0.9980.This means that if chicken increases in volume so does beef and pork. Likewise a p value of 0.0001 suggest that we can reject the idea that the correlation is due to random sampling. A negative relationship exists between the tweet volume and the meat price index ranging from -.469 lamb to -.4855 beef. So a price increase will most likely mean a decrease in tweet volume. Generally speaking we observer a stronger correlation for the meat categories (e.g. beef, chicken) as for meat products (e.g. chorizo, salami). With a p value of ca. 0.003 we again conclude that the correlation is real.

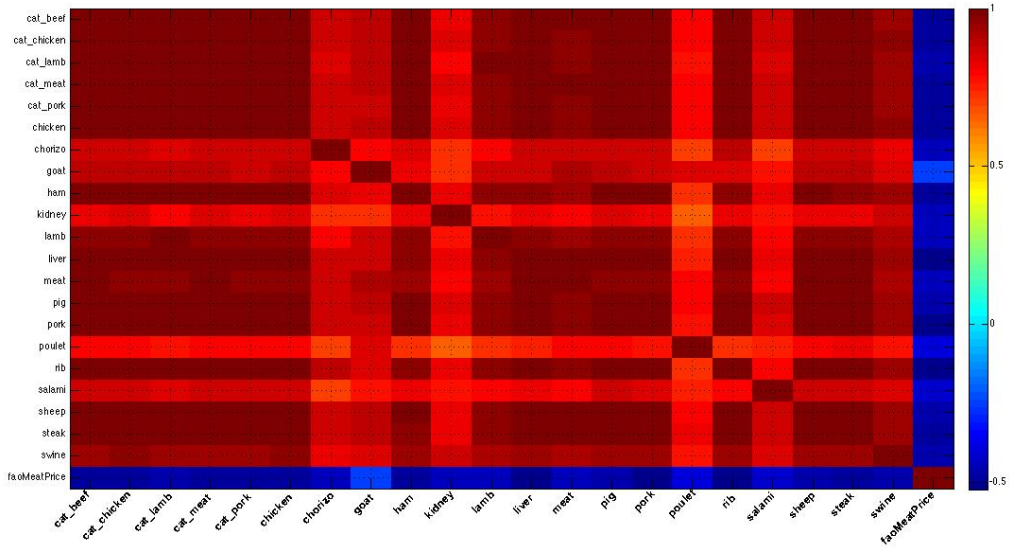


Figure 3.3: Heatplot Meat: Volume of Tweets per Keyword and per Category

For cereals similar to meat we likewise see a high correlation in volume of around 0.95 between the different products, the only exception being flower. Interestingly both bread and noodles are made from flower. We can only assume that flour producers hedge the price of wheat and do not pass the price on to bread or noodle

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

producers. Unlike meat products, when we observe an increase in tweet volume for cereals we also observe an increase in the price. The correlation of around 0.65 suggests a stronger relationship between tweet volume and price of cereals then those of meat.

