ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Title of Thesis

Master Thesis

S. Tudent

January 19, 2038

Advisors: Prof. Dr. A. D. Visor, Dr. P. Ostdoc
Department of Computer Science, EPFL

Abstract

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

Contents

Contents **iii**

1 Introduction **1**

2 Data **3**

2.1 Hyperspace Analogue to Language 3

2.2 Food Keyword Selection 4

2.2.1 Our Approach 5

2.3 Predictor Keyword Selection 5

2.3.1 Motivating a Semantic Approach 5

2.3.2 Evaluating HAL 6

2.3.3 Results 8

2.3.4 Discussion 9

2.4 Factor Keyword Selection 9

2.5 Filtering 9

2.5.1 Evaluation 10

2.6 Discussion 10

3 Analysis **11**

3.1 General Stats 11

3.2 Food Category Analysis 13

3.2.1 Discussion 15

A Data **17**

A.1 Processing and Storage 17

A.2 Crowd Flower 18

A.2.1 Categorise: Food Price 18

A.2.2 Categorise: Food Supply 18

A.2.3 Categorise: Food Poverty 19

A.2.4 Categorise: Food Needs 20

Bibliography **21**

Chapter 1

Introduction

Chapter 2

Data

In this section we describe the filtering process of the tweets and the creation of three 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*). The external factors lexicon similar to the former predictor lexicon tries to capture the price and supply of different food commodities (e.g. *oil, unemployment, flood*). They differ that the former looks explicitly at keywords directly associated with food terms which give an indication about the price fluctuation. The later is concerned with keywords describing external factors such as *oil*. We downloaded 2 TB of tweets from the internet archive [1] over a span of October 2011 - September 2014. The filtering process resulted with 1047698 food relevant tweets and 523549 tweets of influencing factors.

Firstly, we detail an algorithm Hyperspace Analogue to Language (HAL) [7] which was used to find relevant keywords for our lexicon. We then describe our framework for retrieving food related keywords that form our food lexicon. The subsection Feature Definition describes how we define different predictor categories, followed by an illustration of the procedure we applied to retrieve the predictor keywords. In the chapter Factors Keyword Selection we motivate the term selection of the external factors lexicon. Lastly, we describe the filtering algorithm used to create our dataset.

2.1 Hyperspace Analogue to Language

HAL creates a semantic space from word co-occurrences [7]. 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 [4] 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 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.

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

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

2.2 Food Keyword Selection

The filtering of the dataset was initially performed 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]. 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. The result was a collection of 1047698 tweets posted by 949085 user.

Looking at the distribution of the food related tweets we realised that we would have 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* → *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 [5] 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 $\in K_e$ 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 $k \in K_e$. For all keyword $k \notin C$ the condition of it being frequent is evaluated and if true added to K_e.

[2] http://www.imf.org/external/np/res/commod/index.aspx

[3] http://knowledge.allianz.com/demography/health/?767/the-worlds-staple-foods

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

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_e 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 Our Approach

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

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

2.) Further we add all keywords $k \in K_f$ to K_e 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 $k \in K_e$ 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_e was used to perform exact term matching on the tweets collected from the internet archive. The resulting set of keywords in K_e forms our Food Lexicon.

2.3 Predictor Keyword Selection

From our basic food lexicon we proceeded to extract features that we could use to predict the price and the global food security index. 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 [9].

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 [6].

We define the predictor category *Access* by looking for tweets containing price as a keyword as in [6] 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 [2], 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 [10] [2].

2.3.1 Motivating a Semantic Approach

In this section we try to comprehend which words are associated with the above-mentioned categories in C_p. More specifically what words are represented in the context of *food supply, food price, food needs* and *food poverty*. 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.

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_e 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_e 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_e 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_e Sugar	sugar (i), sugarcane (i), syrup (e), energy drink (e), cola (e), chocolate (e), nestle (e), cookies (h), cupcakes (h)
K_e 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_e 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

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. [4] 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.

