Analysis of customers' suggestions for product improvement in Amazon reviews

GROUP 6

ELINA
TUGAEVA

ELIANA
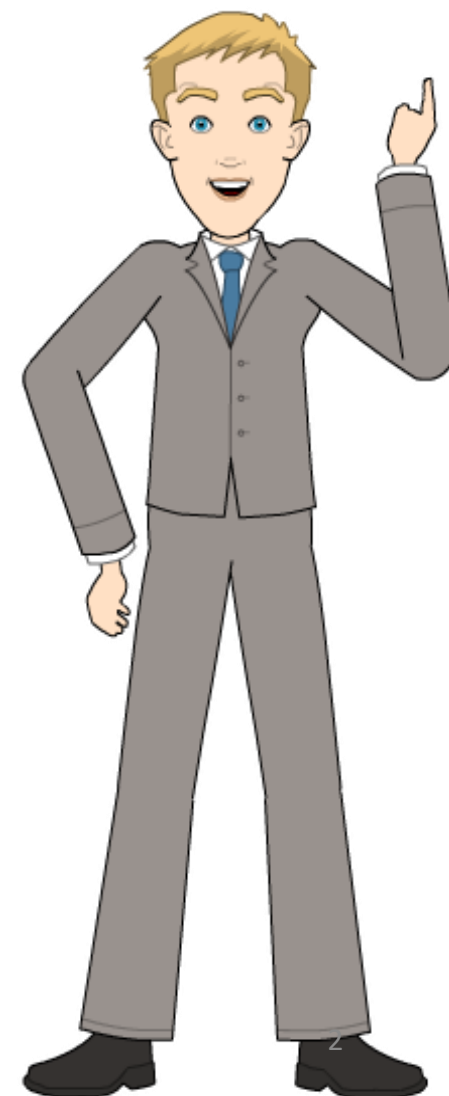RUSLANOVA

MARLENA KUC

SAMAIL
GULIYEV

NIKOLAJ
KOLBASKO

04/12/19

Application area

	⭐⭐ I hate the phone.	⭐⭐ The battery discharges super quickly. Black screen happened only after 1 month of usage.
Sentiment	negative	negative
Information retrieved	Customer did not like the product	Customer did not like the product and we also know why. We also know that this consumer is an engaged consumer.

GOALS

1. Distinguish reviews that contain suggestions about the product

The battery discharges super quickly.

Suggestion?

Yes

2. Determine the features the customer complains the most about

The battery discharges super quickly.

Features?