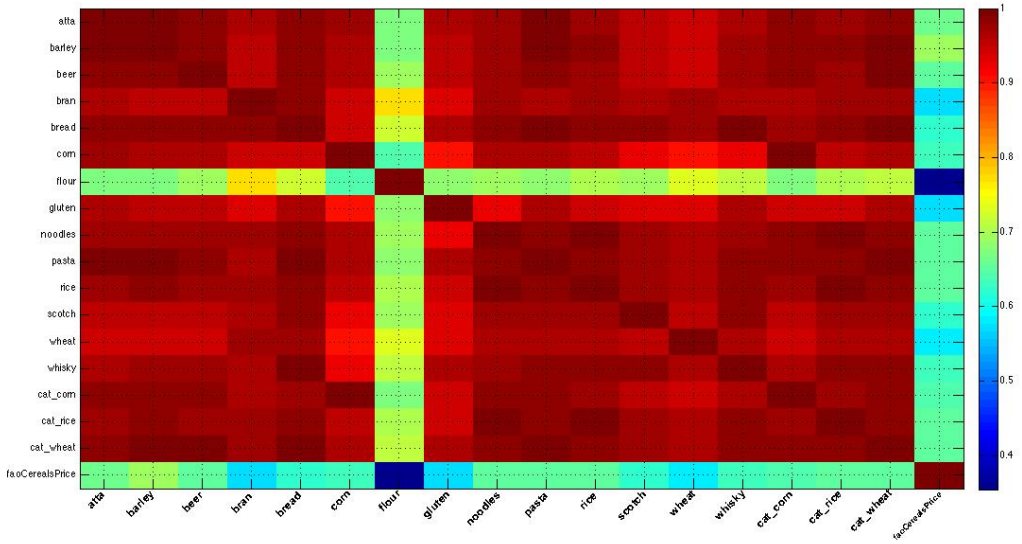


Figure 3.4: Heatplot Cereals: Volume of Tweets per Keyword and per Category

The heat plot of the dairy products is very similar to the one we observed for meet and has been added to the appendix for reference. An increase in volume of tweets suggests a decrease in price for dairy prices with a correlation of around 0.68.

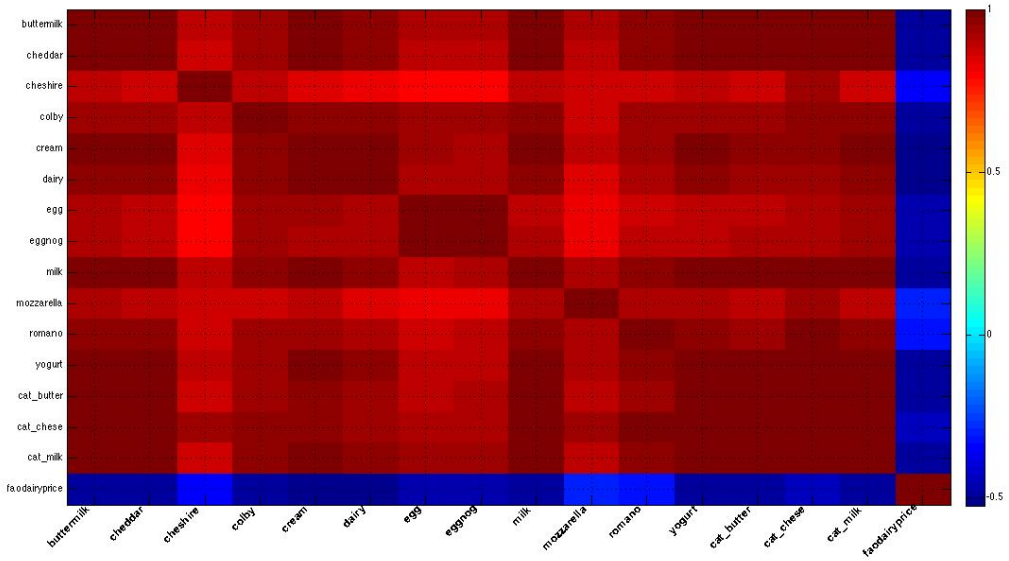


Figure 3.5: Heatplot Dairy: Volume of Tweets per Keyword and per Category

The heat plots of sugar and oil reflect a positive correlation of around 0.37.

