



TATIA PROJECT



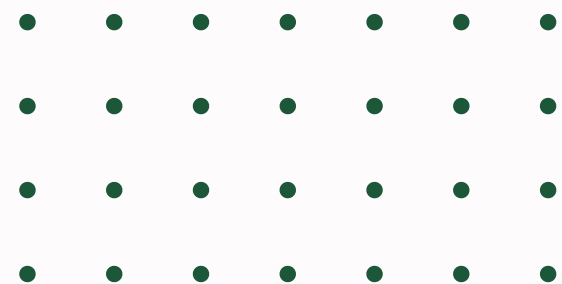
Sentiment Analysis on Books Reviews

BOUDIAF YASMINE
ALIREZA FOROUTAN TORKAMAN

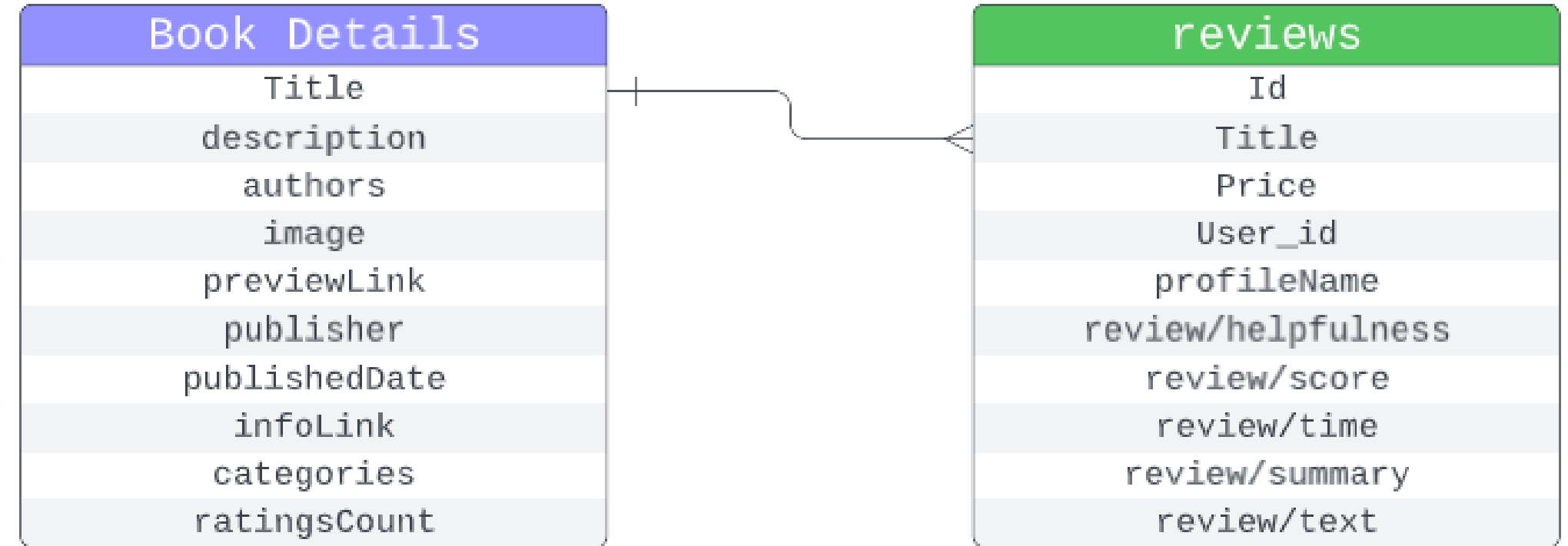
2023/2024

Content

- 01** Overview
- 02** project steps
- 03** Data Preparing
- 04** Models
- 05** Training and Evaluation
- 06** Rating Classification
- 07** Discussion
- 08** Conclusion



Overview



objective

- guessing the rating from the text reviews

dataset

- Amazon books reviews from kaggle

DATA PREPARING


Data reduction & distribution

- dataset too large (3 000 000 rows) and different proportions concerning the ratings
- we took 10 000 instances of each rating (1 to 5)

3000000 rows × 6 columns

Adding 'scale' column

with the aim of making a first classification of text into 3 levels: positive, neutral and negative



review/score	review/scale
4.0	positive
5.0	positive
5.0	positive
4.0	positive
4.0	positive

lemmatization & text vectorization

- we used WordNetLemmatizer() from nltk for lemmatization, the goal is to transform different inflected forms of a word into a common base form
- we choose Tfidf because it fits our data : automatically downweights common terms and giving more importance to terms that are discriminative

MODELS

testing multiple models to choose the more efficient

SVC model

The primary objective of an SVM is to find a hyperplane that best separates the data points of different classes in a feature space.

