Deloitte.

Classifying Amazon Reviews

AI Academy Capstone Project

Frida Martinez Flores - Analyst

01 Overview

02 Analysis

03 Modelling

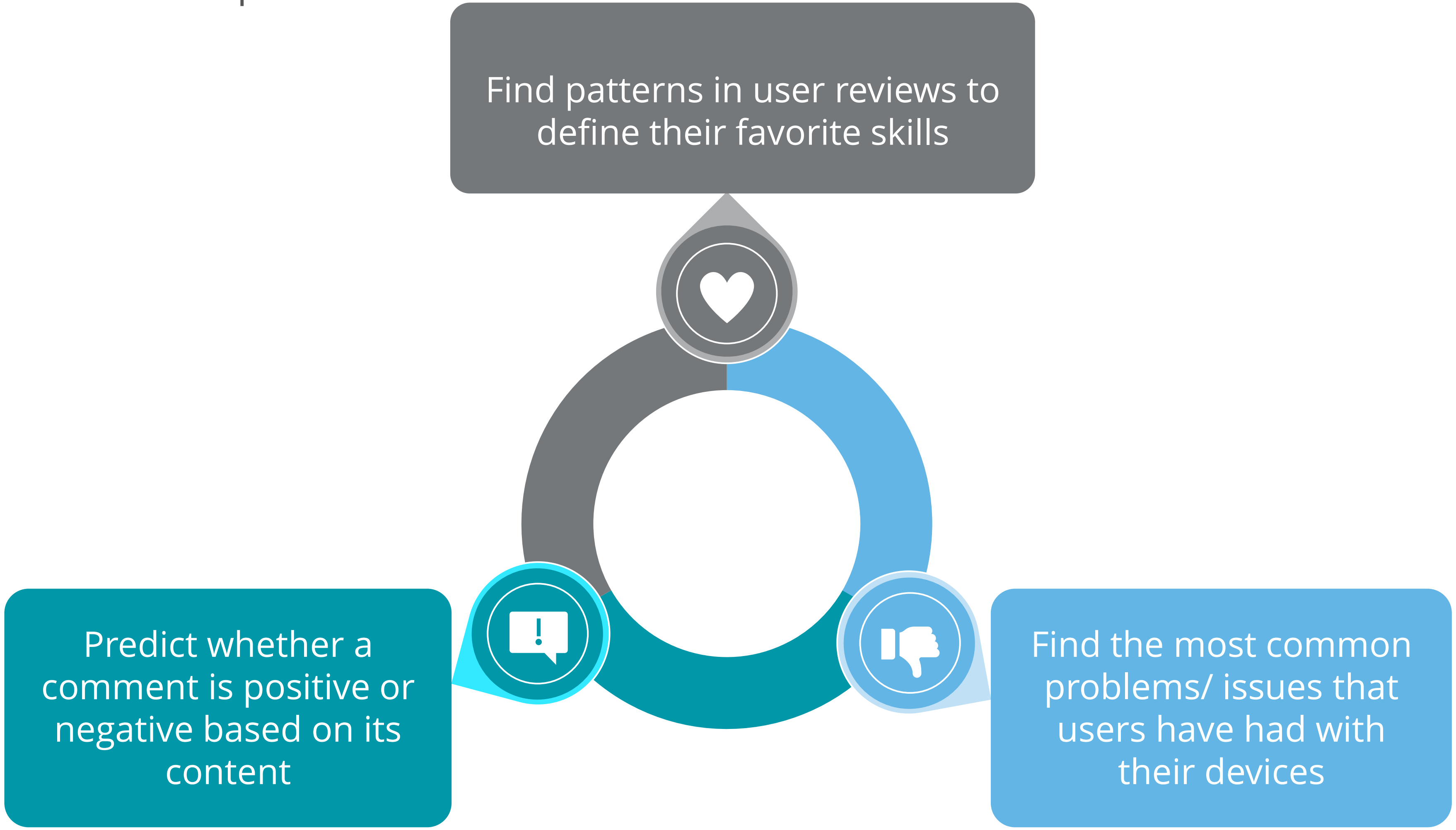
04 Conclusion and Next Steps

05 Q&A

Overview

Goals

How can Classification help?



