Web and Social Media Analytics

Social media (sentiment analysis I)

Dr Philip Worrall

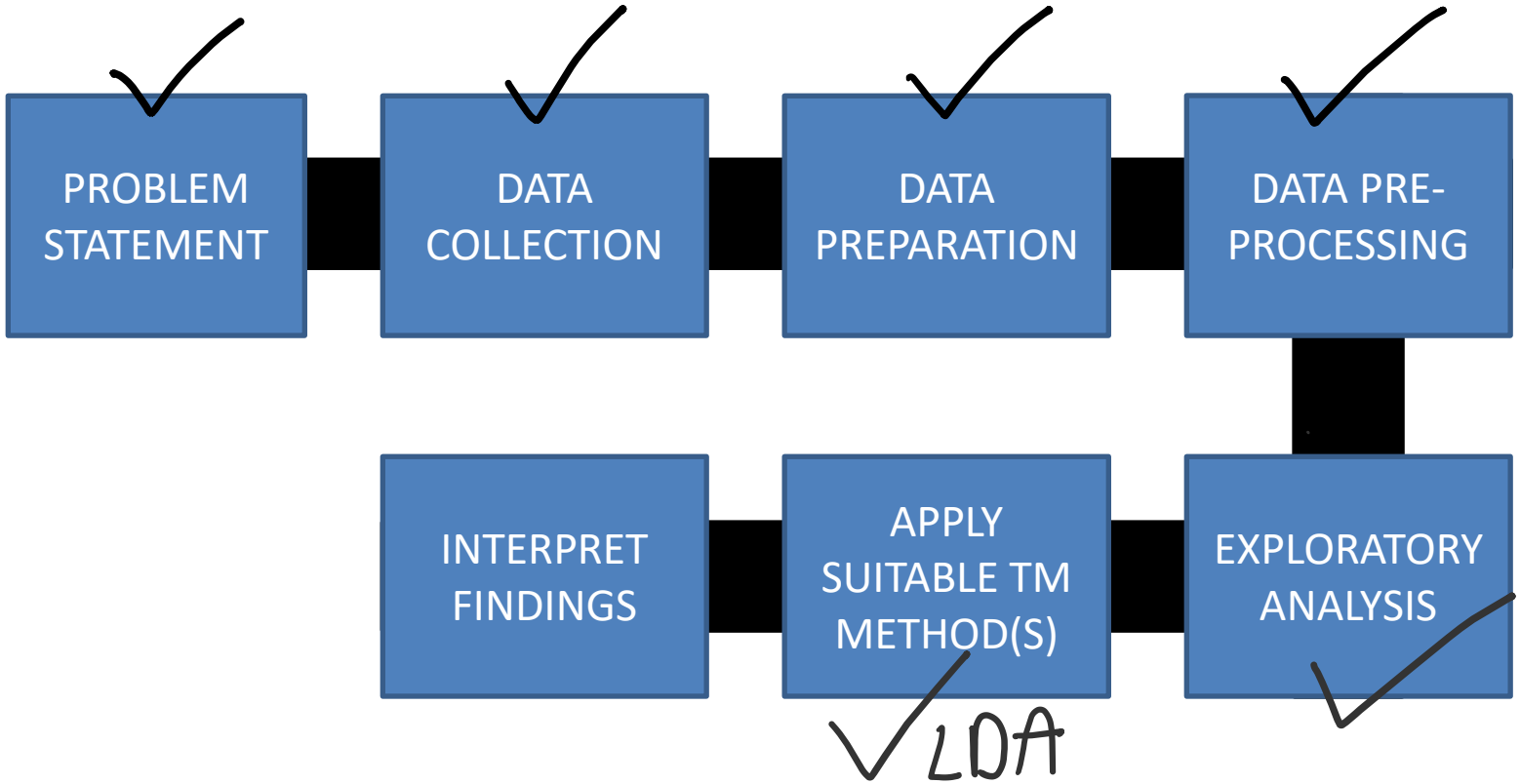
School of Computer Science and Engineering
115 New Cavendish Street
University of Westminster
London, W1W 6UW
worralph@westminster.ac.uk

LW11

Outline

- Recap of last week's material
- Sentiment analysis
  - Lexicons
  - Polarity
  - Valence
  - Context-aware
  - NGrams
  - VADER model

The text mining process…



Sentiment analysis.