We implemented 2 versions of this model with different kernel function parameters:

linear and rbf

MLP Classifier

Multi-Layer Perceptron Classifier is characterized by its architecture, which consists of multiple layers of interconnected nodes or neurons.

We trained this model with different parameters to find the best results :

**hidden_layer_sizes
early_stopping
activation**

Multinomial Naive Bayes

It is a variant of the Naive Bayes algorithm designed for classification tasks where the features are assumed to follow a multinomial distribution. It is particularly suitable for text classification problems

Models Training

We planned to train the models based on different columns of the dataset: the review summary and the text review to see what works better

review/summary	review/text
Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
Really Enjoyed It	I don't care much for Dr. Seuss but after read...
Essential for every personal and Public Library	If people become the books they read and if "t...
Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka "D...
Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...

MLP Training Results

	precision	recall	f1-score
negative	0.75	0.77	0.76
neutral	0.60	0.60	0.60
positive	0.88	0.87	0.87
accuracy			0.80
macro avg	0.74	0.75	0.75
weighted avg	0.81	0.80	0.81

Naive bayes

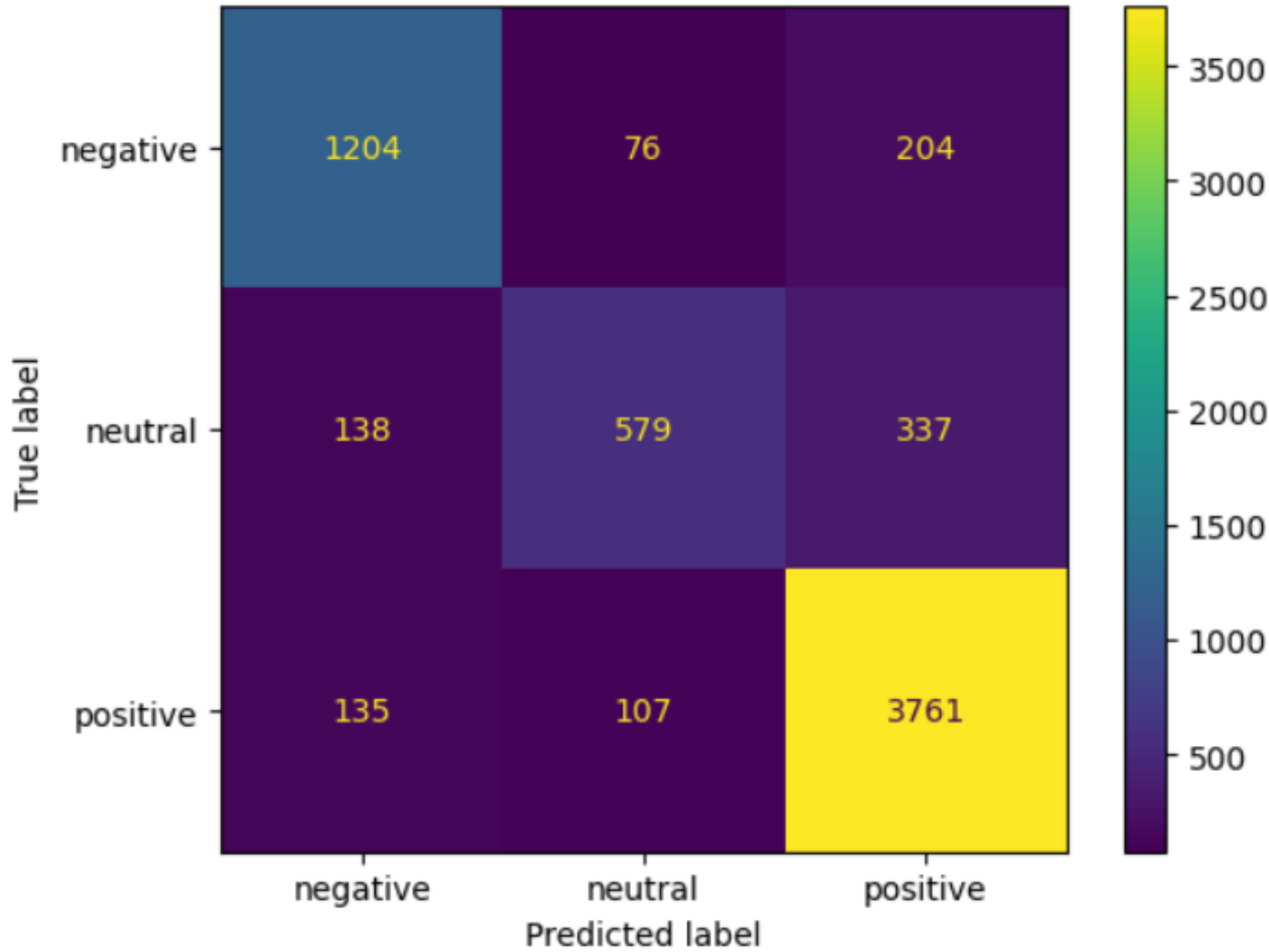
Training Results

Classification Report:				
	precision	recall	f1-score	
negative	1.00	0.04	0.07	
neutral	1.00	0.00	0.00	
positive	0.62	1.00	0.76	
accuracy			0.62	
macro avg	0.87	0.35	0.28	
weighted avg	0.77	0.62	0.48	