Find patterns in user reviews to define their favorite skills

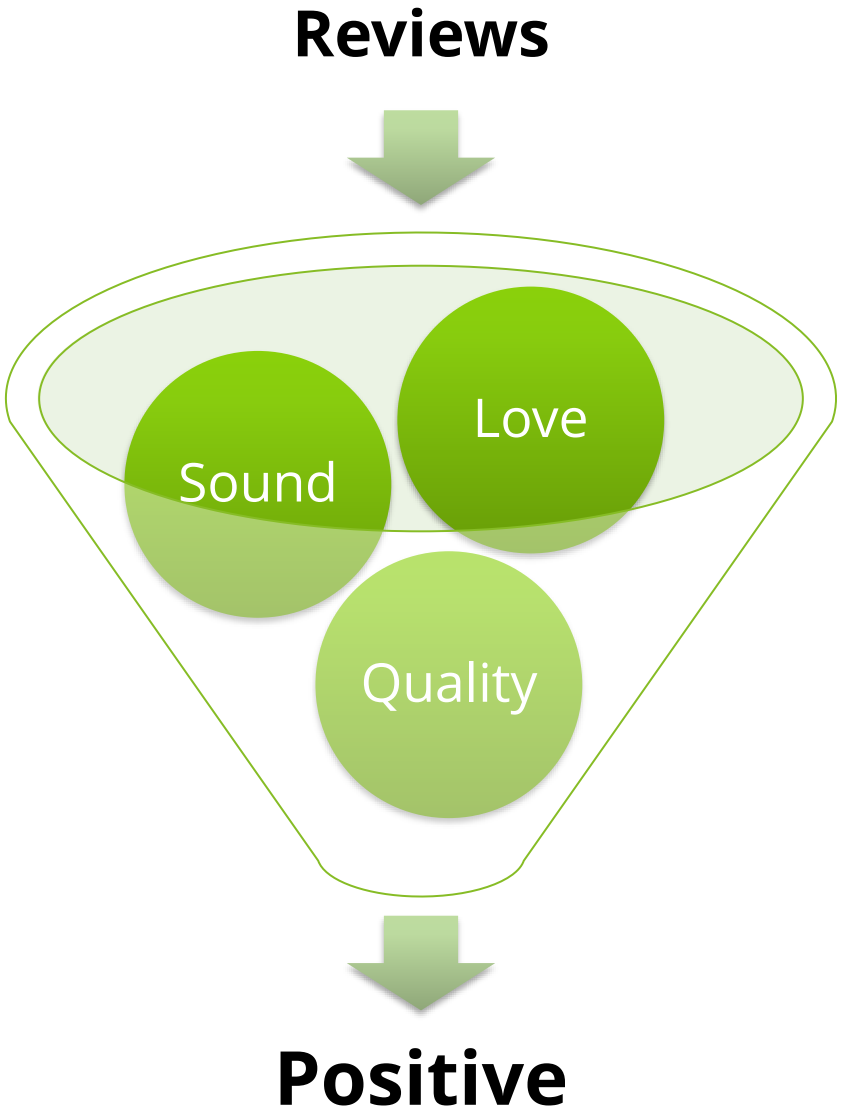
Predict whether a comment is positive or negative based on its content

Find the most common problems/ issues that users have had with their devices

Classification

Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data.