Sentiment analysis…

- Specialist field within NL processing.
- Also known as "opinion mining".
- Infer views, attitudes, feelings, emotions expressed.
- Can be applied across a range of commercial and non-commercial contexts.
  - Elections
  - Brand image and PR
  - 1st class flight experience
  - Movie reviews

Not limited to just SM

Traditional approaches…

- Polls and Surveys (YouGov, ComRes)
- Telemarketing
- Questionnaires
- Delphi/Expert panels

Drawbacks

cost? time lag? response bias?
leading the witness?

A sentiment classifier…

Three basic sentiments:

- Positive
- Negative
- Neutral

"It is a very cold day"

BoW

1,0	1,1	1,2	class
very	cold	day	?

Examples…

Ex 1

1,0	1,1	1,2	class
movie	not	terrible	?

Ex 2

1,0	1,1	1,2	1,3	class
fun	expensive	try	again	?

Lexicons…

Sentiment Lexicon

A word or phrase that is labeled according to its sentiment orientation

e.g., Beautiful → Positive, Tasteless → Negative

✓ ✗

Many of the standard models for sentiment analysis work based on the presence of one or more sentiment lexicons

discovered, ride, explained

Lexicon orientation…

- Language dictionaries (Mondly, 2023)
  - English (~1.4-million-word definitions)
  - Turkish (~150k)
  - Italian (~270-400k)
- Standardised subsets of sentiment lexicons (Harvard, Linguistic Inquiry and Word Count [LIWC])
- ”I really enjoyed the football match at the weekend despite the ticket being so expensive”
- Enjoyed is Positive, Expensive is Negative
- Overall Sentiment Score: 0.5
- NP: football match ticket

Polarity Lexicons…

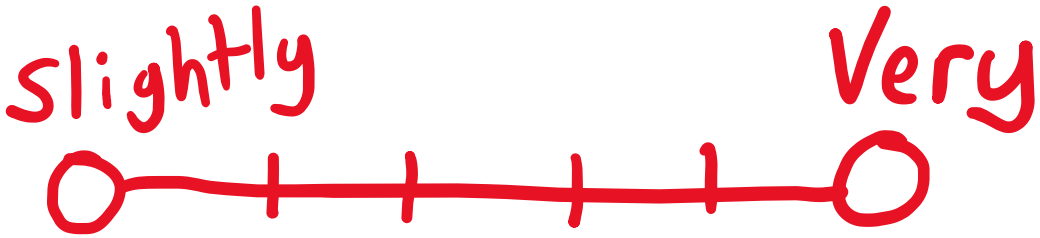
- One issue with using sentiment lexicons is that it is not always obvious to which sentiment class each term belongs.
- Humans express a much wider range of emotional and cognitive states besides positive and negative.
- A **second** class of models is more strongly rooted in the use of polarity-based lexicons.

Polarity Based Lexicon

A word or phrase that is assigned a broader sentiment class (think emotion or state of mind). e.g. Happy, Sad, Optimistic, Pessimistic…

Valence-based lexicons…

HAPPY



Valence-based lexicon

A word or phrase that is assigned a sentiment strength or intensity, for example if "good" receives a score of 1 towards being positive "excellent" would receive a valence score of 2.5

A **third** category of models recognises even within the same emotional state, individuals may lie on a spectrum between the two extremes

Context-aware lexicons…

- “I really like the new Top Gear”
- In this case ”really” is important here since it magnifies the impact of a weakly positive lexicon ”like”

Context-Aware Lexicon

A word or phrase whos sentiment is interpreted in light of the surrounding lexicons and the way in which the word is used

"I very much enjoyed that coffee"

Context-aware lexicons…

- Negation and affirmation

"The experience was not very interesting"

- what happens to the sentiment score?
- does it have to appear earlier?

