Assignment 8: Case Study

Text Mining Tools for Analysing Twitter

Group No: 5

Madhulika Sharma & Janiece Pandya

★ Introduction

Contribution by social media these days is enormous towards the growth of the data written by each individual who share it across the web. It presents different opinions, believes, attractions, preferences and more.

However, data from users does not occupy storage space without a purpose but it is very important for many people and organizations including firms, manufacturers, politicians, policy makers, marketers, intelligence agencies and even armed forces once the useful facts for them are extracted from the data.

There are over 554 million users and around 58 million tweets everyday. Twitter has become important source of information at a place for most of the topics people are looking for.

★ Objective

The main objective of this project is to gather Tweets with a single hashtag , we have used "#Vegan"and then identifying the opinions about that people about vegan food, whether it is positive, negative, neutral or maybe irrelevant of Tweets is identified depending on the user's view that is expressed in those Tweets. However, Tweets are restricted to the length of 140 characters which might make the analysis harder and motivates users to produce shortened forms of words for that the content of Twitter is considered as a noisy content.

Following are the steps we performed for discovering and extracting the information from unstructured data:

★ Information Retrieval (IR)

The process of gathering related text about a particular subject in order to be analysed afterward.

IR systems can be seen everywhere e.g. search engine, library catalogues, cookbook indices, and so on. The main objective of any IR system is to retrieve whatever is useful and to ignore whatever that is not. IR family can be summarised through figure. Traditional IR systems retrieve the information from unstructured texts (raw text) such as documents, comments, etc.,

while the advance IR systems extract them from structured texts (marked up/ tagged) texts such as XML1 documents.

```
# CLEANING TWEETS
tweets.df$text=gsub("&amp", "", tweets.df$text)
tweets.df$text = gsub("&amp", "", tweets.df$text)
tweets.df$text = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tweets.df$text)
tweets.df$text = gsub("@\\w+", "", tweets.df$text)
tweets.df$text = gsub("[[:punct:]]", "", tweets.df$text)
tweets.df$text = gsub("[[:digit:]]", "", tweets.df$text)
tweets.df$text = gsub("http\\w+", "", tweets.df$text)
tweets.df$text = gsub("[ \t]{2,}", "", tweets.df$text)
tweets.df$text = gsub("^\\s+|\\s+$", "", tweets.df$text)
tweets.df$text <- iconv(tweets.df$text, "UTF-8", "ASCII", sub="")
```

★ Natural Language Processing (NLP)

The process of processing human natural language in order to be understood and analysed by a computer. This is one of the hardest artificial intelligence problems.
Natural Language Processing refers to the practices performed on sentences written in a natural language such as the English language.

	Finding answers and information that already exist in a system		Creating answers and new information by analysis and inference – based on query
	Search by navigation (following links, as in a subject directory and the Web generally)	Search by query (as in Google)	
Unstructured information (text, images, sound)	**Hypermedia systems** (Many small units, such as paragraphs and single images, tied together by links)	**IR systems** (Often dealing with whole documents, such as books and journal articles)	
Structured information		**Database management systems (DBMS)**	**Data analysis systems Expert systems**

Figure (3): *The IR system family*.

	RDB search	**Unstructured retrieval**	**Structured retrieval**
Objects	Records	Unstructured documents	Trees with text at leaves
Model	Relational model	Vector space and others	?
Main data structure	Table	Inverted index	?
Queries	SQL	Free text queries	?

Figure (4): *Comparison between IR methodologies and traditional relational database search.*

It includes many subtopics such as:

- Signal processing: dealing with spoken language (out of scope).
- Syntactic analysis: dealing with how the sentences are written and their grammars.
- Semantic analysis: dealing with meanings of words within the sentence.
- Pragmatics: deals with the correlation between the meaning of a sentence and its daily usage.