Classifying Amazon Reviews



Reviews

Sound

Love

Quality

Positive

Amazon Reviews Data Set

3150 Amazon customers reviews for Alexa Echo, Firestick, Echo Dot etc.

Rating	Date	Variation	Verified Reviews	Feedback
5	31-Juli-18	Charcoal Fabric	Love my Echo!	1
2	31-Jul-18	Walnut Finish	Without having a cellphone, I cannot use many of her features	0

Feedback = Labels
1: Positive Review
0: Negative Review

Rating:

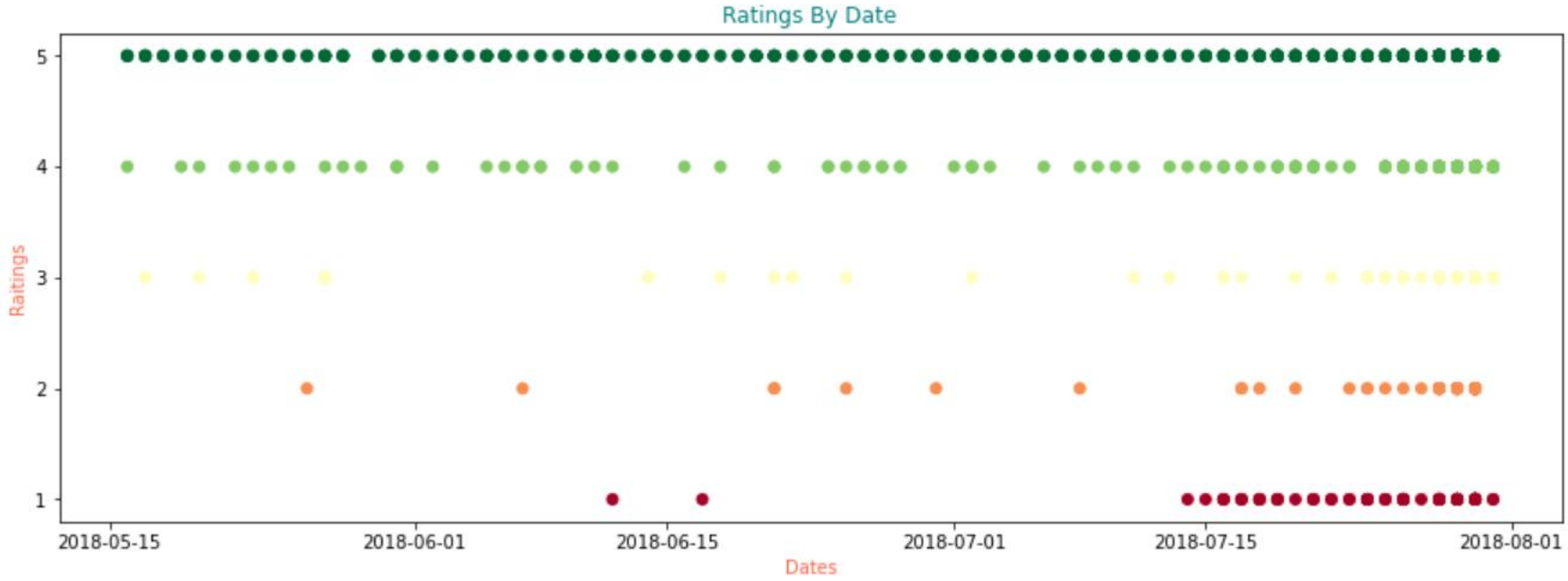
★★★★★	5
★★★★	4
★★★	3
★★	2
★	1

Variation:
Characteristic differing the models of the product.