Bigrams, trigrams, ngrams…

- A common way of deriving context-aware lexicons is using bigrams or more generally ngrams.

```
import nltk
from nltk.collocations import *
tokens = nltk.wordpunct_tokenize("I really like the new Top Gear")
finder = BigramCollocationFinder.from_words(tokens)
```

```
tokens
```

```
['I', 'really', 'like', 'the', 'new', 'Top', 'Gear']
```

```
for k,v in finder.ngram_fd.items():
    print(k,v)
```

```
('I', 'really') 1
('really', 'like') 1
('new', 'Top') 1
('like', 'the') 1
('the', 'new') 1
('Top', 'Gear') 1
```

VADER - 1

- Valence Aware Dictionary for sEntiment Reasoning
- One of several algorithms developed around social media data (2014)
- Uses of combination of qualitative and quantitative methods to produce a sentiment lexicon that is especially designed to be used in microblogging contexts
- It was developed by analysis of a sample of several hundred tweets that were manually reviewed to identify common sentiment features
- VADER uses a mixture of standard sentiment and valence lexicons together with custom context-aware lexicons

VADER - 2

VADER's 5 context-aware lexicons

1) **Punctuation**, namely the exclamation point (!), increases the magnitude of the sentiment intensity without modifying the semantic orientation. For example, "The food here is good!!!" is more intense than "The food here is good."

2) **Capitalization**, specifically using ALL-CAPS to emphasize a sentiment-relevant word in the presence of other non-capitalized words, increases the magnitude of the sentiment intensity.

VADER - 3

3) **Boosting modifiers** impact sentiment intensity by either increasing or decreasing the intensity. For example, “The service here is extremely good”.

4) **Contrastive conjunctions** i.e., “but” signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. “The food here is great, but the service is horrible”

5) **Extraction of tri-gram features** to identify cases where sentiment flips due to negation, i.e. The food here isn’t all that great

Exploring the VADER code…

https://github.com/nltk/nltk/blob/56bc4af35906fb636c11d0cbc3c8ea54447def24/nltk/sentiment/vader.py#L596

from the code we can see that hashtags are not included in the sentiment score.

VADER - 4

- VADER has been shown to provide good classification performance when compared with the grouped sentiment opinion of human reviewers.
- VADER is good in picking up negation and emphasis.
- Results obtained through the applications of VADER to non-microblogging data, such as movie reviews, have been found to be less correlated with the opinion of human reviewers.
- This is perhaps not as surprising given that VADER has been largely developed to analyse data from Twitter.

Hutto, C.J. Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014

Applying the VADER model…

```
import nltk

nltk.download('vader_lexicon')

message = "I didn't enjoy the film but at least the ticket was cheap!"

from nltk.sentiment.vader import SentimentIntensityAnalyzer
sia = SentimentIntensityAnalyzer()
sia.polarity_scores(message)
```

```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
{'neg': 0.174, 'neu': 0.826, 'pos': 0.0, 'compound': -0.2746}
```

$-1 \longrightarrow 1$

neg pos

VADER model (Last of Us data)…

```
import pandas as pd
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer

nltk.download('vader_lexicon')

sia = SentimentIntensityAnalyzer()
df = pd.read_csv("lastofus.csv", index_col=0)
df = df[df["body"].str.contains("Joel")]

def score(row):
  return sia.polarity_scores(row["body"])["compound"]

df["sentiment"] = df.apply(score, axis=1)
```

VADER model (Last of Us data)…

```
df
```

	body	created_utc	author	upvotes	sentiment
1	Joel's post-apocalyptic apartment is better th...	2023-01-17 02:59:55	pyRSL64	276	0.4404
3	I just finished the first episode, and I reall...	2023-01-17 01:51:03	folder_finder	105	0.9198
4	Anybody else notice the bookmarked page of the...	2023-01-17 05:08:04	dolpgg	99	0.0000
7	Why was Joel able to cut the line for the towe...	2023-01-17 05:12:37	TraditionalContest6	47	-0.7987
9	Tremendous first episode. It's almost like the...	2023-01-17 10:02:59	Jedi_Mindtrix53	48	-0.5489
16	I watched the first episode twice and I'm prob...	2023-01-16 20:59:13	stillwatersrunfast	77	0.9565

```
df["sentiment"].describe()
```