Topic	Explanation
Phonetic and phonological knowledge	How words relate to their sounds.
Morphological knowledge	How words are built from more primitive morphemes e.g. how "sunny" comes from "sun".
Syntactic knowledge	How a sequence of words makes a **correct** sentence. (Knowledge of the **grammar rules**)
Semantic knowledge	How words have meaning. How words have denotations (references) and connotations (associated concepts).
Pragmatic knowledge	How sentences are used in different situations and how the different usage can affect the meaning of the sentence. Involves intentions and context.
Discourse knowledge	How previous sentences can affect the meaning of a sentence. When referencing to a pronoun.
World knowledge	General knowledge such as knowledge about others involving in the conversation

Figure(6): *Knowledge Relevant to Natural Language Understanding*.

★ Information Extraction (IE)

The process of articulating entities, events, facts and the connections between them. After that, they are extracted and can be processes later by the next step different techniques.

Information extraction is done by extracting structured data from unstructured or semistructured machine-readable documents which was traditionally relying heavily on human involvement. Structured data that can be extracted includes:

• Named Entities. For example, persons, locations, organisations, products, etc. They can be extracted using Named Entity Recognition (NER) Techniques.

• General Entities. They are hard to disambiguate and in order to be identified, contextual information is required e.g. knowledge bases or encyclopedias.

• Characteristics and Attributes of Entities. They can be gathered using entity-centred web search (object level vertical search). An example of this is Microsoft's EntityCube project which summarises information about the entities regardless of their presence level on the web.

★ Data Mining (DM)

The process of finding various relationships between pieces of information extracted from different data sources such as databases, collections of documents, blogs, webpages and/or various combinations of them.

There are many techniques used in DM, here is some of the commonly used techniques:

a) Artificial neural networks: structured as biological neural networks to learn in a non-linear predictive way.
b) Decision trees: structured as trees that resembles different sets of decisions. These decisions produce the rules that are used for classification. For Instance, Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
c) Genetic algorithms: uses some processes for optimisation such as genetic combination,natural selection and mutation. It is designed based on evolution concepts.
d) Nearest neighbour method: classifies each element in the dataset based on combination of the classes of the kth element(s) which is the nearest to it. Also, it is called the k-nearest neighbour technique.
e) Rule induction: using statistics to extract the useful if-then relationships.

★ Text Mining with Twitter

Using TM approaches with twitter introduces many challenges and problems and requires usage of already available libraries. Using what is already available helps in advancing the available cutting-edge technologies and adding extra features once they are needed.

★ Design Overview

This project's goal is to build a system that can retrieve Tweets from twitter, filter them, normalise them, give them polarity scores and visualize the produced results for the user. Activity diagram illustrates the flow of the application

★ Implementation Technologies

This project was implemented using R. The following are the steps that are needed:

- **Twitter API:** To use Twitter API in R, we use library called “twitteR”, we need an authorization request which we get by creating an application in our twitter account which gives us Consumer API keys and access tokens and secrets which we need to get access to twitter.

```
library("tm")
library("twitteR")
library("httr")
library("RCurl")
library("syuzhet")
library("RColorBrewer")
library("wordcloud")

consumer_key <- 'bjeLK6wxISeNFGbN2sPSKhUEV'
consumer_secret <- 'hZ7CgfgOt1WOcqlEYebGsnolqve1ayg6GgrWOujKFKKMYeJ7DE'
access_token <- '1129554063799083008-silDDpNIn11Boo0DiqFPzojtSFZiXv'
access_secret <- 'fM7u3j9fPlVB1pVZ2sBtXbSOqZPvEdRFyWs0rFroKWOdy'

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

- Extract tweets for the sentimental analysis. Here we have used a single hashtag #Vegan and extracted recent 5000 tweets.

```
# extract tweets
tweets = searchTwitter("#vegan", lang="en", n=5000, resultType="recent")
```

- Store tweets into a data-frame for further use.