2.3.2 Evaluating HAL

Here we describe an evaluation of HAL and two different frameworks we applied to create the predictor lexicon. 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. We will refer to this set of words as F_c. 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 [7] 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 evaluated the following two frameworks to retrieve the desired keywords for our four categories. The manual selection of the keywords was done through crowd flower [5]. The works 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 pice 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.

Framework 1

1.) $\forall k \in K_e$ 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.

In our evaluation 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.

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

2.3.3 Results

Other than the words for our categories of interest HAL highlights some clear topics associated around food. As expected other contextual similar words were other food items building the clear majority of the retrieved words. Interestingly there was also a high percentage of country names in the retrieved results. Looking more closely at the retrieved countries we could see that most of them have a clear association to food. Where the majority of the retrieved countries such as Thailand, Bali [6] or the cities Singapore and Paris [7] are considered to be famous holiday destinations for food lovers other retrieved countries such as Pakistan, Syria, Jakarta India or the Philippines [8] are cities with a clear history of food insecurity and political unrest.

Words that fell into our categories of interests were words such as available, profit, price, sustainability, progress, sales, war, easy. The four words we used to model the predictor categories were *available* for supply, *price* for price, *children* for poverty and *help* for needs. Where the first two terms are self explanatory the term children was selected because in [8] anthropological studies have shown that children are seen as a symbol of poverty and social exclusion . For the category needs keywords such as *yum* or *foodporn* had a high similarity with the term *food* but those terms are nearly exclusively used to express a positive sentiment. Words that also showed high similarity such as *abuse* and *protest* only retrieved a small amount of relevant keywords. We hence explored the results of the other categories and found *help* which was retrieved from the results of *price*. It was deemed as a good indicator as it can express both a positive and negative sentiment associated with *needs*.

[6] http://www.nomad4ever.com/2008/08/24/top-10-popular-foods-of-asia-explained/

[7] http://www.hellotravel.com/stories/best-food-cities-in-world

[8] http://foodsecurityindex.eiu.com/Country

Lexicon / Subset *s*	Keywords (i: from initial set, e: from K_f , h: from HAL space)
Food Supply	*available*, giveaway, coupons, told, growth, sustainability, nomnom, receive, indonessia, berlin, program, institute, survey, news, farming, journal, strategy, price, india, check, canada, production, campaign, protection, imports, launched, rating, storage, nutrition, restaurant, resources, trends, container, stall, government, distribution, processing, impact, policy, consumption, stores, exports, opportunities, harvest,price,savings,discount, budget, profits, increase, rise,relief
Food Price	*price*, issue, india, coupon, health, children, news, malaysia, discount, benefit, syria, asia, indonessia, philippines, grothw, dubai, consumer, campaign, sold, agriculture, available , sustainability, thailand, farming, markets, harvest, program, success, foundation, crops, politics, demand, purchase disaster, rates, safe, cost, association, nutrition, nation, sponsored, fundraiser, protest, deal, giveaway, growing, dangerous, threat future, programs, fighting, farms, consumers, support, jakarta, pakistan, africa, curtesy, poverty, exports drought, funding, bill, summit, delhi, rating, priced, justice, avoid
Food Poverty	*children*, community, source, future, issue, safe, project, growing, support, benefit, india, health, asia, baby, government, dangerous, area, agriculture, politics, poverty, cultures, obesity, tax, changes, program, freedom, price, impact, news, report, nutrition, help, country, syria, sustainability, philippines, success, awesome, farm, donate, diet, foundation, indonesia, summit, supplies, israel, farms, farming, kills, cash, crops, confernece, projects, seeking, nation, fight, protection, courtesy
Food Needs	*help*,power, amazing, thanks, future, children, beyond, yummy, issue, death, killing, helping, brilliant, delicious, awesome, tasty, freedom, kill, needed, nice, healthier, benefits helps, feeding, love, tax often, health, incredible, politics, destroy, expensive, increase, yum, heavenly, trash, necessary, cheap, enjoy, smiling, struggle, disaster, stress

Table 2.3: Keywords of Predictor Categories

2.3.4 Discussion

for 40 we had 119024420 words

2.4 Factor Keyword Selection

2.5 Filtering

Following the creation of the four categories we will use polarities to model the price variation. For example, the category *price* has two polarities: *high* and *low*. In [2] we worked on a filtering mechanism to extract relevant tweets and to assign them to the relevant polarities. In order to achieve this goal, a prediction lexicon with a total of four categories was built (*price, poverty, needs, supply*). We will refer to it as D.