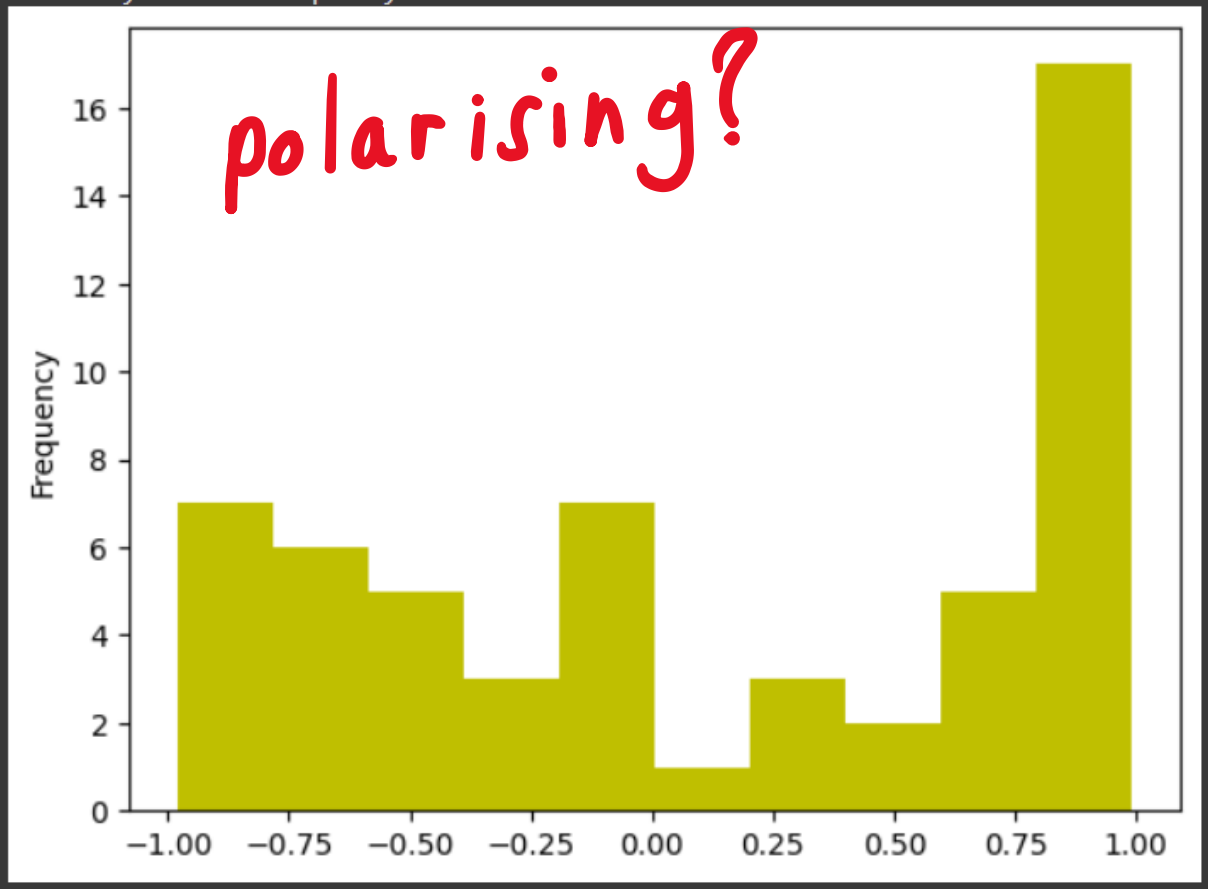
```
count    56.000000
mean      0.118670
std       0.700232
min      -0.981500
25%      -0.541775
50%       0.076550
75%       0.862700
max       0.992300
Name: sentiment, dtype: float64
```

how to improve?

Plotting the sentiment scores…

```
df["sentiment"].plot(kind="hist", color="y")
```

```
<Axes: ylabel='Frequency'>
```



VADER model (Last of Us data)…

```
df.sort_values(by="sentiment", ascending=False).head(10)[["body","sentiment"]]
```

	body	sentiment
60	I saw some ads on Instagram about this show bu...	0.9923
82	Like Joel and Ellie, Bill is the protector, Fr...	0.9791
23	I loved the first episode! I was scared severa...	0.9770
117	Honest question for game players (I'm a non-ga...	0.9678
16	I watched the first episode twice and I'm prob...	0.9565
50	Im belatedly getting into the show, which mean...	0.9538
465	That's still a really high rate. Kansas City w...	0.9522
114	I was reeeeeeeeeeeally hoping they would copy ...	0.9274
207	Yeah that's always been the real deciding fact...	0.9231
53	This is amazing and:\n\nI don't recall HBO ser...	0.9230