```
# store the tweets into dataframe
tweets.df = twListToDF(tweets)
```

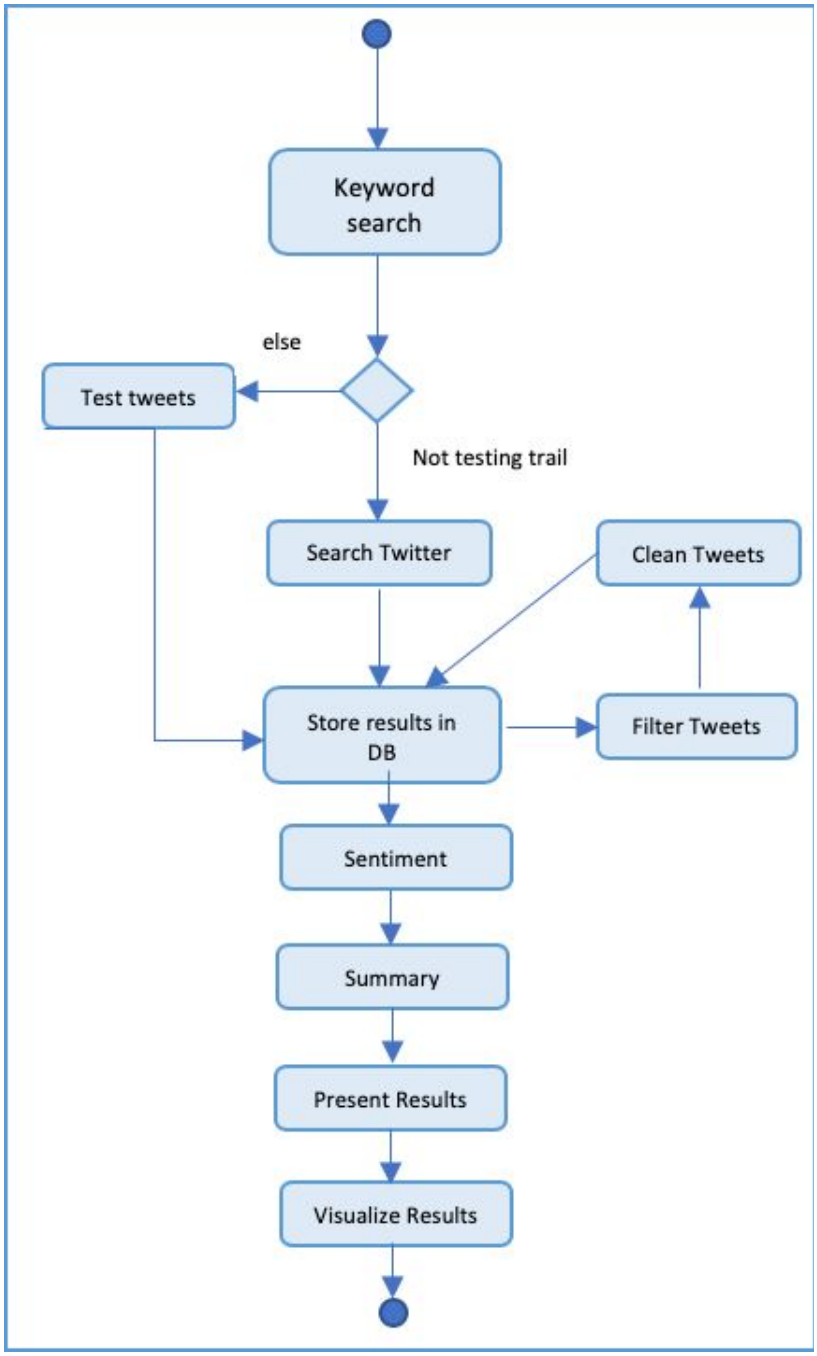
- Cleaning Tweets by removing whitespaces, numbers, url, punctuations and other irrelevant things in our extracted tweets.