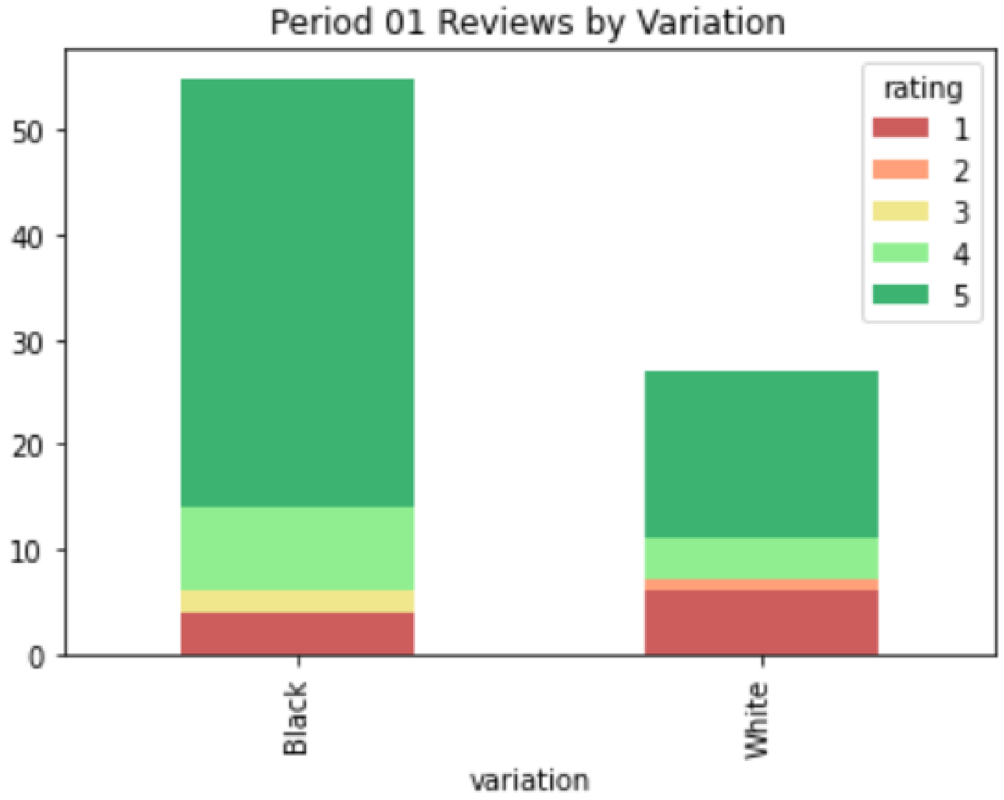
Analysis

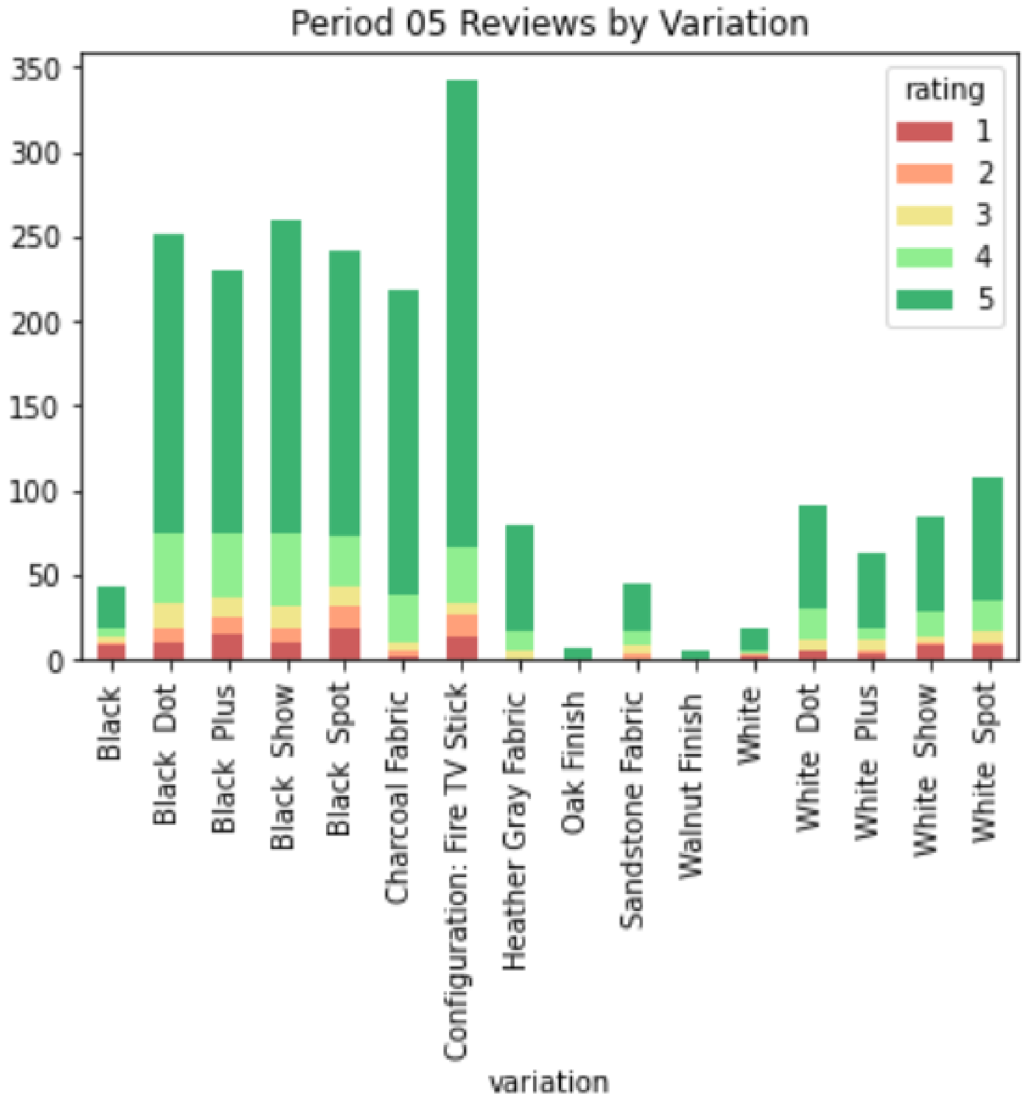
Numeric Data Analysis

The number of positive reviews keeps constant while the negative reviews have increased in the last analyzed period.



The number of variations (sources from which users have uploaded their reviews) has