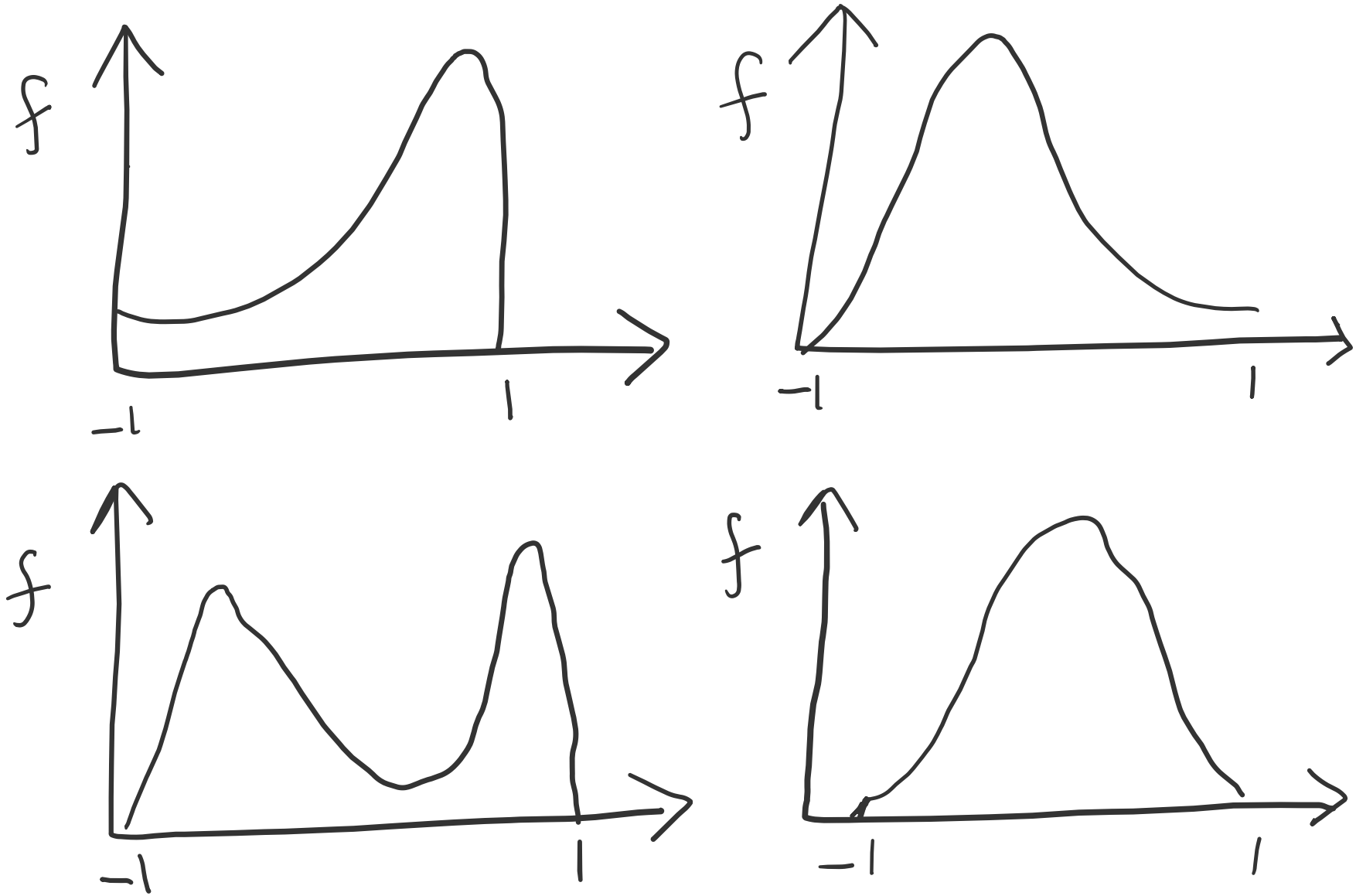
View in collab

VADER model (Last of Us data)…

```
df.sort_values(by="sentiment", ascending=True).head(10)[["body","sentiment"]]
```

