




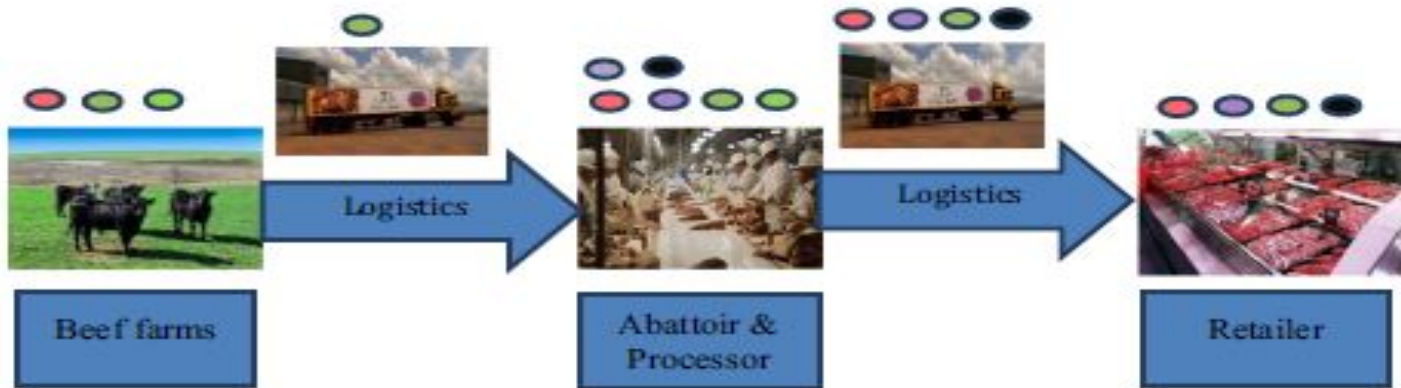
# SUPPLY CHAIN MANAGEMENT IN FOOD INDUSTRIES

# CONTENT

- 
1. Introduction
  2. Why food supply chain management?
  3. Objective
  4. Measures Taken
  5. Twitter Data Analysis Process
  6. Sentimental Analysis
  7. Hierarchical Clustering
  8. Mitigations required for improvement
  9. Conclusion
  10. References

# INTRODUCTION

A food supply chain or food system refers to the processes that describe how food from a farm ends up on our tables. The processes include production, processing, distribution and consumption.



# WHY ?


- Perishable nature of food products
- To make it more consumer centric
- Ineffective decisions made, based on smaller feedbacks of consumers
- For large audiences and hence more profit

## **OBJECTIVE**

**The objective of this study is to use text data analytics and find out how the supply chain can be improved at various stages to get a consumer -centric supply chain.**