```
# cleaning tweets
tweets.df$text=gsub("&amp", "", tweets.df$text)
tweets.df$text = gsub("&amp", "", tweets.df$text)
tweets.df$text = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tweets.df$text)
tweets.df$text = gsub("@\\w+", "", tweets.df$text)
tweets.df$text = gsub("[[:punct:]]", "", tweets.df$text)
tweets.df$text = gsub("[[:digit:]]", "", tweets.df$text)
tweets.df$text = gsub("http\\w+", "", tweets.df$text)
tweets.df$text = gsub("[ \t]{2,}", "", tweets.df$text)
tweets.df$text = gsub("^\\s+|\\s+$", "", tweets.df$text)
tweets.df$text <- iconv(tweets.df$text, "UTF-8", "ASCII", sub="")
```

- Dividing cleaned tweets into emotions using NRC dictionary to calculate the presence of ten different emotions and their corresponding valence in a text file. The ten columns are as follows: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative", "positive".

```
# Emotions for each tweet using NRC dictionary
emotions <- get_nrc_sentiment(tweets.df$text)
emo_bar = colSums(emotions)
emo_sum = data.frame(count=emo_bar, emotion=names(emo_bar))
emo_sum$emotion = factor(emo_sum$emotion, levels=emo_sum$emotion[order(emo_sum$count, decreasing = TRUE)])
```

The following is the flow chart of the Sentimental Analysis using Twitter:



- Lexicon : Stores 11816 different terms along with their sentiments.
- Modifier : Stores modifiers that change the sentiments of the words in lexicon.
- Negation : Stores negation words which reverse the sentiment of a word in lexicon.
- Abbreviation : Stores some abbreviations along with their full form.

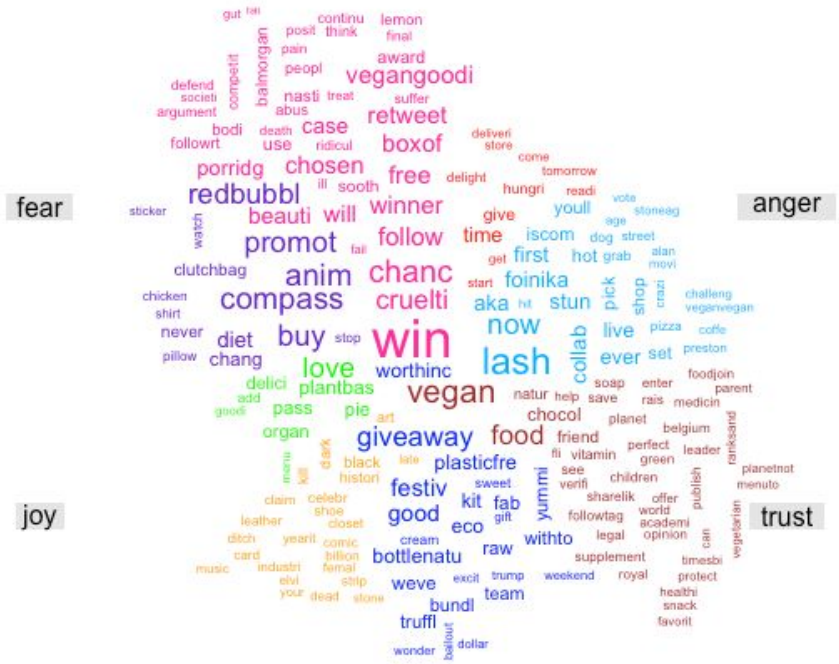
★ Generation of Results

Tweets are stored for the user as soon as they are retrieved from Twitter. After that, all Tweets are normalised and each Tweet that is expanded because of the presence of abbreviation will be printed for the user. Final number of Tweets after filtering i.e. after removing spams, duplicates and non-English Tweets. Each sentiment analysis method mentioned above will classify Tweets into positive, negative and possibly neutral classes. Then percentage of each class is calculated for each method and printed at the end. Also, each Tweet is evaluated using the 4 methods and the resulting sentiments are viewed for the user.

★ Visualisation

Resulting classes from classifying Tweets using NRC dictionary will be visualised using WordClouds such that each class will be represented by an individual WordCloud represented by a separate tap in a window so the user can move between each tap and see the differences.