	body	sentiment
374	I never played and only knew that it was mostl...	-0.9815
151	I figured they were trying to land at the Aust...	-0.9578
54	When Joel was beating up the soldier I know I'...	-0.9552
92	I don't think joel should have killed the doct...	-0.9451
475	Literally had a nightmare last night about it....	-0.9001
448	It is not just the mediums. It works because t...	-0.8853
7	Why was Joel able to cut the line for the towe...	-0.7987
37	Tbh the doctors plan sounds so suss, it's hard...	-0.6969
84	After Joel and Ellie meet, and he says he's go...	-0.6908
271	In the game you meet Bill, who has a town set ...	-0.6858

Sentiment distributions…



TextBlob sentiment classifier...

```
import nltk
nltk.download('movie_reviews')
nltk.download('punkt')

from textblob import TextBlob
from textblob.sentiments import NaiveBayesAnalyzer

analyser = NaiveBayesAnalyzer()

message = "I didn't enjoy the film but at least the ticket was cheap!"

blob = TextBlob(message, analyzer=analyser)
blob.sentiment
```

training dataset

tokenisation

```
[nltk_data] Downloading package movie_reviews to /root/nltk_data...
[nltk_data]   Package movie_reviews is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
Sentiment(classification='neg', p_pos=0.42862491881565234, p_neg=0.5713750811843475)
```

Naïve Bayes…

- Each lexicon contributes equally to the sentiment score.
- Each lexicon is assumed to occur independently of all others.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) \Rightarrow P(B|A) \cdot P(A)$$

Classifying all the comments…

```
import nltk
nltk.download('movie_reviews')
nltk.download('punkt')

from textblob import TextBlob
from textblob.sentiments import NaiveBayesAnalyzer

analyser = NaiveBayesAnalyzer()

def bayes_sentiment_class(row):
  blob = TextBlob(row["body"], analyzer=analyser)
  if blob.sentiment.p_pos >= 0.6:
    return "Positive"
  if blob.sentiment.p_neg >= 0.6:
    return "Negative"
  return "Neutral"

df["bayes_sentiment"] = df.apply(bayes_sentiment_class, axis=1)
```