---

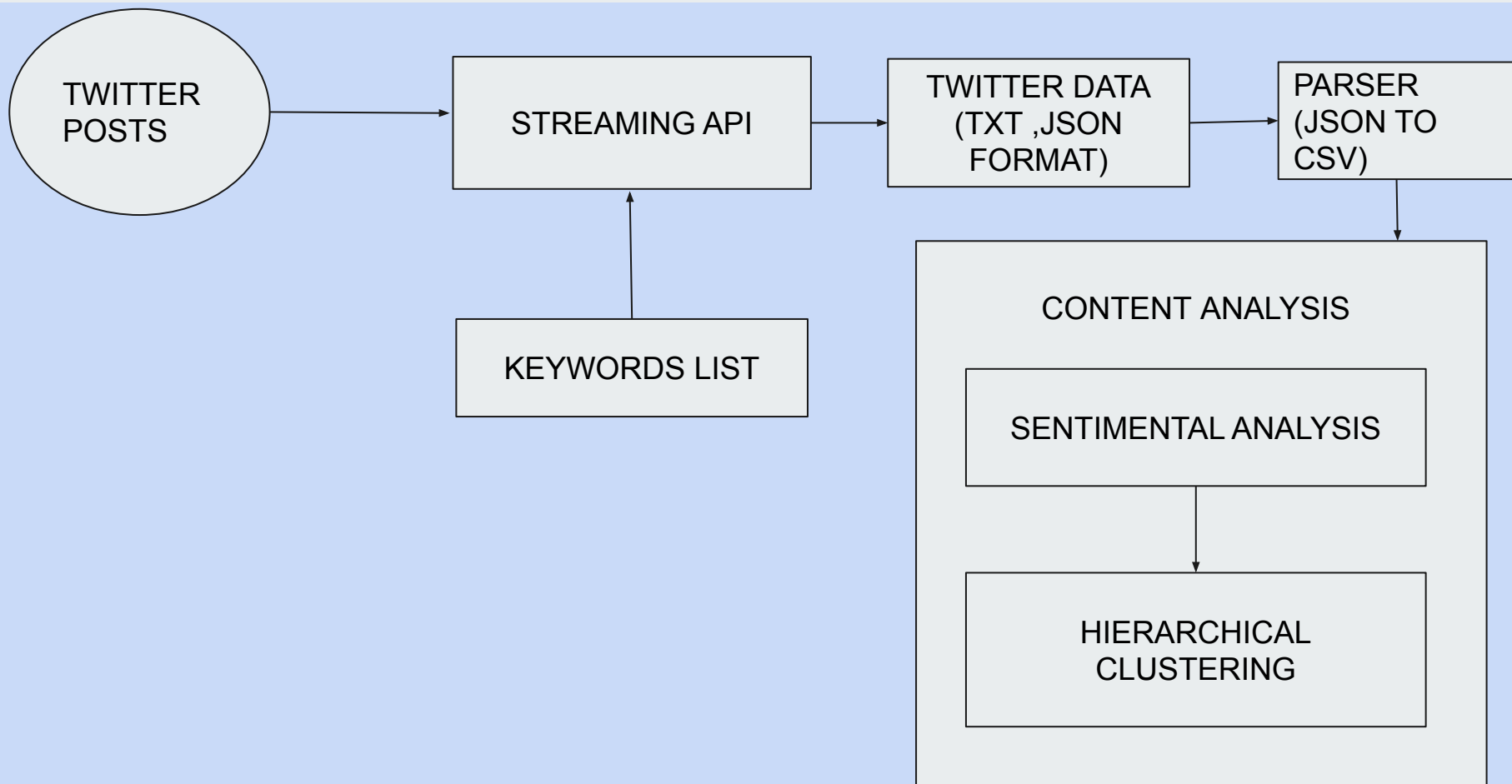
## Measures Taken

- 
- Twitter data extracted using Twitter API and R.
  - Then, sentimental analysis was done to investigate the positive and negative sentiments of tweets, using SVM.
  - Hierarchical clustering of tweets from different geographical locations using multiscale bootstrap resampling.



# TWITTER DATA ANALYSIS PROCESS

t.







# SENTIMENTAL ANALYSIS



## **Data Collection**

We have collected the Dataset from twitter streaming API using R, which gives the all tweets of last one week searched with specific keyword. Author has given the keyword which are occurred most in their data so we have used only those keyword for extraction.



## **About the Data**

The Data collected after parsing have 3032 rows (tweets) and 16 attributes but we are interested only the text i.e the tweet so we extracted the tweets from the parsed CSV.



## Preprocessing

1. Extracting the text data into new Dataframe.
2. Dropping duplicates.
3. Tokenize the tweets.
4. Removing stop words from tweets.
5. Converting them into lower case.



## Making Data Suitable for Modeling

As we have clean tweets now but for training any model using supervised learning method we need target variable also so for that we have adapted two methods.

1. Find the :) and :( emoticons in the tweets and assign +1 to :) and -1 to :( (used by Author).
2. For the rest of the data we used Textblob Library of python for finding polarity which give a number between -1 to 1 and assign sentiment +1 if polarity > 0 else assign sentiment -1.



