Sentiment Analysis of Amazon reviews

Edo Spigel Emmerich
Capstone Sprint 2

Why reviews?

Consumers are choosing **online platforms** over traditional retail.

Drive sales

Build trust

Provide insights

How can we use machine learning to understand the contents of reviews?

Utilise **Natural Language Processing** to analyse the **sentiment** of reviews.

Compare sentiment analysis results with star ratings to **find discrepancies**.

The Data

Column name	Datatype	Measures
overall	float	Overall star rating of review
verified	boolean	Whether the review has been verified as real or not.
reviewTime	object	Time of review
reviewerID	object	Unique ID of reviewer
asin	object	Product metadata
style	object	Product metadata
reviewerName	object	Name of reviewer
reviewText	object	Textual contents of review
summary	object	Textual summary of review
unixReviewTime	int64	Time of review since Unix Epoch on January 1st, 1970
vote	object	Count of usefulness vote
image	object	Image of product reviewed

Review Star rating

Product and reviewer metadata

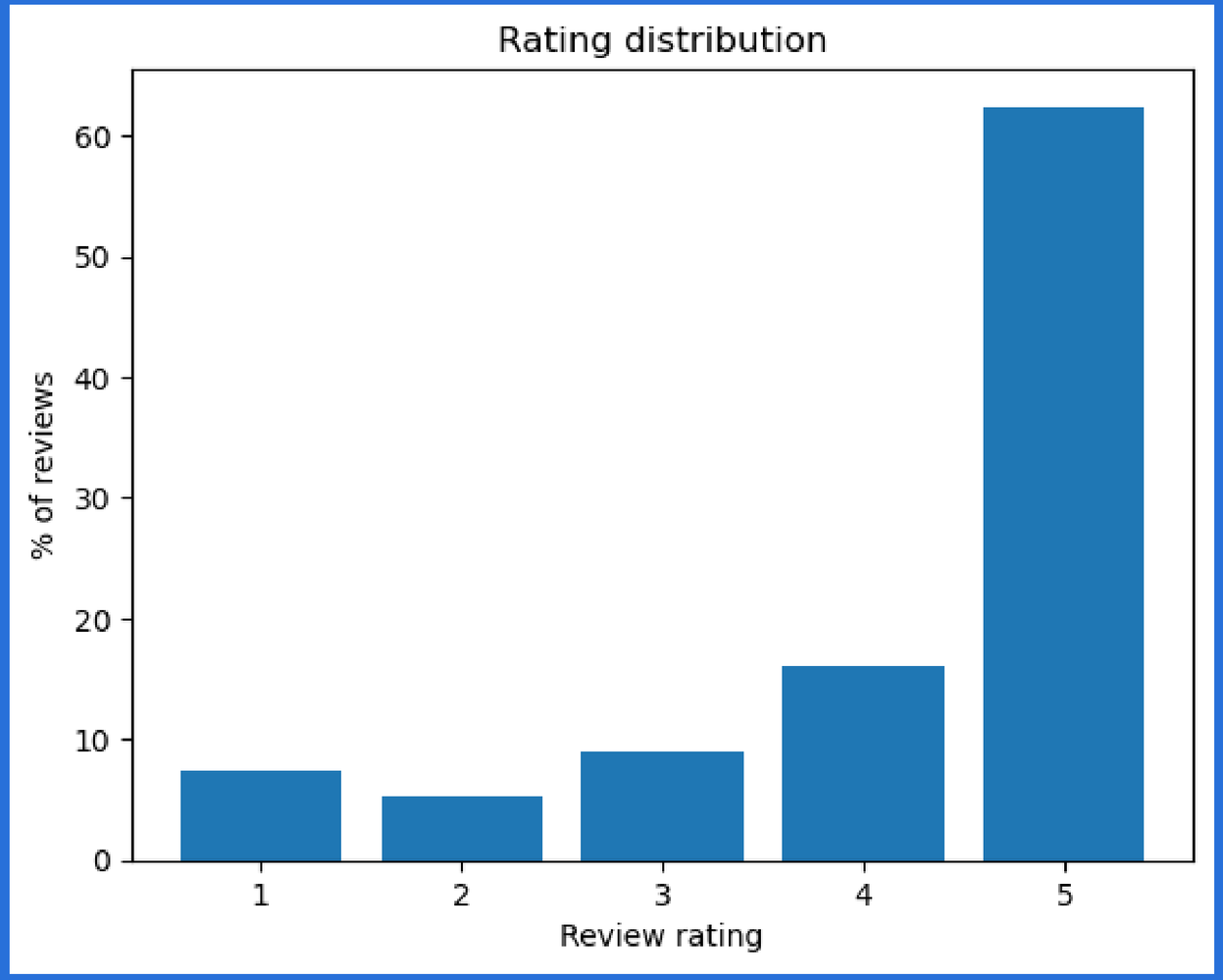
Text review and summary

Extra information

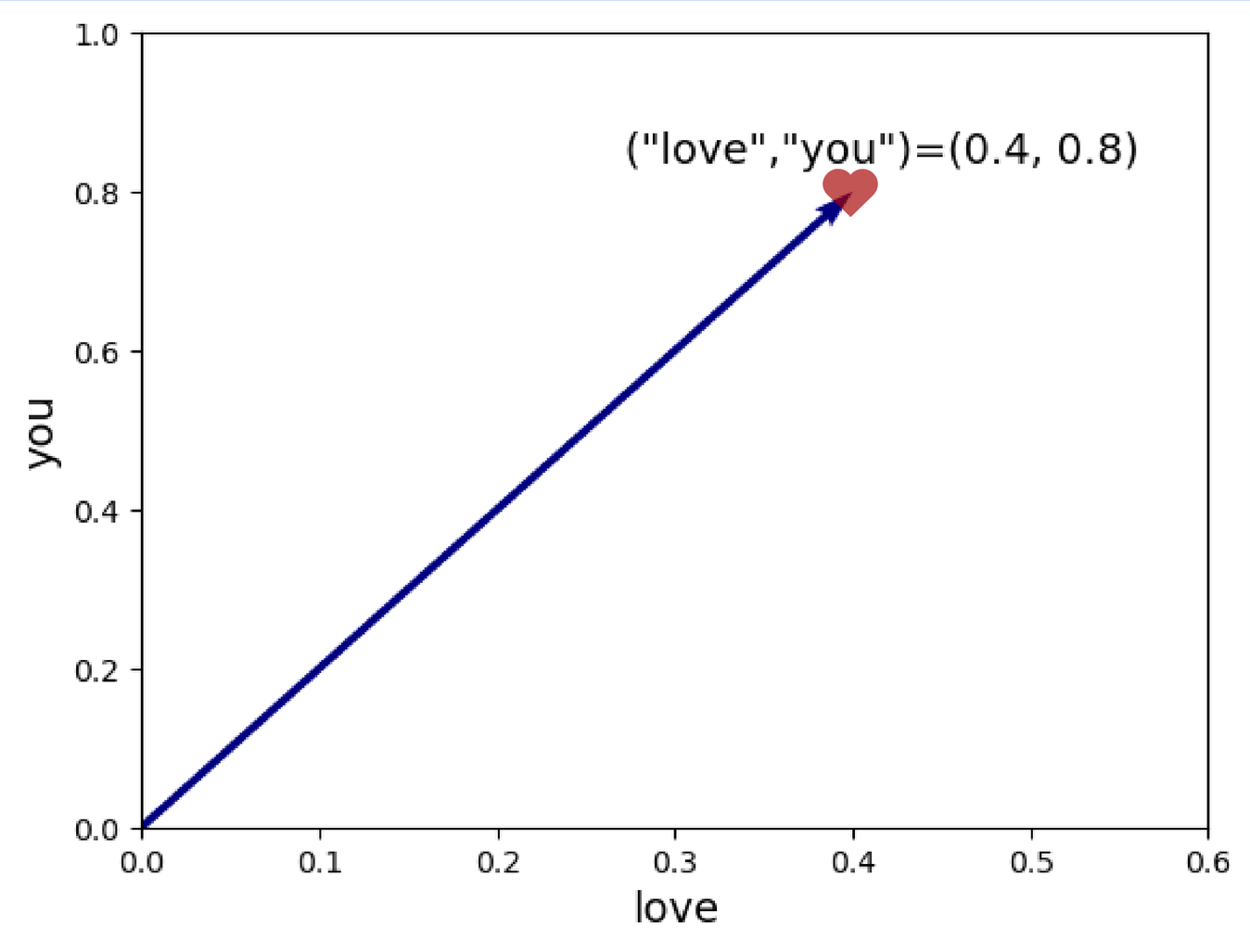
0h9ne
0hone

stormtrooper
stormtrouper

ccover
stornger

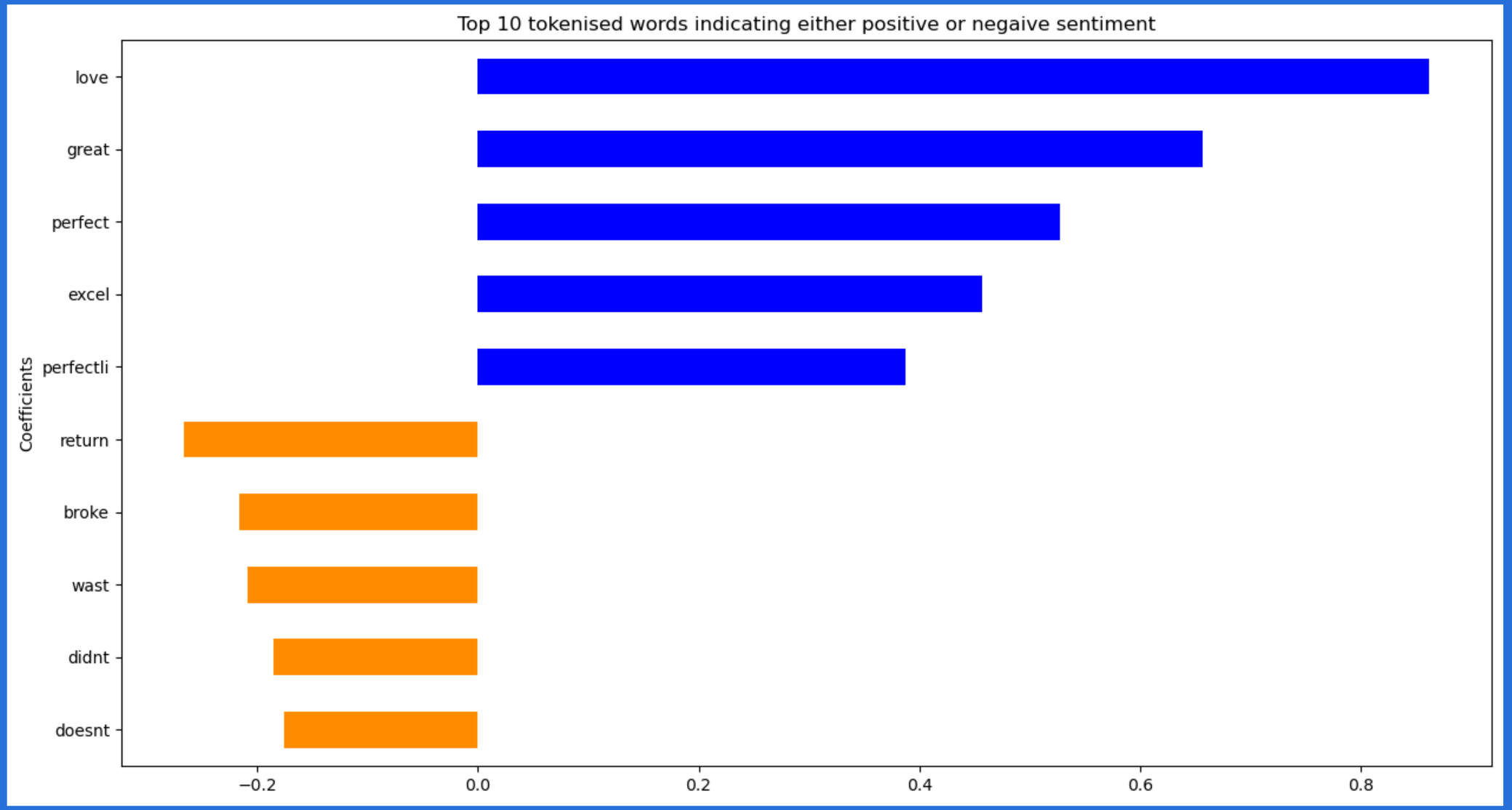


TF-IDF Vectorization



	tfidf
had	0.493562
little	0.493562
tiny	0.493562
house	0.398203
mouse	0.235185
the	0.235185
ate	0.000000

Top words from logistic regression



Top Positive Words

love
great
perfect
excel
perfectli

Top Negative Words

return
broke
wast
didnt
doesnt

I used this case for not even a week and the bow came off. I loved it so pretty, but wish it would o stayed together.

I used this case for not even a week and the bow came off. I loved it so pretty, but wish it would o stayed together.

Guess: Positive

I used this case for not even a week and the bow came off. I loved it so pretty, but wish it would o stayed together.

Guess: Positive

Overall: 2/5

Actual: Negative

I used this case for not even a week and the bow came off. I loved it so pretty, but wish it would o stayed together.

Contribution	
love:	1.53
wish:	0.56
pretti:	0.25
use:	0.074
doesnt:	0.036
return:	0.033
didnt:	0.031
dont:	0.029
broke:	0.029
month:	0.024

Confidence: 0.58

Logistic regression performs well!
Not great at predicting positive reviews

Decision Tree doesn’t do as well
okay with positive reviews, awful at negative

Random forest has best performance but
still struggles with negative reviews

Next steps: Pre-processing

spaCy for better text preprocessing

Text embeddings and neural network