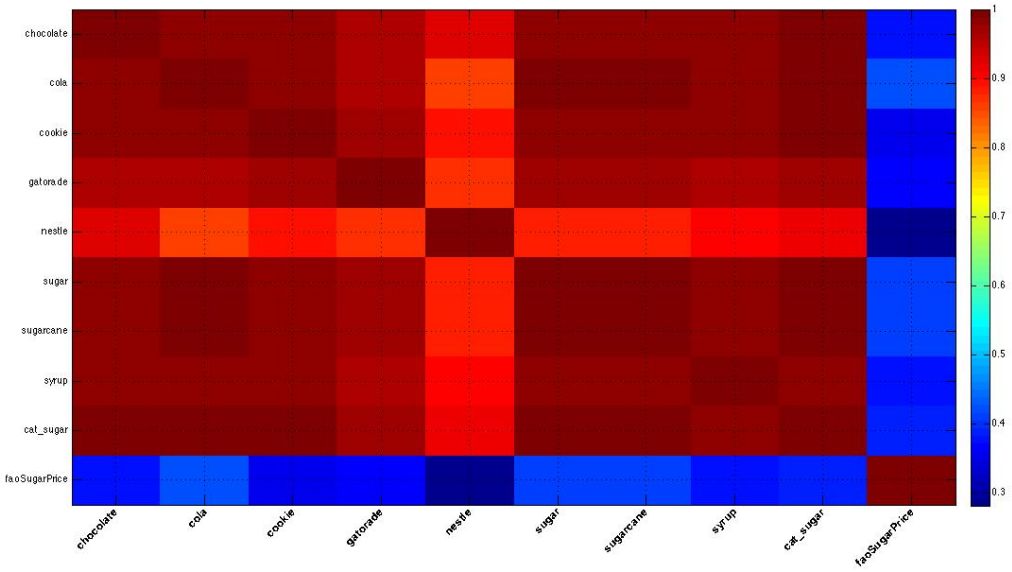


Figure 3.6: Heatplot Sugar: Volume of Tweets per Keyword and per Category

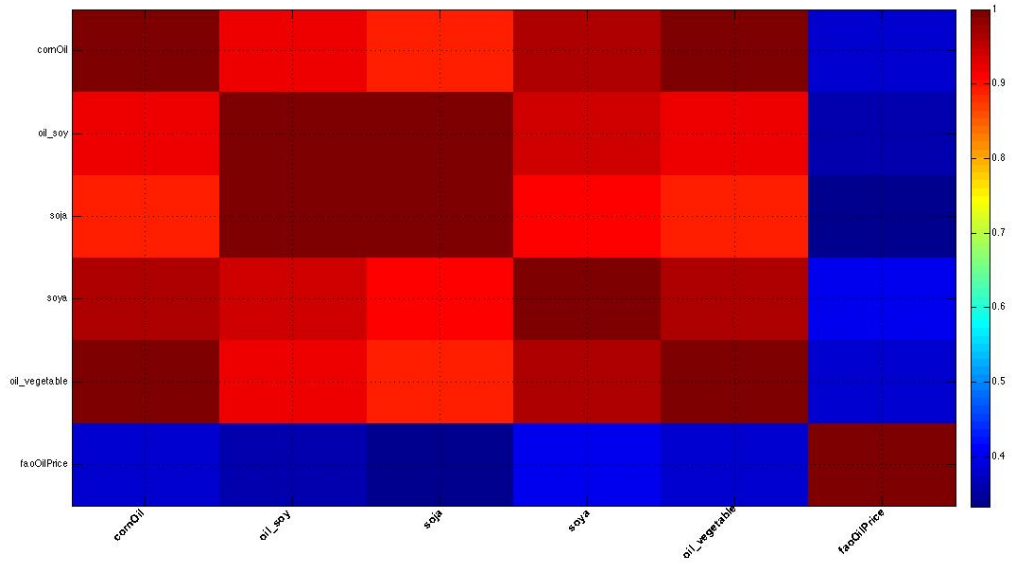


Figure 3.7: Heatplot Oil: Volume of Tweets per Keyword and per Category

3.2.1 Discussion

A smilier correlation analysis has been made in [9]. They however used more contextual sensitive tweets i.e. instead of just 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 looking at the simple raw volume of the tweets we perform significantly better with an average correlation of 0.65 and a p - value of 0.015. Where the correlation for the category prices is significant there is hardly any correlation between the international Food Price Index and the tweet

volume of the different categories. This was expected since the International Food Price Index is calculated by a weighted average of the price indices of the five food categories.

	Category Price Index	Food Price Index
Meat	-0.4802 **	0.1611
Dairy	-0.7256 ***	0.1388
Cereals	0.6489 ***	0.1543
Oil	0.3804 *	0.1485
Sugar	0.3897 *	0.1881
General	-	0.1685

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

3.3 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.

Traditional market fundamentals such as demand and supply factors were found to be inadequate to explain the recent food crises in 2007 - 2008 and 2010-2011 [2]. Recent research has been centered around defining causes of soaring food prices such as biofuel demand, trade restrictions or commodity futures markets. In [17] 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.

According to [17] , exogenous shocks are expected to generate food price spikes and volatility. The other factors determine the magnitude of the volatility and rely heavily on the political and economic condition of the country.