For every word in a tweet and for every word in D the stem is computed. This is 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.

Experiments showed that sentiment analysers such as SentiStrength [12] or Stanford CoreNLP [11] performed poorly on microblog content. Hence, 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. There were 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.

2.5.1 Evaluation

2.6 Discussion

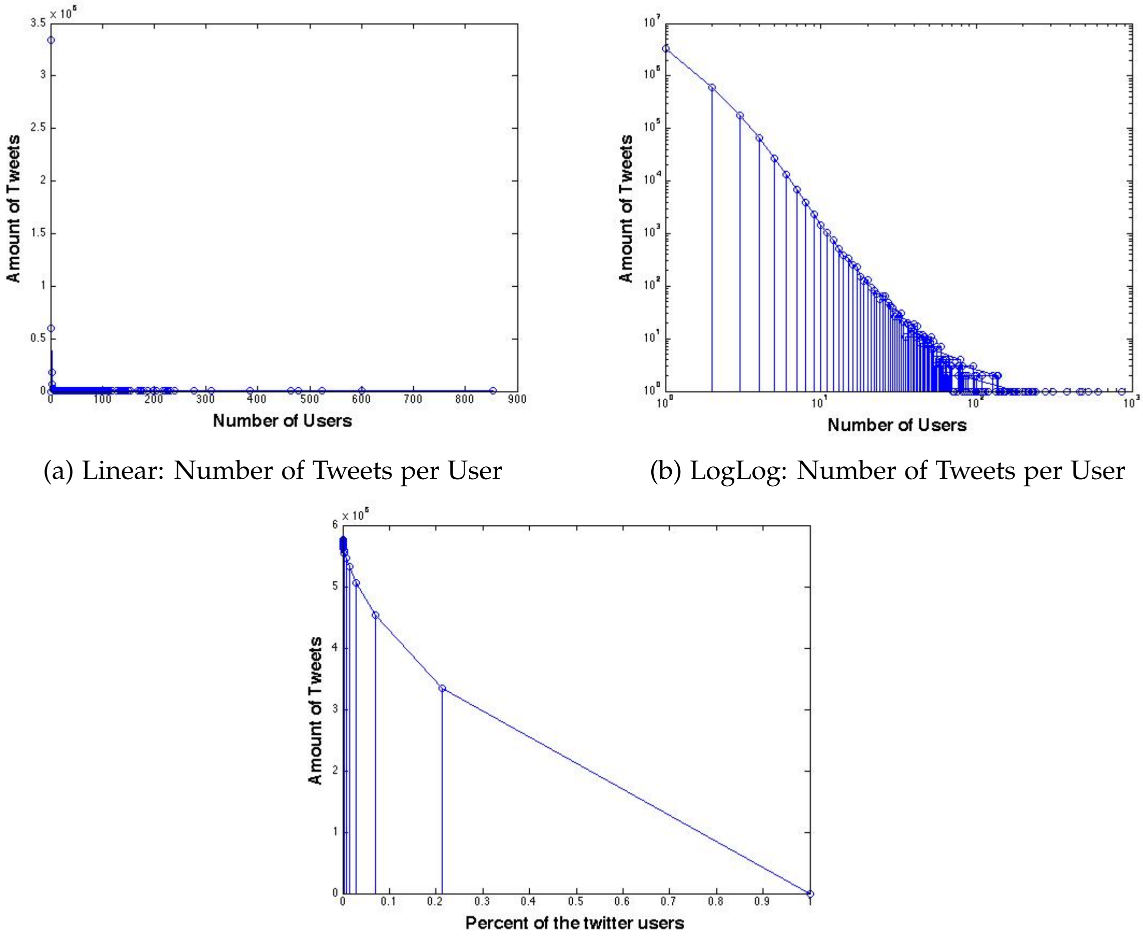
Chapter 3

Analysis

3.1 General Stats

Twitter is a social network and in general such networks follow a power law distribution [13]. We see in the bellow Figure **3.1a** and Figure **3.1b** that the distribution of the number of tweets per user slightly deviates from a normal power law. A lot of individuals have sent only a few tweets about the subject and only a small number of users have sent a large amount of tweets. Unlike [3] 