Food Habits of the People of Chicago
[Analysis of Organic Food Trends using Twitter Data]

Shashank Sharma (A20330372)   Sanket Nitin Wagh(A20330391)

INTRODUCTION:

The Organic food industry is a small yet up and coming industry. The choices made on the food you eat have major impacts on one’s lifestyle. More and more people seem to want organic food on the table rather than processed food. Yet the choices people make lean more towards the side of processed food. This may be because processed food comes cheap and is more abundantly available than the organic foods.

We decided to analyse the major factors that play a role in this particular industry. Everyone likes to post and share their views on the internet. We thought the microblogging site Twitter would be a good source to obtain data regarding this topic on how people state their views about food. We decided to build a classifier that predicts whether or not a tweet has an organic-food-influence or not.

Sentiment analysis helps us predict this influence. We have built a hand-labelled custom classifier. This classifier is going to mark the tweet as ‘having organic food influence’ and ‘not having organic food influence’.

Apart from this we have applied a few machine learning techniques like logistic regression and Support Vector Clustering to predict whether or not the tweet is indeed having an impact on this field.

DATA:

We used the Twitter API to pull data from the website. We performed custom searches using hastags as search criteria. The hashtag #Organic was the main tag used to extract information to help create our own training data.

The *search/tweets* filter to get data corresponding to food and food related tweets in *7 cities: Chicago, New York, Boston, Houston, Austin, Los Angeles and San Francisco .* This data was dumped into MongoDB which was used as and when required to build the classifier and perform the experiments. There were 5437 tweets downloaded 380 tweets using #Organic.

METHODS:

Pre Processing:

Once we collected the data, we began the cleaning process. The data was not raw per say, but still had material other than the pure text which is what we wanted. The first step was to finish cleaning the data obtained using #Organic tag. We removed the 'mentions' tag and URLs from our data after which tokenization was performed.

The data also contained stopwords that would hinder the performance. Stopwords are words that occur very frequently but are not important for the analysis. The next step was to tokenize our proposed training data. The function removeStopWordsperTweet(tokenizedTweets) removes the stopwords.

Sentiment Analysis:

We labelled the training data using a gradient with the range [0.0, 0.25, 0.5, 0.75, 1.0] . This is going to help classify tweets. This gradient was chosen by analysing articles published by the OTA and the USDA. The concept here is to use this gradient to determine the sentiment of the tweet; whether or not they have an 'organic' influence or not.

The gradient is applied to each word in the tweets. Taking the average weight(or grade) of the tweet and checking it with the threshold will determine the sentiment of the tweet.

Machine Learning Models:

The python package used is sklearn.

1) ***Logistic Regression:***

For logistic regression we import LogisticRegression from the sklearn.linear_model to build a model.

2) ***Support Vector Machines:***

For SVM we have implemented support vector clustering. We have imported LinearSVC from sklearn.svm to build a model.

We set a threshold value for the classification. This threshold value decides whether or not the tweet is organically influenced. Once applied along with these models we got some interesting results. At first there was overfitting with manual classification. There was some improvement once the models were used. To validate the accuracy of the model, we apply logistic regression and svc on the test day.

Cross Validation:

We have used standard K-Fold cross validation. We have set to 5 folds. Accuracy values are obtained for these folds for the training data.

EXPERIMENTS and CONCLUSIONS:

Table 1: Average accuracy values:

Threshold Value	*Average Accuracy for Logistic Regression*	*Average Accuracy for SVM*
0.5	*89.50%*	*98.16%*
0.6	*69.56%*	*92.38%*
0.7	*96.05%*	*97.63*

Table 2: Train data vs. Test data; predicted value: (Validation for trained model)

Threshold Value = 0.6	Trained Data	Test Data
Logistic Regression:	69.56%	85.63%
SVM:	92.38%	86.66%

Initially due to heavy bias of labelled data we were getting a 100% accuracy(as mentioned in our presentation). On training our model with different threshold values over a small sample of data, we were able to get conclusive results.

Table 3: N-fold cross validation. n=5.

	Logistic Regression	*SVM*
Fold 0	63.64%	*94.81%*
Fold 1	*71.05%*	*98.68%*
Fold 2	*57.89%*	*97.37%*
Fold 3	*84.21%*	*100%*
Fold 4	*71.05%*	*71.05%*

We found that there was excessively bias effect because of the small size of the dataset used while training. We have tried to follow the train, transform and classification procedures correctly as learned through the coursework. We assume that training over a larger dataset would yield optimal results.

The graphs supporting these results are in the ipython notebook submitted along with this document.

Few observations made while collecting data:

- The training data resulted in 7.61% of people tweeting about organic food.
- Based on the cities we can say New York users are most aware of the Organic Food Industry.

Figure 1:

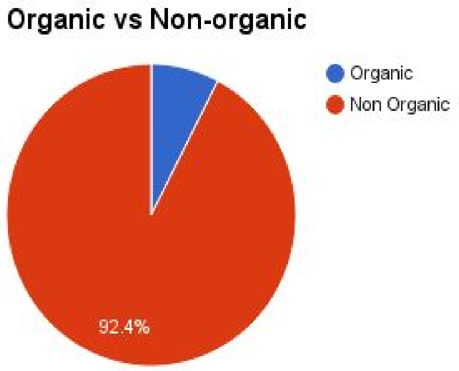
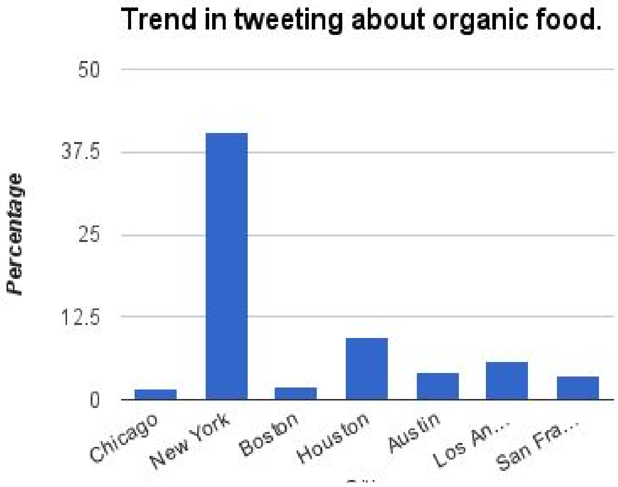


Figure 2:



RELATED WORK:

Usually analysis on this type of topic is done by simply analysing general data about food and making comparisons. Our methodology involves segregating the general data from the data pertaining exclusively to organic data. We felt using only one platform of social media, in the form of Twitter, was good enough for what we had in mind. Using the results and applying the concepts in a better way we can recommend users to use organic food by promoting it. Since a very small portion of users actually speak about this topic it may be useful to promote its importance.

FUTURE WORK:

We now know the limitations of our data collection approach. Rectifying this and collecting more data over a period of time will help us perform good experiments and build a better classifier.

Gender classification based on the amount of people of different age groups who tweet about organic food can help understand how men and women view the importance of a healthy lifestyle, based on which we can make recommendations.

We can build a recommendation engine for food suppliers like Native Foods and Trader Joe’s that market organic food. Also, we could use this information to help smaller markets grow.