## **Bag of Words Representation of tweet**

1. We created a Vocabulary set  $V$  based on the total all tweets which contain each unigram with their respective frequency.
2. Now, a  $|V| * 1$  sized binary vector is created for each tweet, with ones at the words it does contain, zeros at others.
3. The CountVectorizer library from sklearn has been used to create these vectors.



## Tf-Idf Representation

1. This approach is absolutely similar to binary bag of words, with the only difference being that instead of creating binary vectors of 1s and 0s, the  $|V| * 1$  sized vectors we create for each review have the tf-idf of the corresponding word.
2. We use the TfidfVectorizer library from sklearn to create these vectors.



## **About Preprocessed data**

After all the preprocessing and using Bag of Words and Tf-Idf representation the final data we have is 3332 rows and 8830 column where the last column is the assigned target variable. In the Final Dataset we have 1593 positively labelled tweets and rest are negatively labelled tweets.



## Results:

Model Name	Document Representation	Accuracy		AUC		Cross Validation	
		Train	Test	Train	Test	Train	Test
Neural Network	BOW	96.86	73.26	0.9565	0.8126	99.85	72.42
	Tf-Idf	99.14	74.14	0.999	0.8194	99.83	72.33
SVM	BOW	96.56	76.51			59.62	73.68
	Tf-Idf	59.49	57.38			99.68	58.23
Random Forest	BOW	80.76	70.56	0.8955	0.7955	99.89	70.97
	Tf-Idf	83.56	70.25	.9260	0.7669	99.42	69.64

Model Name	Document Representation	Accuracy		AUC		Cross Vaildation	
		Train	Test	Train	Test	Train	Test
Naive Bayes	BOW	97.84	67.91	0.9796	0.6748	98.21	66.49
	Tf-Idf	98.17	66.75	.9841	.6704	97.97	66.72
AdaBoost	BOW	94.31	74.27	0.9934	0.8014	76.88	68.46
	Tf-Idf	83.90	71.63	.9452	.7753	76.71	69.31

What to do  
with  
Sentiments  
???