```
# column name binding
colnames(tdm) = c('anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust')
colnames(tdmnew) <- colnames(tdm)
wordcloud(tdmnew, random.order=F,max.words=80, col=rainbow(50), scale=c(2,0.5))
comparison.cloud(tdmnew, random.order=FALSE,
          colors = c("#00B2FF", "red", "#FF0099", "#6600CC", "green", "orange", "blue",
"brown"),
          title.size=1, max.words=200, scale=c(2, 0.4),rot.per=0.1)
```



★ Results:

Twitter Search

Keywords entered by the user are used to make the query which then used to perform a search on Twitter. While Tweets are being pulled from Twitter, they will be printed to the user along with the user who posted them and the time they were created at.

Sentiment Evaluation

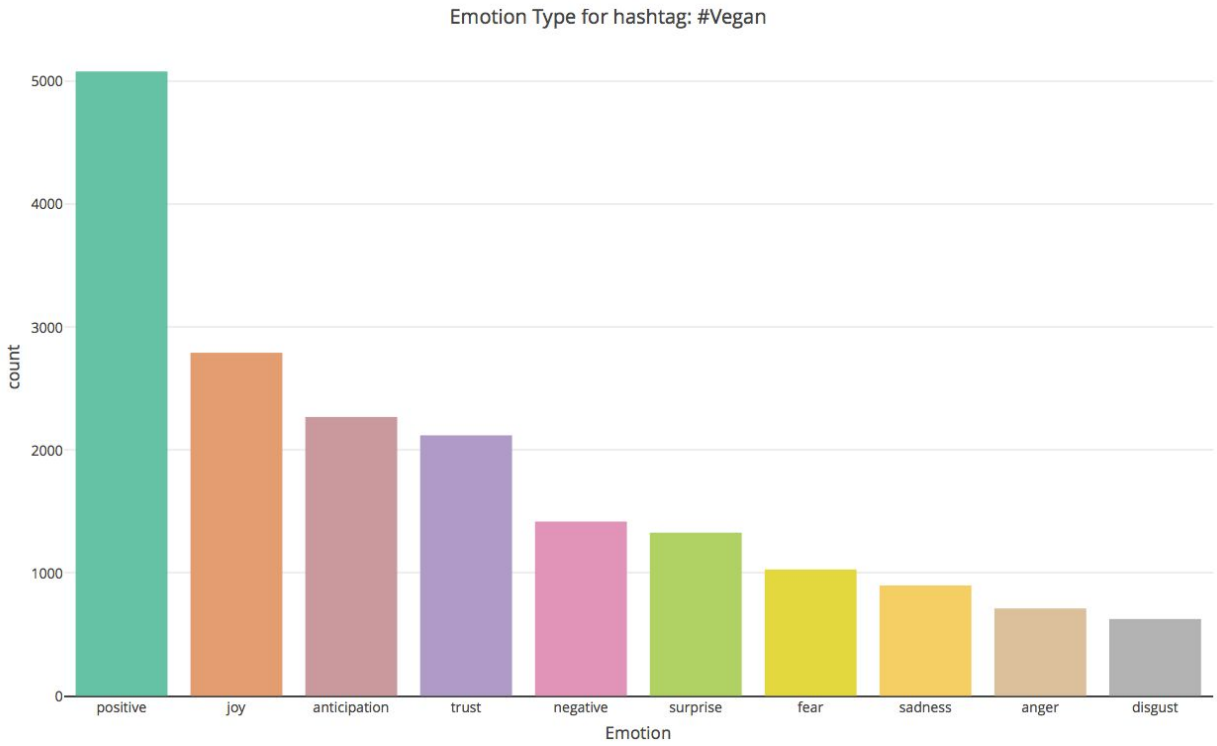
Before performing Sentiment analysis on Tweets, Classification and Stanford pipeline methods will construct the required models i.e. NRC dictionary. Sentiment of each Tweet will be evaluated using the 4 methods discussed above and the result from each evaluation method will be shown to the user. Sentimental analysis can be performed by:

- ❖ Extracting tweets using Twitter application
- ❖ Cleaning the tweets for further analysis.
- ❖ Getting sentiment score for each tweet.
- ❖ Segregating neutral, positive and negative tweets.

Figure below scores the emotions on each tweet as syuzhet breaks emotion into 10 different categories.