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 very flat. About 40 % of the tweets are caused by 20 % 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.

Our framework for the data acquisition successfully increased the total volume of food related tweets. From an initial 2.6 M tweets we raised the entire volume by 110% to a total of 5.6 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 [9]. We assume this is due to the highly popular products *coca cola* and empty chocolate which caused alone 70 % of the sugar related tweets.

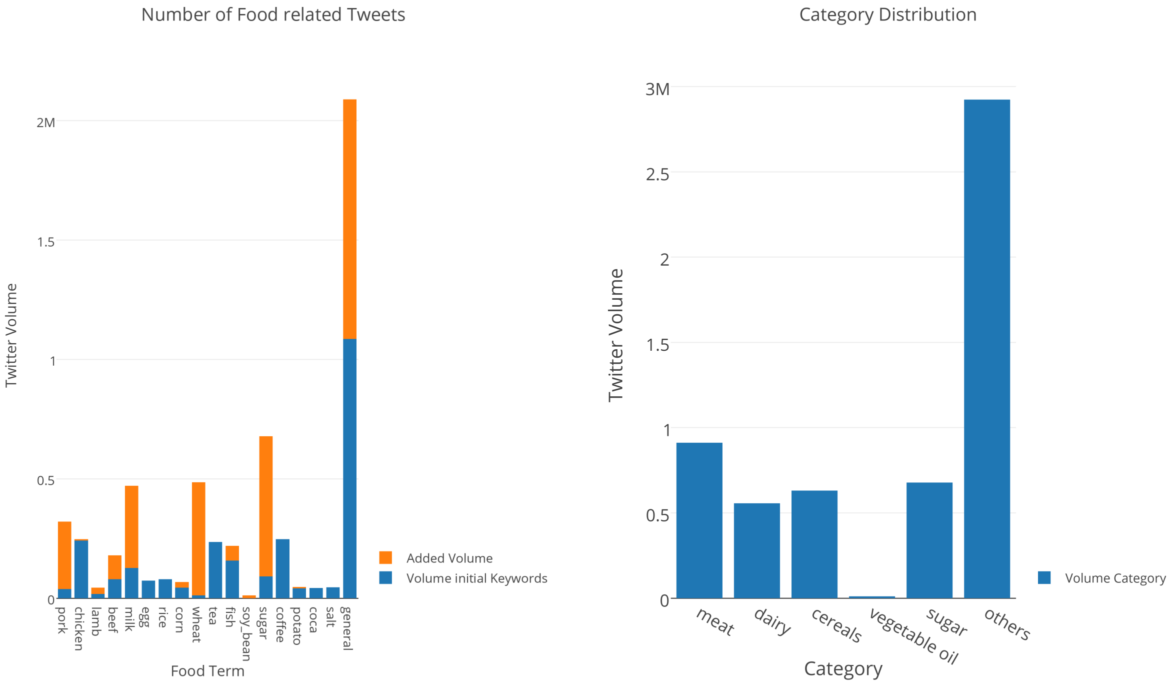


(a) Linear: Number of Tweets per User

(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.

Between the meat categories there is a strong positive linear relationship in the range of 0.9914 and 0.9980. 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. 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.