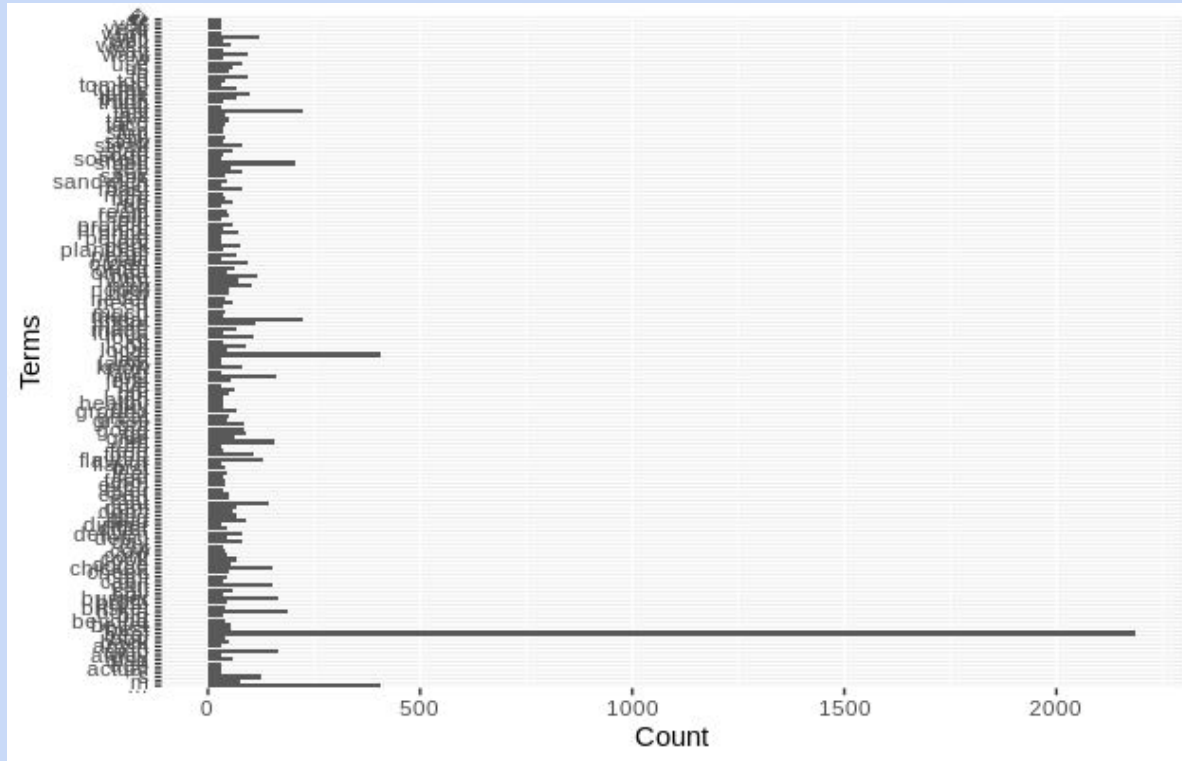


Clustering





# Why ?





# How ?

- Clustering has been done on positive tweets , negative tweets and all tweets separately .
- We have used pvclust package in R for clustering
- Method used for clustering : Hierarchical Clustering with Multi Scale Bootstrap Resampling.

# Multi Scale Bootstrap Resampling



- Thousands of bootstrap samples are generated by randomly sampling elements of the data, and bootstrap replicates of the dendrogram are obtained by repeatedly applying the cluster analysis to them
- The bootstrap probability (BP) value of a cluster is the frequency that it appears in the bootstrap replicates.



# Multi Scale Bootstrap Resampling



- The multiscale bootstrap resampling was developed (Efron et al., 1996; Shimodaira, 2002, 2004) for calculating approximately unbiased (AU) probability values.
- In the multiscale bootstrap resampling, we intentionally alter the data size of bootstrap samples to several values. Let  $n$  be the original data size, and 'nb' be that for bootstrap samples.

# Multi Scale Bootstrap Resampling

- For each cluster, an observed BP value is obtained for each value of  $n'$ , and we look at change in  $z = N^{-1}(\text{BP})$  values, where  $N^{-1}(\cdot)$  is the inverse function of  $N(\cdot)$ , the standard normal distribution function.
- Then, a theoretical curve  $z(n') = v \sqrt{n'/n} + c \sqrt{n'/n}$  is fitted to the observed values, and the coefficients  $v, c$  are estimated for each cluster.
- $AU = N(-v+c)$

## Procedure :




- First we divided the tweets on the basis of negative and positive sentiments .
- Then extracted text from all tweets .
- Created Corpus


## Procedure :



- Little bit of preprocessing.
- Created Term-to-Document matrix.
- Calculated distance matrix , which contains distance of each words from another words .
- Applied Clustering Methods.