```
# Emotions for each tweet using NRC dictionary
emotions <- get_nrc_sentiment(tweets.df$text)
emo_bar = colSums(emotions)
emo_sum = data.frame(count=emo_bar, emotion=names(emo_bar))
emo_sum$emotion = factor(emo_sum$emotion,
levels=emo_sum$emotion[order(emo_sum$count, decreasing = TRUE)])

install.packages("plotly")
library(plotly)
plot_ly(emo_sum, x=~emotion, y=~count, type="bar", color=~emotion) %>%
layout(xaxis=list(title="Emotion"), showlegend=FALSE,title="Emotion Type for hashtag:
#Vegan")
```

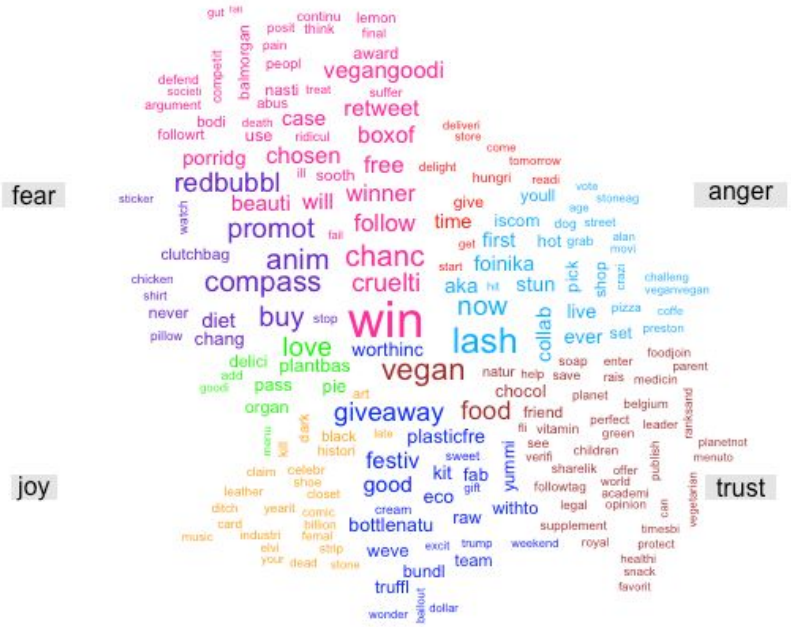


Results Summary

Statistics about the analysed Tweets are viewed to the user after sentiment analysis is done. Tweets are classified into 2 or 3 classes using each sentiment analysis method. After that, the percentage of Tweets of each class is calculated and presented to the user. An example of the summary results is shown

Lastly, lets see which word contributes which emotion:

Example:



Calculation of all the emotions for all the tweets:

	tweets.df2	anger	anticipation	disgust	fear	joy	sadnes
19...	zero sugar + organic ingre...	0	0	0	0	1	
48...	zenberry power packs are ...	0	1	0	0	1	
445	yummy vegan smoothie bo...	0	0	0	0	0	
19...	yummy burger	0	0	0	0	0	
12...	yummy breakfast! homem...	0	0	0	0	0	
42...	yum!	0	0	0	0	0	
41...	youtube gold award. we g...	0	1	0	0	1	
44...	your pupils will encounter ...	0	0	0	0	0	
47...	your lips new obsession, b...	0	0	0	0	1	
27...	your daily skincare regime...	0	2	0	0	1	
47...	you're a	0	0	0	0	0	

Creating Sentiment Value for each emotion per tweet :