Battery

AGENDA

1. Structure and size of the data set
2. Preprocessing
3. Classification
4. Clustering
5. Results

DATA SET

400 000+

reviews extracted
by PromptCloud
web crawler
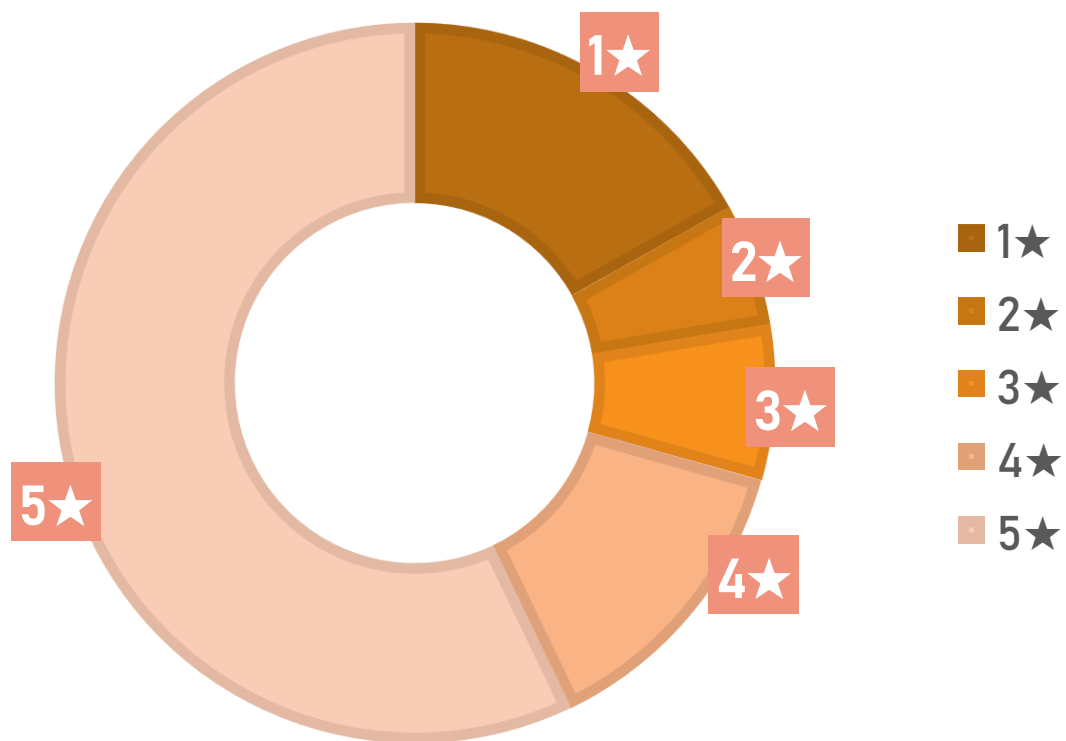
10 000
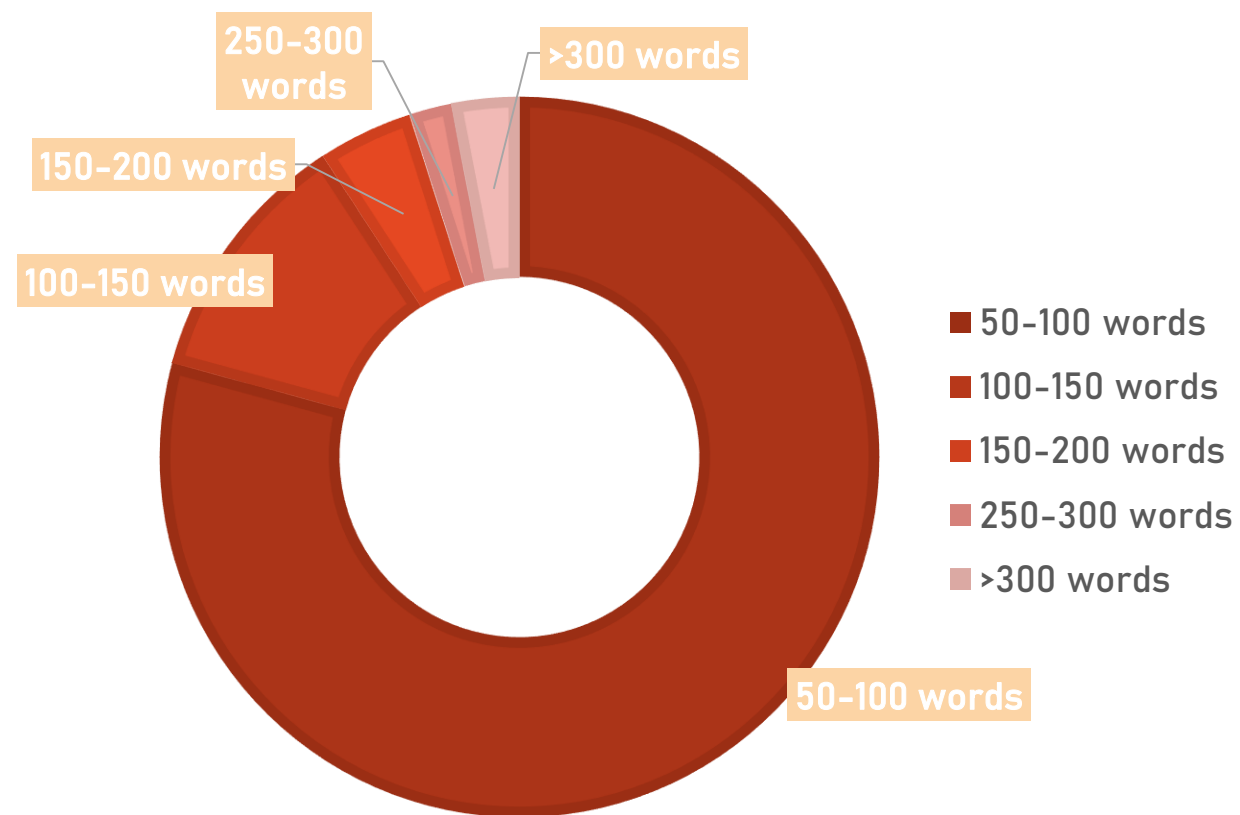
used due to
computational
constraints

Fields:

- Product Title
- Brand
- Price
- Rating
- Review text
- Number of people who
found the review helpful

DATA SET

RATING DISTRIBUTION

1★
2★
3★
4★
5★

Legend:
- 1★
- 2★
- 3★
- 4★
- 5★

LENGTH OF REVIEW

250-300 words
>300 words
150-200 words
100-150 words
50-100 words

Legend:
- 50-100 words
- 100-150 words
- 150-200 words
- 250-300 words
- >300 words

PREPROCESSING

1 Sentence tokenization

2 Manual labelling

3 Remove punctuation

4 Word tokenization

5 Remove the stopwords

6 POS tagging

CLASSIFICATION

	SpaCy	InferSent	Naïve Bayes
Validated on	*5-folds cv*	*1-fold cv*	*5-fold cv*
Accuracy	0.68	0.65	0.70
Precision	0.47	0.52	0.69
Recall	0.58	0.28	0.63

CLASSIFICATION

	SpaCy	InferSent	Naïve Bayes
Validated on	5-folds cv	1-fold cv	5-fold cv
Accuracy	0.68	0.65	0.70
Precision	0.47	0.52	0.69
Recall	0.58	0.28	0.63

1 410 "suggestion" sentence

4 500 classified sentences

10 000 randomly chosen reviews (29 400 sentences)