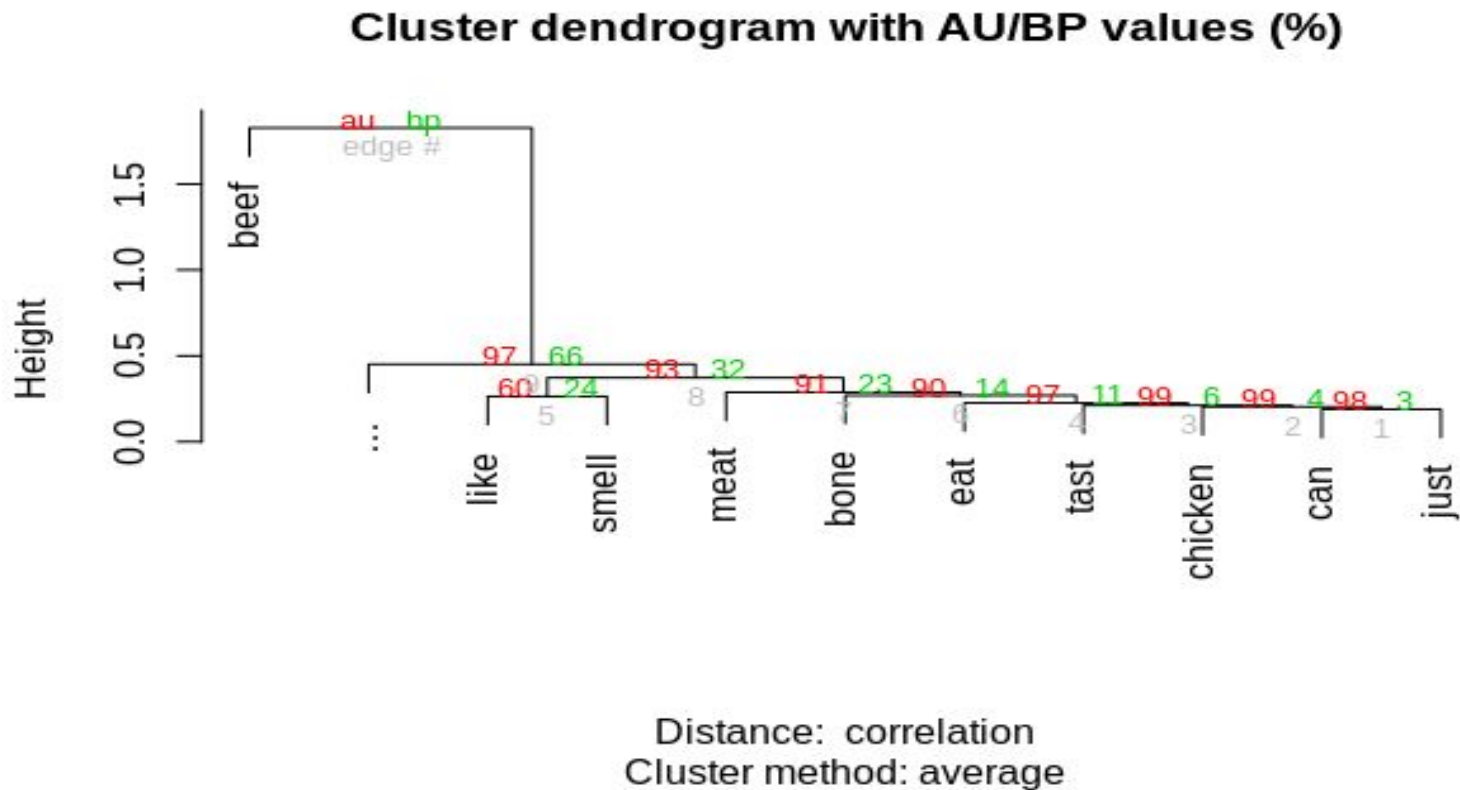


During resampling, the replicating of sample sizes was changed to multiple values including smaller, larger, and equal to the original sample size. Then, bootstrap probabilities are determined by counting the number of dendrograms which contain a particular cluster and by dividing it by the number of bootstrap samples.