```
> sent.value
   [1]  2.85  0.00  1.50  0.40  1.00  1.40  1.20  1.40 -0.25  1.00  1.00
  [12]  1.20  0.75  1.00  0.85  0.80  1.10  2.25  0.90  0.40  1.30  0.80
  [23]  0.00  0.10  0.00  1.00  1.20 -1.00 -0.10  2.15  0.00  2.75  2.75
  [34]  0.50  0.50  0.50  0.50  0.50  0.50  0.50  0.50  0.50  0.50  0.50
  [45]  0.50  0.50  0.50  0.50  0.50  2.15 -0.20  0.50  1.30  1.50  1.65
  [56] -1.20  0.80  2.90  1.30  1.25 -0.20  1.30  2.05  1.50  0.00  1.40
  [67]  1.25  0.10  0.85  2.30  0.00  2.60  0.80  2.10 -1.60  1.55  2.75
  [78]  0.60  1.50  0.10 -0.15  0.75  0.15  1.75  1.75  1.65  1.15  0.90
  [89]  1.40  1.10  0.75  1.35  1.00  3.00  0.50  0.75 -2.10  0.00  0.75
 [100]  0.05 -0.50  0.60  0.75 -0.50  1.75  1.10  0.75  0.80  0.75  0.35
 [111]  1.10 -2.10  0.50  2.10  0.50  0.80  2.25  0.50  1.20  2.60  0.15
```

Positive Tweets :

1	logging out.. i think my car might have turned into a p...
2	big fan of pickles? then you'll love these
3	the latest the mindfulness post!
4	57 easy
5	keeping the trend going at
6	🌱💪 cauliflower, beetroot & asparagus salad w ...

Most Positive Tweet :

```
> most.positive
[1] "these beautiful little feet need extra care 😉 learn about our gentle and cal
ming mogi mousse® baby butter + baby w… "
```

Negative Tweets :

1	all this drama going on in the
2	mice: the biggest losers with vegetarianism
3	fiber is killing your gains! you think i'm crazy:
4	we're hungry for change,
5	we're so hungry,
6	dinner time is here [illegible] don't forget to fuel up before t...
7	on sunday we will be serving miso with tofu and carrot...
8	meet the vegan-aires: alt-meat frenzy boosts tech, fo...
9	introducing our cbd cream! it may help with pain caus...
10	i was recently talking about how badly i wanted to find a
11	living candida free: conquer hidden epidemic that's ma...
12	ruthless man: you begin by slaying the animal and the...

Most Negative Tweet :

```
> most.negative
[1] "thinking animal abuse is wrong but being so adamant that veganism is stupid i
s a mind fuq "
```

Neutral Tweets :

3230	ciabatta rolls - baking improvised -
1970	close up of my rainbow cakes that are about to be iced...
2555	find us today in derby for veganmarketsuk. we're next ...
1051	getting serious about plant-based eating
1010	if you're thinking about trying a
1222	just pinned to soup recipes: autumnal root vegetable s...
2213	our tea lights are made using only 100% eco soya wax ...
1508	rabbit: the cyberpunk 2077 gal crossbody bag
3346	thinking about going
2652	this week it's all about the
138	vegan icecream bonabonaicecream at

Alternatively creating a table for all the total number of positive, negative and neutral tweets:

```
#Alternate Option
category_senti <- ifelse(sent.value < 0, "Negative", ifelse(sent.value > 0, "Positive", "Neutral"))
head(category_senti)

category_senti2 <- cbind(tweets.df2, category_senti, sent.value)
head(category_senti2)

table(category_senti)
```

```
> table(category_senti)
category_senti
Negative  Neutral Positive
     235     3638     1127
```

★ Conclusion:

From this it can be concluded that most of the people upload neutral tweets for #Vegan. Sentiment analysis of tweets or social media posts can help companies better analyze customer feedback and opinion, and better position their strategy. Businesses are trying to unlock the hidden value of text in order to understand their customers' opinions and needs and make better, more informed, business decisions. Traditionally businesses relied on surveys, workshops and focus groups to gain insight into their customers opinions and feelings, but today with modern technology we are able to harness the power of Machine Learning and Artificial Intelligence using NLP to extract meaning from text and dive into opinions of customers and see them outside of the often controlled environment of a survey.