Motivated by this research we try to discover spikes in our twitter conversations centered around Food Security and create a link to events that might have caused them. Such events can play a significant role in explaining food price volatility.

3.3.1 Methodology

We investigate the four food categorie's 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. Lehmann et al. [11] defines there categories of temporal behaviours. Continuous activity, periodic activity or activity concentrated around an isolated peak. Where continues activities are topics that are of daily interest such as weather periodic actives reoccur with a fixed pattern such as the release of a new episode of a popular tv show. The later is event driven and usually occurs once during a very short period such as a national holiday.

In order to detect anomalies in our food topics we applied a similar approach as in [3] [11]. A fixed window size of 2m + 1 was defined where $m = 15$ to build a month long time 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 1.5 MAD from the mean. For this analysis we only consider rapid increases and ignore anomalies in form of a steep descent.

The discussion centred around food price showed 147 events, tweet activities for food supply resulted in 160 spikes. 159 anomalies were detected for food needs and lastly 153 for food poverty.

3.3.2 Social Attention

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 3.8. 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.

For price and supply we observe a similar temporal pattern. Both show a continuous activity with one extremely prominent event (boost) where the discussion about food supply is followed by a boost in food price. Both of the peaks show abnormal activities for around 9 days before and after the most prominent peak. The discussion for the price category was driven by a popular Korean pop band *T-ara* and the high volume was causesd by the tweet *[T?R? - Sugar Free] SBSPopAsiaTARA* . The music video was released on the 10th of September which caused the first anomaly. It reaches the global maximum on the 16th when they released an announcement about collaborating with a famous european DJ [2]. In the boost from conversations about food supply we observed *coffee, get, fat* to be the most popular terms. A health a lifestyle promoting coffee as a beverage to lose weight caused a high amount of retweets. We see that both boosts contain misleading indicators about pop culture and health & life style.

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

The spikes do not show any obvious seasonal patterns indicating that the conversation is very much event driven.

To investigate the content further we counted the top 50 terms of that day, ignoring stop words. The most frequent terms for the prominent price peak are *Sugar, Free, sbspopasiatara*.

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

Further observation on the other dominant pattern showed more promising indicators. In topic catered around supply and price we detected national holidays such as Thanks giving on the 22h of November 2012. Protests regarding food waste were found on the 16h of October 2013.

Supply: Thanks giving: 2012 October 16:

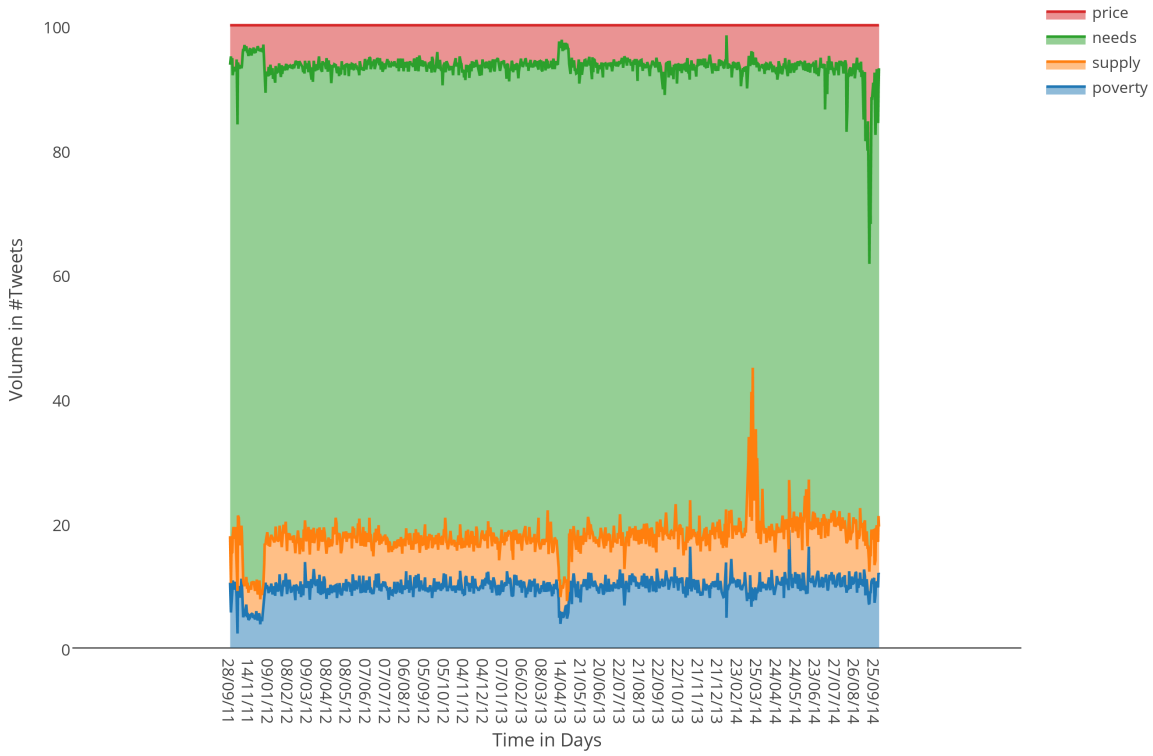


Figure 3.8: Topic Distribution - Food Security

3.3.3 Peak Identification

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 exogenous or conditional causes. For this analysis we consider discussion around food supply, price and poverty. Food price and supply are by the definition of [17] exogenous shocks and poverty is a conditional indicator. We excluded needs because it does not fit into the taxonomy.

To identify what topics spikes 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 then two events caused the peak we marked it as ambiguous.

3.3.4 Event Annotation

We annotate each peak according to the definitions given bellow. 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 - Exogenous: 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 [10] 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 - Exogenous: We define Economic Access according to FAO's [14] definition. Price, expenditure or market indicators fall into this domain.

Government - Conditional: 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 [17].

Stability - Conditional: Poverty, political unrest and topics concerning extreme weather [14] 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 (8.4 %) out of 83 fell into the above given categories, for supply 4 (4.3 %) out of 92 and poverty 13 (13 %) out of 100.

3.3.5 Results

The distribution of the annotations is visualised in Figure 4.2. 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 . 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 [3]. 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 [4]. As a consequence there was also a sharp increase in discussion around food banks. The UK observed a 51 % increase in food bank users [5]. 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 [6]. Lastly, discussion around

[3] http://www.theguardian.com/science/2012/apr/25/mad-cow-disease-us-mutation

[4] http://www.cnbc.com/id/101588110

[5] http://www.bbc.com/news/business-27032642

[6] http://www.theguardian.com/society/2013/dec/17/government-under-fire-eu-funding-food-banks

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. [7]