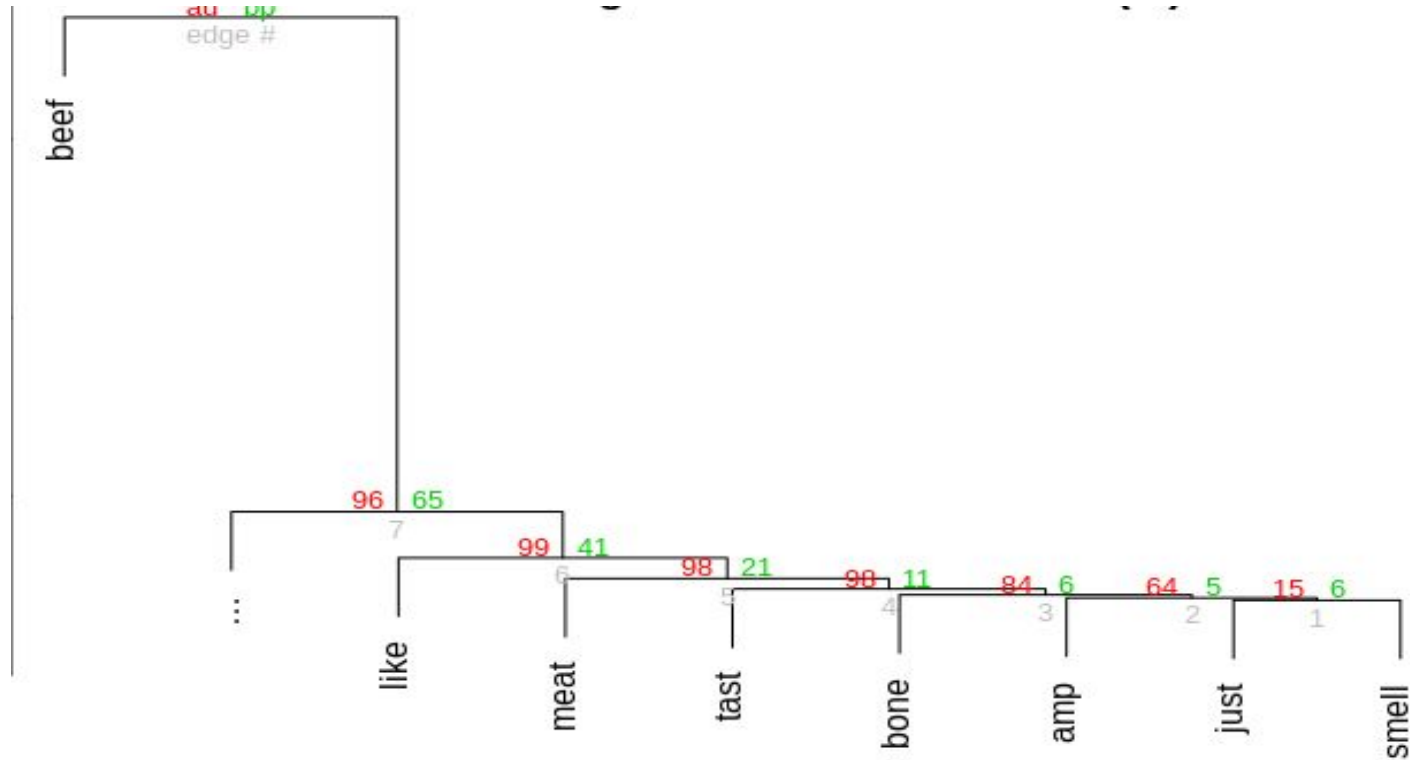


This procedure is performed for all the clusters and sample sizes. Then, these bootstrap probabilities are used for the estimation of the p-value, which is also known as approximately unbiased (AU) value.