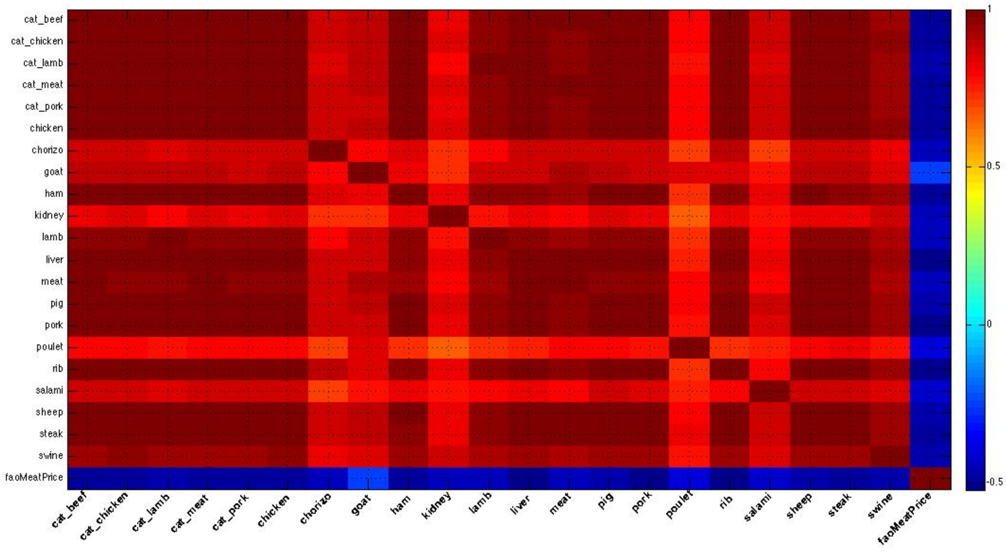


Figure 3.3: Heatplot Meat: Volume of Tweets per Keyword and per Category

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 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.

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

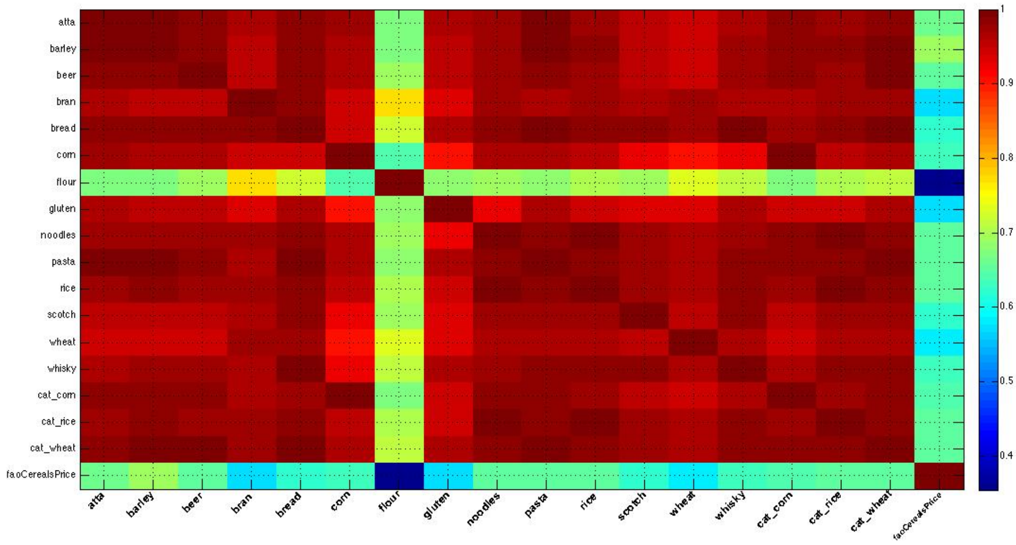


Figure 3.4: Heatplot Cereals: Volume of Tweets per Keyword and per Category

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.