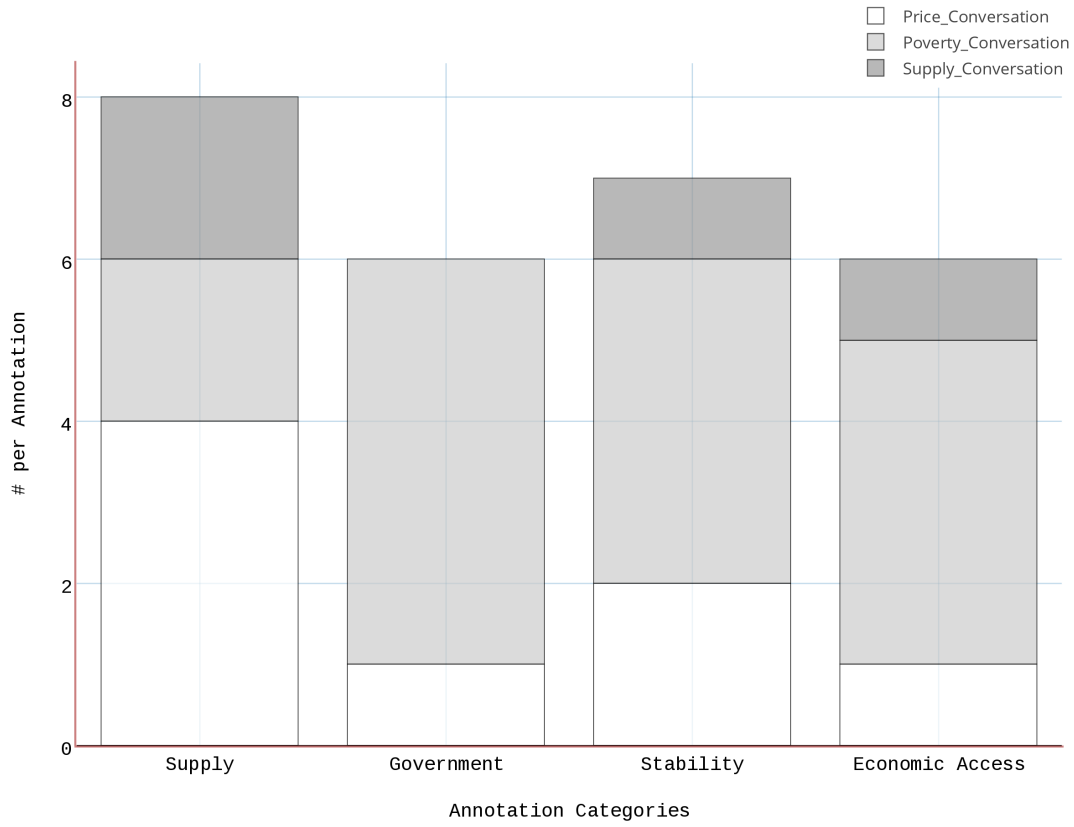


Figure 3.9: Annotation Distribution - Relevant to Food Security

[7] http://www.usatoday.com/story/news/world/2013/07/17/india-children-deaths/2523727/

Chapter 4

Model Building

Commodities are traded over 5 days. The markets are closed during weekends and national holidays. 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, corn and cattle we removed the 12/11/12 and the 8/10/12 which are the veteran day and the columbus day respectively. We preprocessed the time series of the tweets to exclude weekends and national holidays to match the time series of the commodities. The period considered is the 03/01/2012 - 26/09/2014.