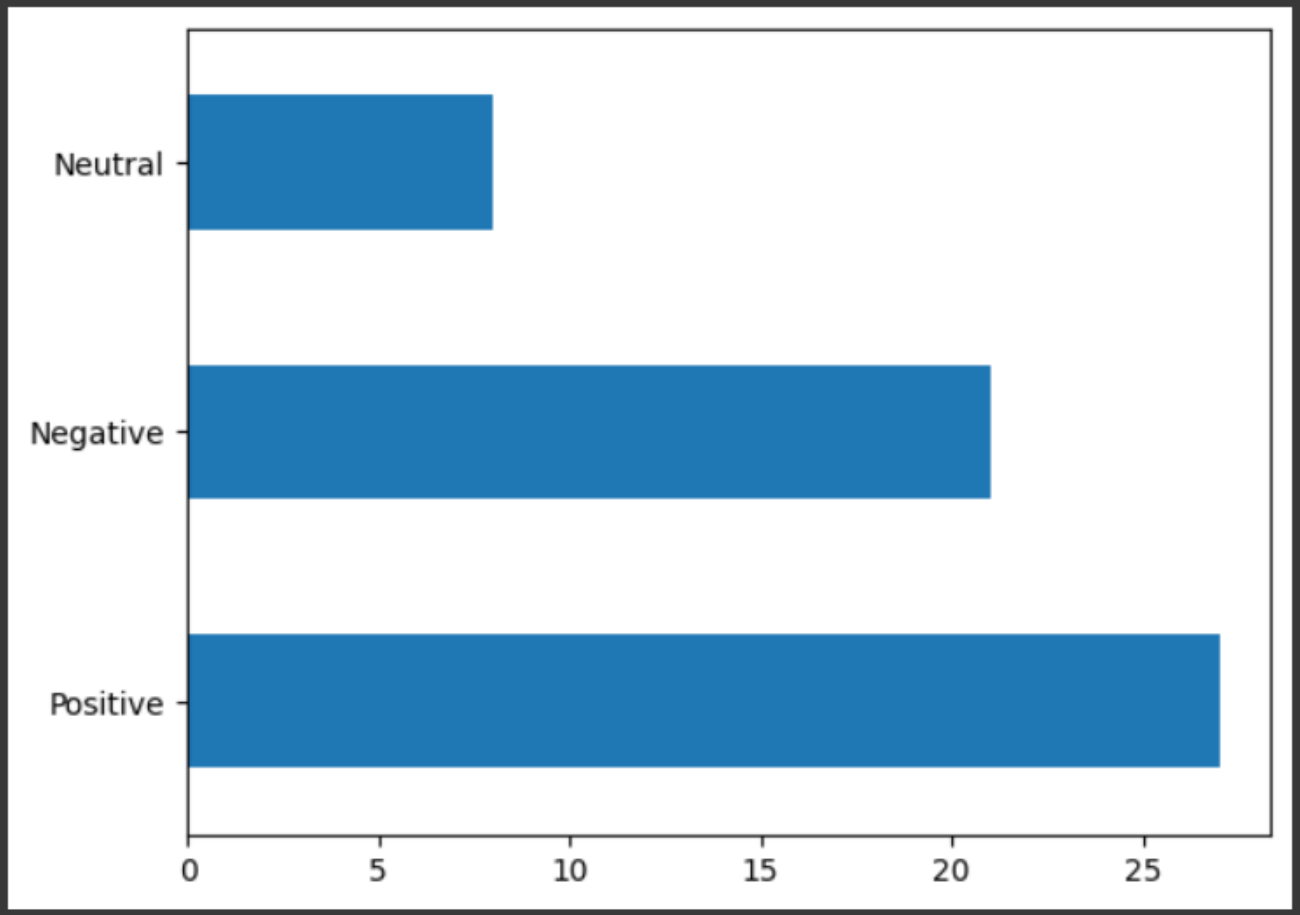
DataFrame with sentiment class…

	body	created_utc	author	upvotes	sentiment	bayes_sentiment
1	Joel's post-apocalyptic apartment is better th...	2023-01-17 02:59:55	pyRSL64	276	0.4404	Neutral
3	I just finished the first episode, and I reall...	2023-01-17 01:51:03	folder_finder	105	0.9198	Positive
4	Anybody else notice the bookmarked page of the...	2023-01-17 05:08:04	dolpgg	99	0.0000	Negative
7	Why was Joel able to cut the line for the towe...	2023-01-17 05:12:37	TraditionalContest6	47	-0.7987	Positive
9	Tremendous first episode. It's almost like the...	2023-01-17 10:02:59	Jedi_Mindtrix53	48	-0.5489	Positive
16	I watched the first episode twice and I'm prob...	2023-01-16 20:59:13	stillwatersrunfast	77	0.9565	Positive
23	I loved the first episode! I was scared severa...	2023-01-17 09:57:40	Isthatanewtie	15	0.9770	Positive
26	This is something I was thinking of today. Da...	2023-03-08 04:04:04	elways_love_child	16	-0.5369	Positive
33	Anyone notice Joel keeps struggling with bolt ...	2023-02-14 23:37:35	TheeOneWhoKnocks	13	-0.4423	Negative
37	Tbh the doctors plan sounds so suss, it's hard...	2023-03-22 01:19:55	Safe-Watercress-6477	13	-0.6969	Negative

Visualising the sentiment class…

```
df["bayes_sentiment"].value_counts().plot(kind="barh")
```

```
<Axes: >
```



In Summary

- Sentiment analysis concerns identifying the opinion, emotion or viewpoint expressed in unstructured text.
- Sentiment analysis has been applied in a range of commercial and non-commercial settings where traditionally surveys or polls might have been utilised.
- Sentiment classifiers use three basic sentiment states: positive, negative and neutral.
- These classifiers determine the sentiment expressed by considering the lexicons used and their associated sentiment orientation.
- More advanced sentiment classifiers make use of additional types of lexicons to understand the emotional state and the strength of emotion expressed.
- VADER is one of the most famous sentiment models developed to model the sentiments of comments posted by users to micro-blogging services.

End