increased. Resulting in an increase in both positive and negative reviews

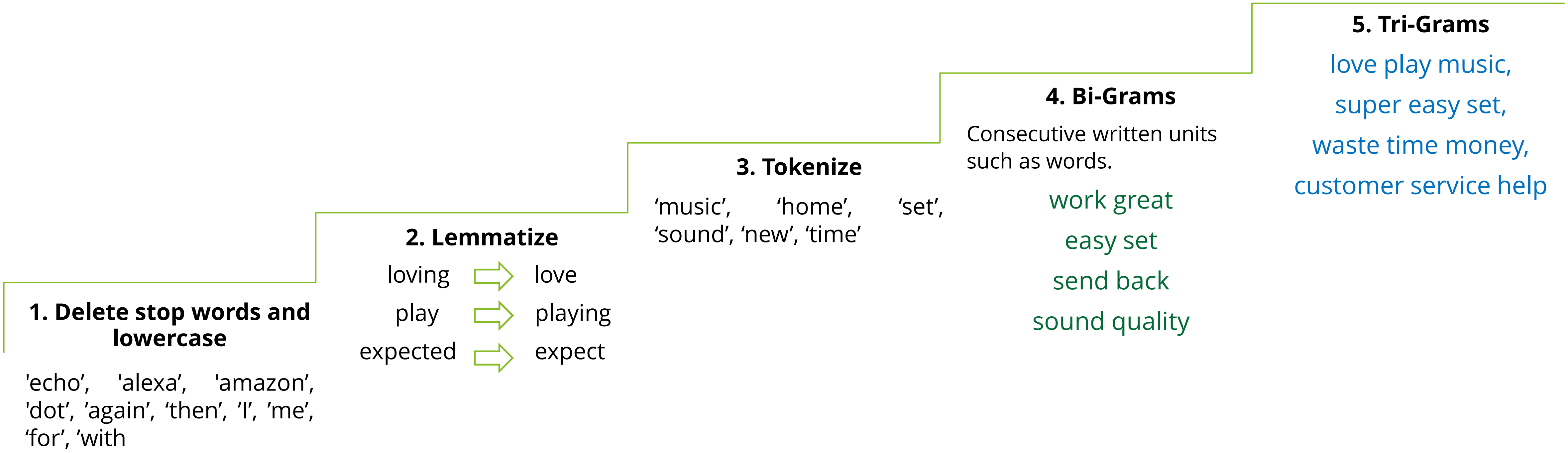




Text Data Analysis

NLP Cleaning and Preprocessing

Removing and transforming certain parts of the text so that it becomes more easily understandable for NLP models that are learning the text.