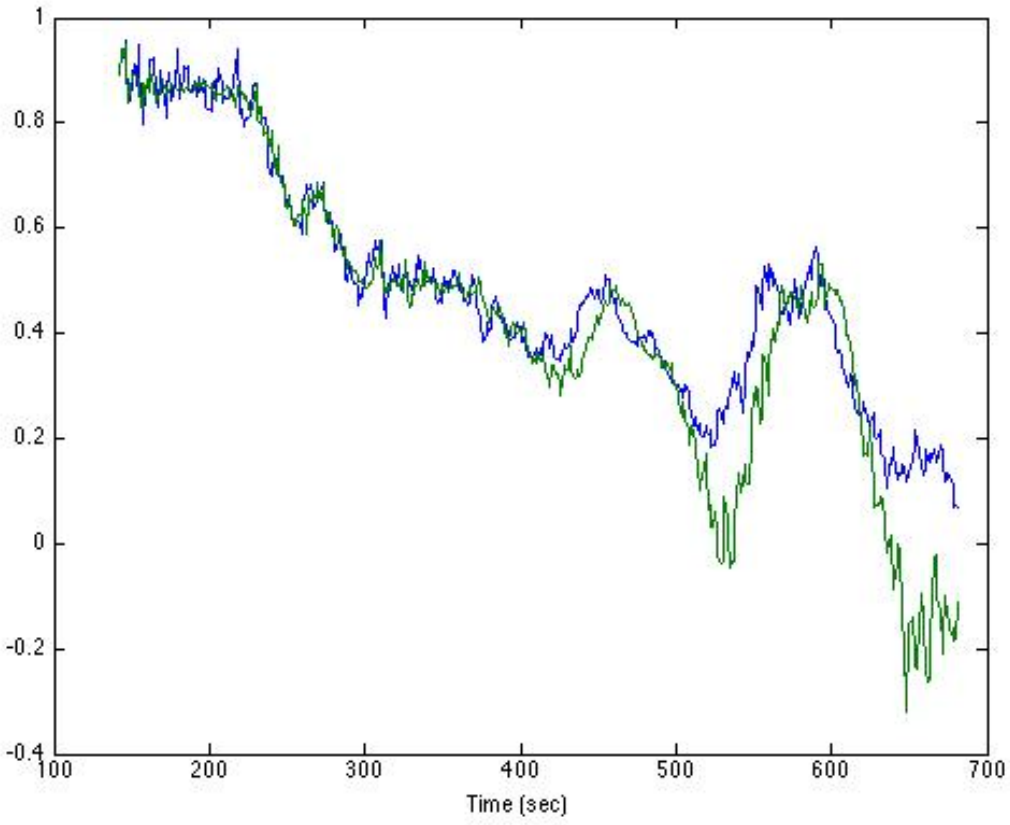


Figure 4.1: 50 % train / 50 % test

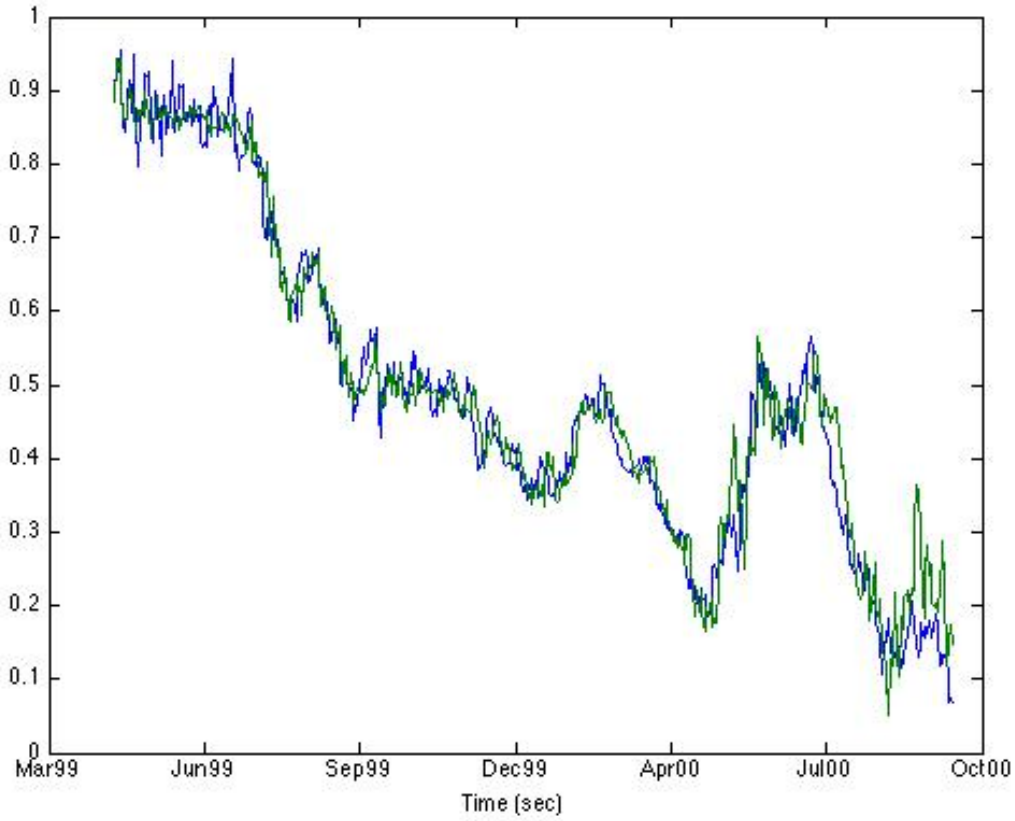


Figure 4.2: 70 % train / 30 % test

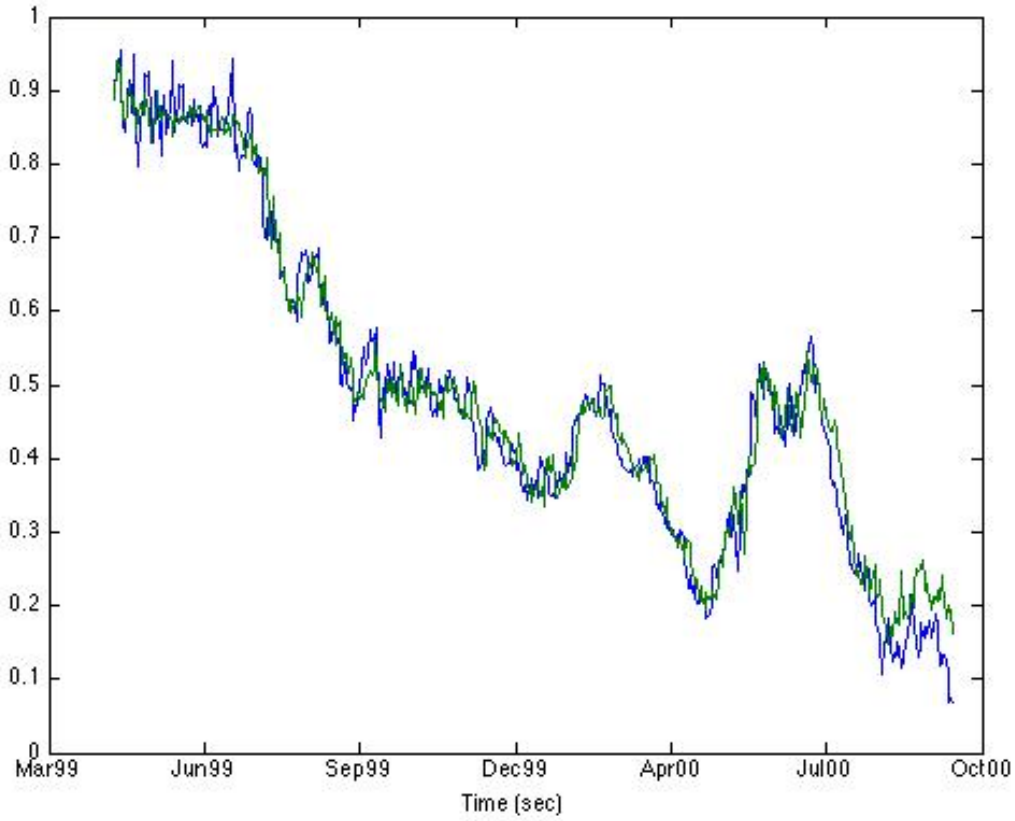


Figure 4.3: 85 % train / 15 % test

Data Source:

Cattle [1] Milk [2]

Corn [3] Wheat [4]

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

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 **A.1** eight cores directly access one out of eight memory blocks. Each dataset was filtered in parallel reducing the 64 dataset to two lexicons.

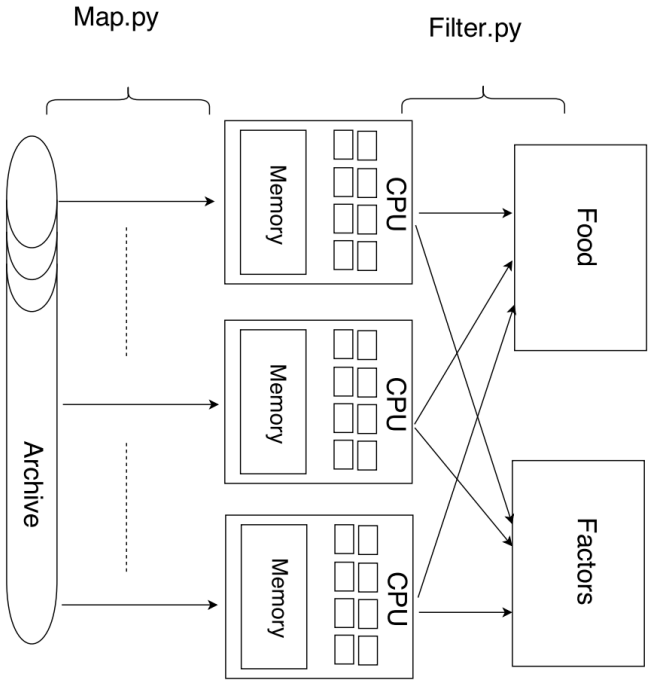


Figure A.1: Data Processing

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.

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

[7] Food and Agriculture Organisation of the United Nations. Faos food price index revisited, 2013.

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

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

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

[11] 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.

[12] 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.

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

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

[15] 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.

[16] Stanford. Corenlp. 2011.

[17] 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.

[18] 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.

[19] 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.