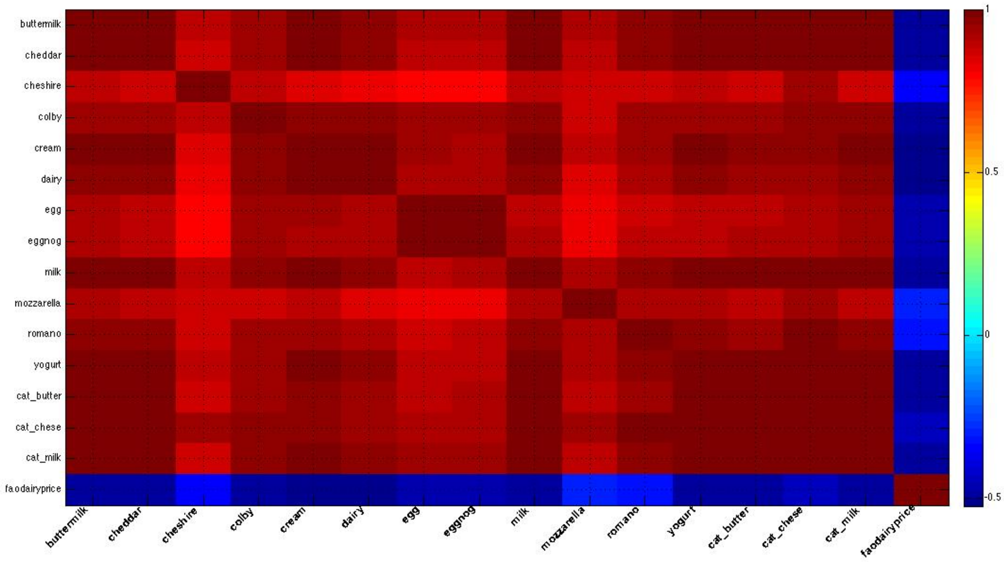


Figure 3.5: Heatplot Dairy: Volume of Tweets per Keyword and per Category

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

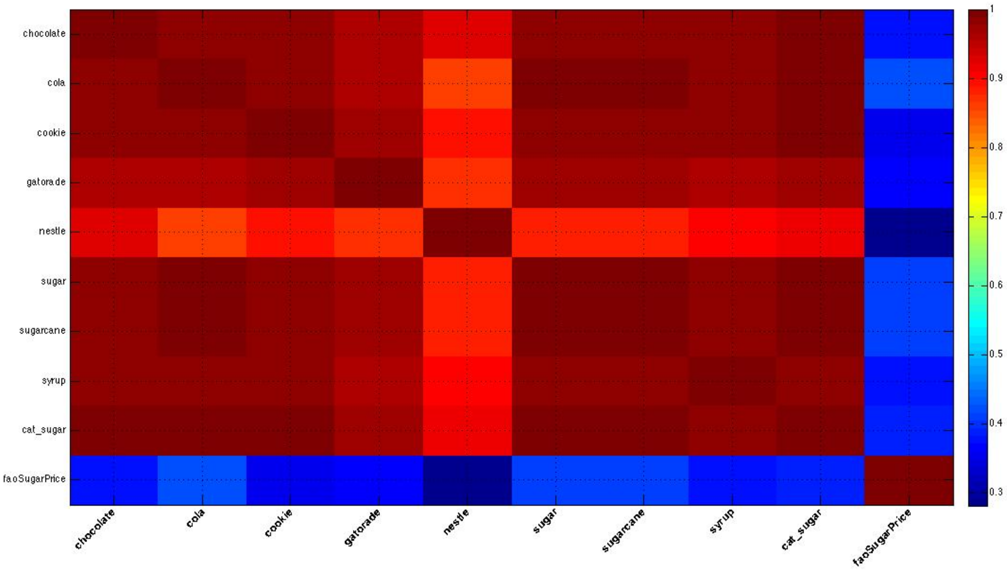


Figure 3.6: Heatplot Sugar: Volume of Tweets per Keyword and per Category

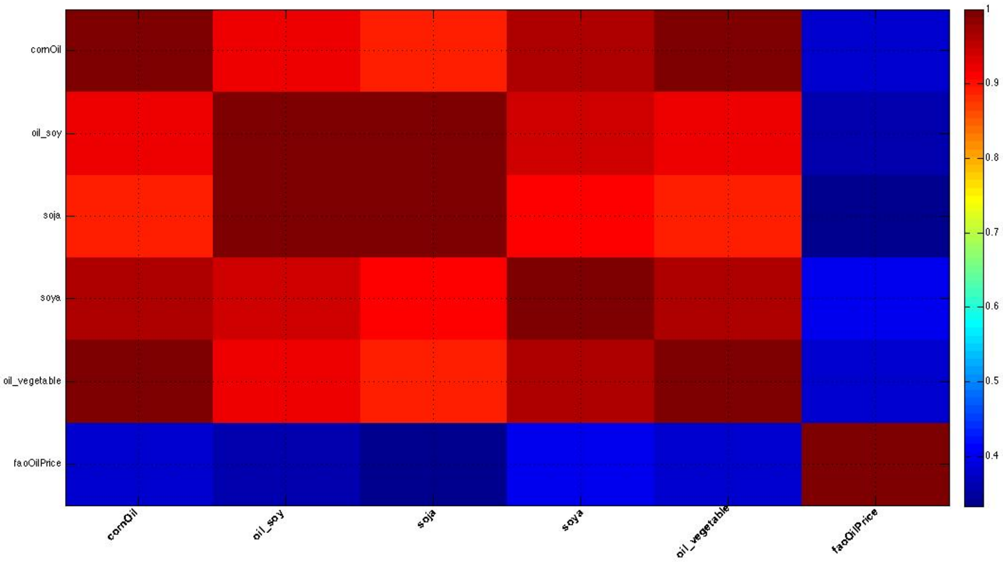


Figure 3.7: Heatplot Oil: Volume of Tweets per Keyword and per Category

3.2.1 Discussion

A smilier correlation analysis has been made in [6]. 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: $p < .0005$ ***, $p < 0.005$ **, $p < 0.05$ *

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.

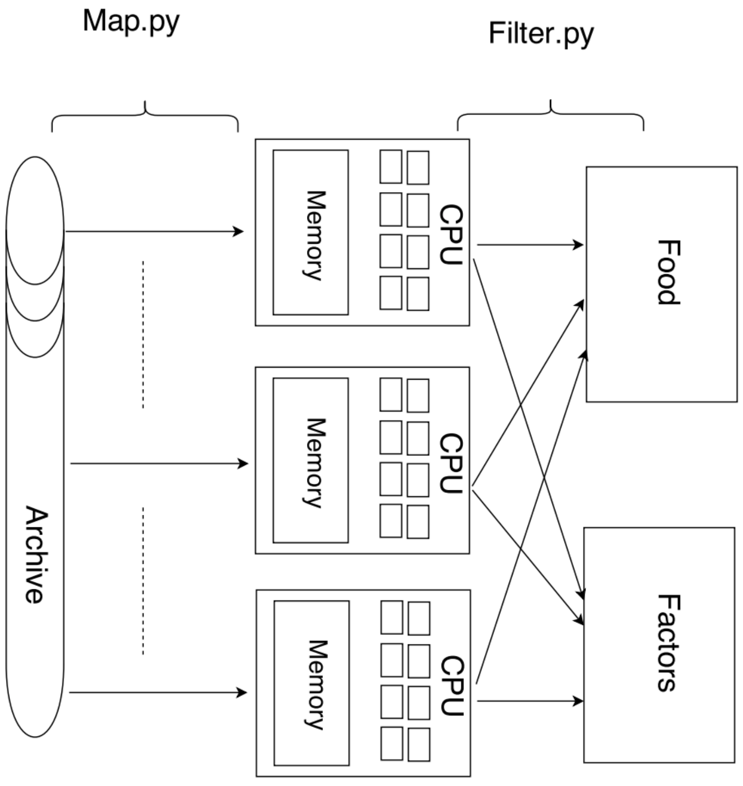


Figure A.1: Dada Processing

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] 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.

[3] 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.

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

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

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

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

[8] C. Panter-Brick. Street children, human rights and public health, 2002.

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

[10] 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.

[11] Stanford. Corenlp, 2011.

[12] 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.

[13] 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.