400 000 reviews Kaggle data set
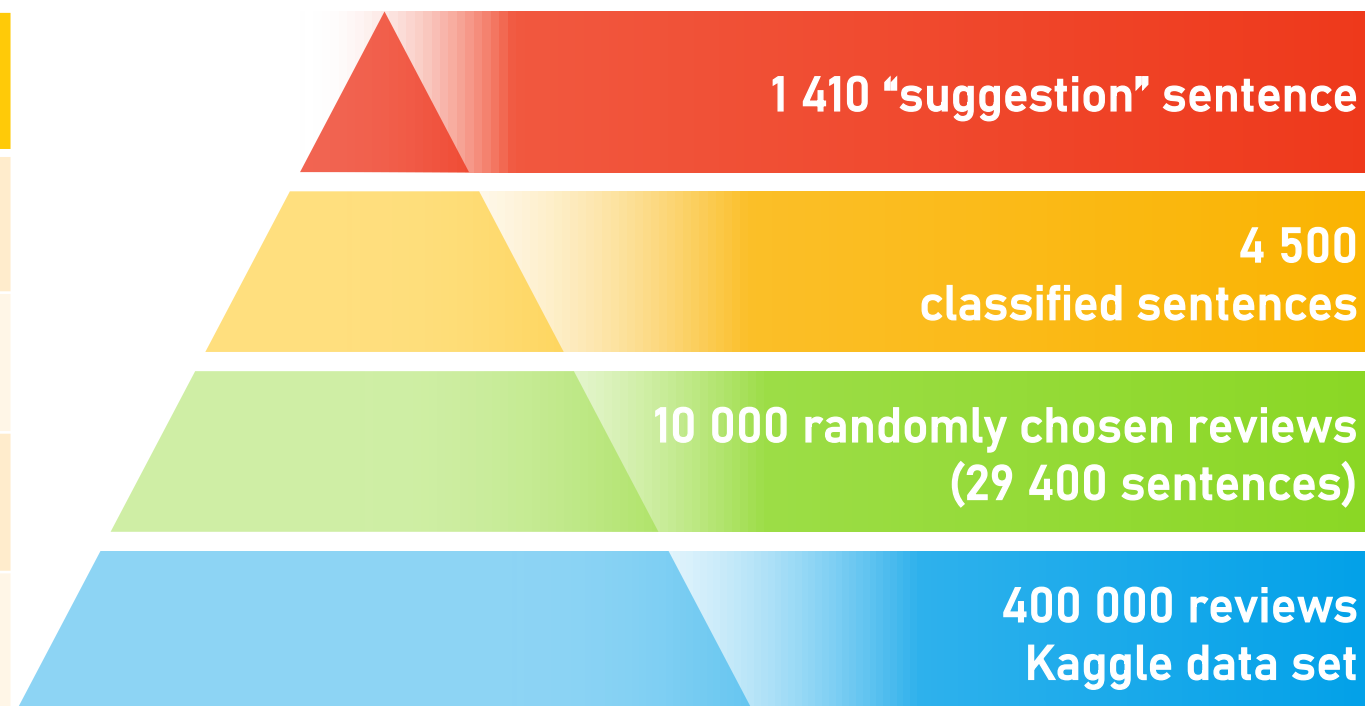
CLUSTERING

Input: TF–IDF vectorized "suggestion" sentences

Method: KMeans

Number of clusters: 12

Elbow Method For Optimal k

GOOD CLUSTERS
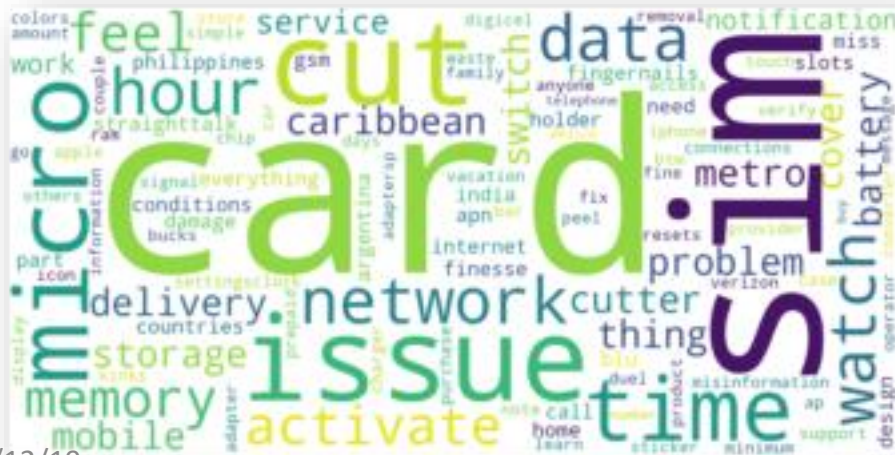
BAD CLUSTERS

Results

	1. Distinguish reviews that contain suggestions about the product	2. Determine the features the customer complains the most about
Solution:	SpaCy	KMeans
Result:	4500 classified sentences	Clusters with identifiable features

Thank you for attention!

Q&A