1. Delete stop words and lowercase

'echo’, 'alexa’, 'amazon’, 'dot’, ’again’, ‘then’, ’I’, ’me’, ‘for’, ’with

2. Lemmatize

loving → love

play → playing

expected → expect

3. Tokenize

‘music’, ‘home’, ‘set’, ‘sound’, ‘new’, ‘time’

4. Bi-Grams

Consecutive written units such as words.

work great

easy set

send back

sound quality

5. Tri-Grams

love play music,

super easy set,

waste time money,

customer service help

Modelling and Evaluation

Modeling Steps

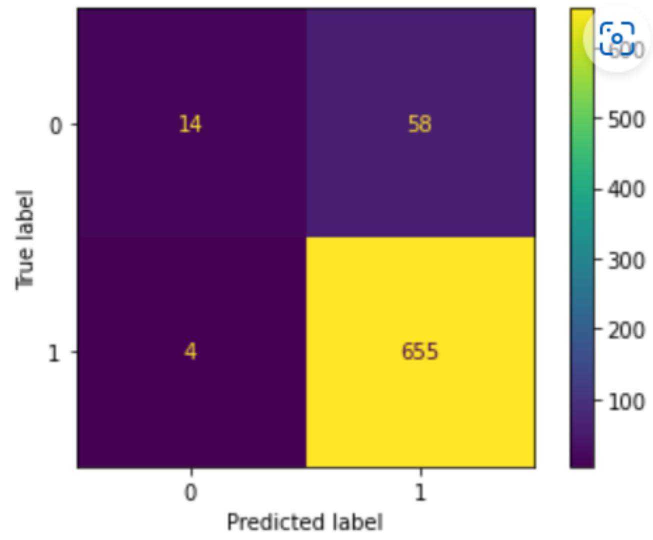
Developed different models to compare its efficiency

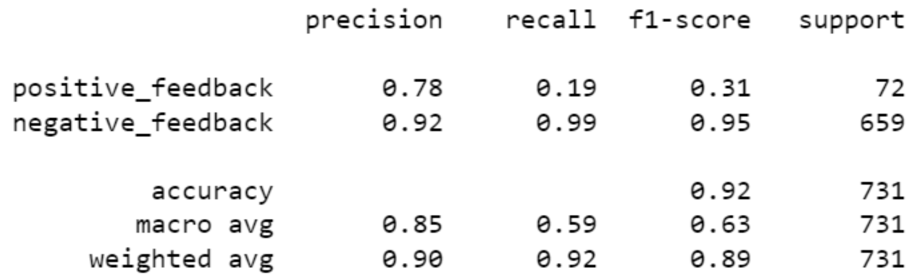
Decision Tree Classification

	Transform Data	Split data in training and test sets	Create Model	Observe Predictions
Model #1	TF-IDF	Train: 60% data Test: 40% data	Criterion = Entropy	Accuracy Score, Confusion Matrix
Model #2	Count Vectorizer	Train: 60% data Test: 40% data	Criterion = Entropy	Accuracy Score, Confusion Matrix
Model #3	TF-IDF	Train: 60% data Test: 40% data	Criterion = Gini	Accuracy Score, Confusion Matrix
Model #4	Count Vectorizer	Train: 60% data Test: 40% data	Criterion = Gini	Accuracy Score, Confusion Matrix
Model #5	Count Vectorizer	Train: 60% data Test: 40% data	Splitter = Random	Accuracy Score, Confusion Matrix

Criterion: Function to measure the quality of a split.
Splitter: Strategy used to choose the split at each node

Model 3





```
                   precision    recall  f1-score   support

positive_feedback       0.78      0.19      0.31        72
negative_feedback       0.92      0.99      0.95       659

         accuracy                           0.92       731
        macro avg       0.85      0.59      0.63       731
     weighted avg       0.90      0.92      0.89       731
```

Conclusion

Q&A