# Dendrogram for Negative tweets :



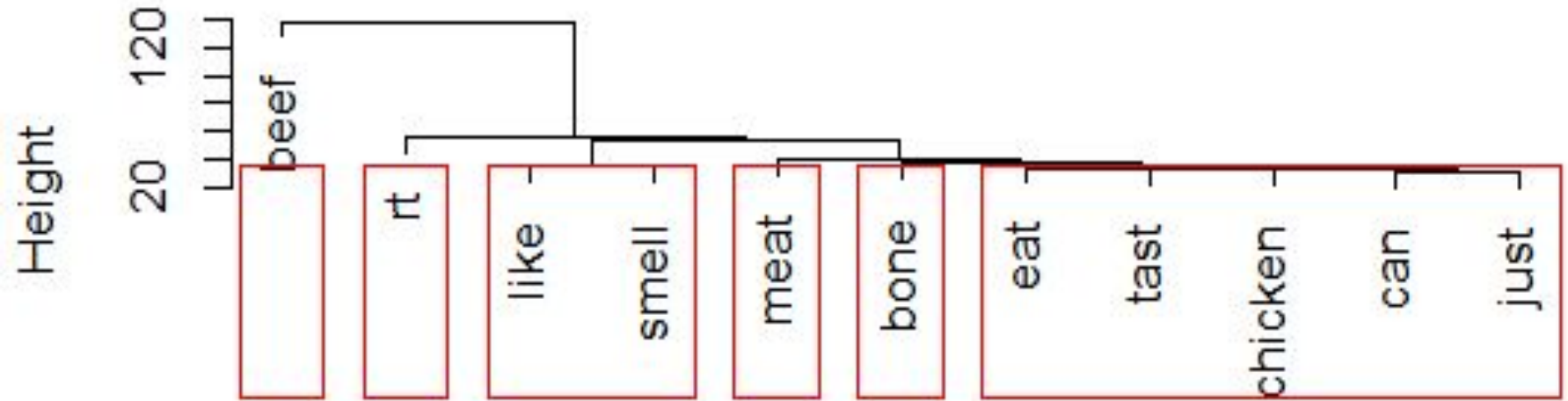
# Dendrogram for All tweets :





# INFERENCE FROM CLUSTERS

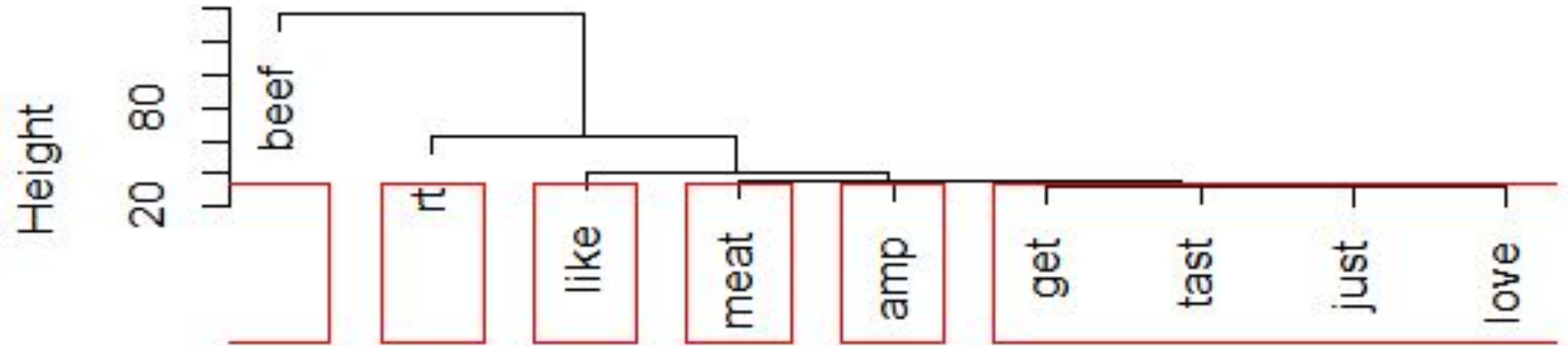
## Cluster Dendrogram (NegativeTweets)



distMatrix  
hclust (\*, "ward.D")

## Cluster Dendrogram


(Positive Tweets)




distMatrix  
hclust (\*, "ward.D")



# Mitigation Of Issues

- 
1. Periodic maintenance of packaging machines at abattoir and processor
  2. Supply Chain Mapping
  3. Raising of cattle as per the weight and conformation specifications of retailer
  4. Maintaining cold chain management
  5. Following renowned food safety process management techniques

# Conclusion

- 
- Both positive and negative sentiments related to a particular product are crucial components for the development of a consumer -centric supply chain.

# References



- [https://hwpi.harvard.edu/files/chge/files/lesson\\_4\\_1.pdf](https://hwpi.harvard.edu/files/chge/files/lesson_4_1.pdf)
- <https://academic.oup.com/bioinformatics/article/22/12/1540/207339>
- <https://projecteuclid.org/euclid.aos/1107794881>