SVC Training Results

results for linear kernel function

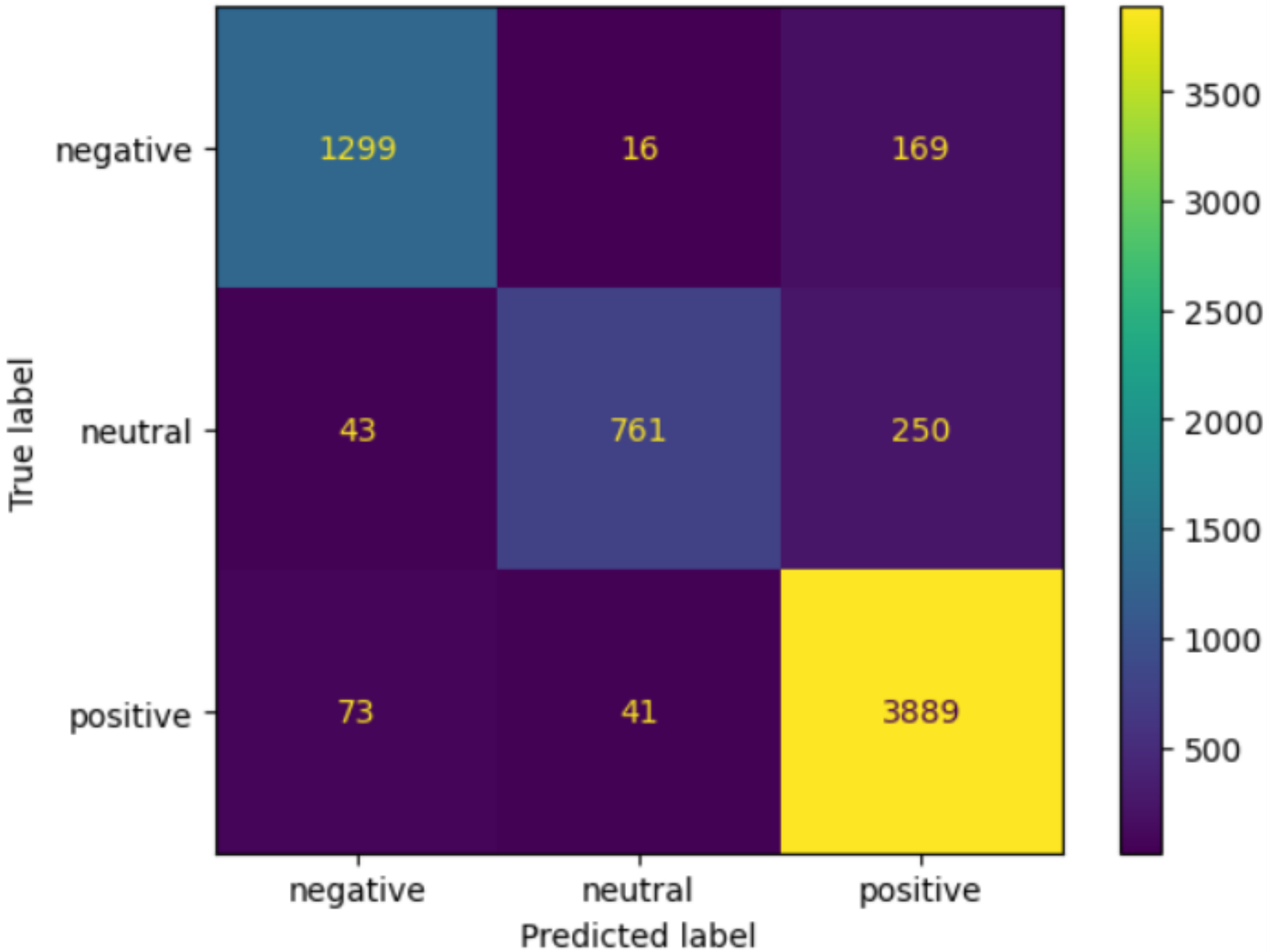
Rapport de classification pour les données d			
	precision	recall	f1-score
negative	0.93	0.92	0.92
neutral	0.92	0.77	0.84
positive	0.94	0.98	0.96
accuracy			0.93
macro avg	0.93	0.89	0.91
weighted avg	0.93	0.93	0.93



SVC Training Results

results for rbf kernel function

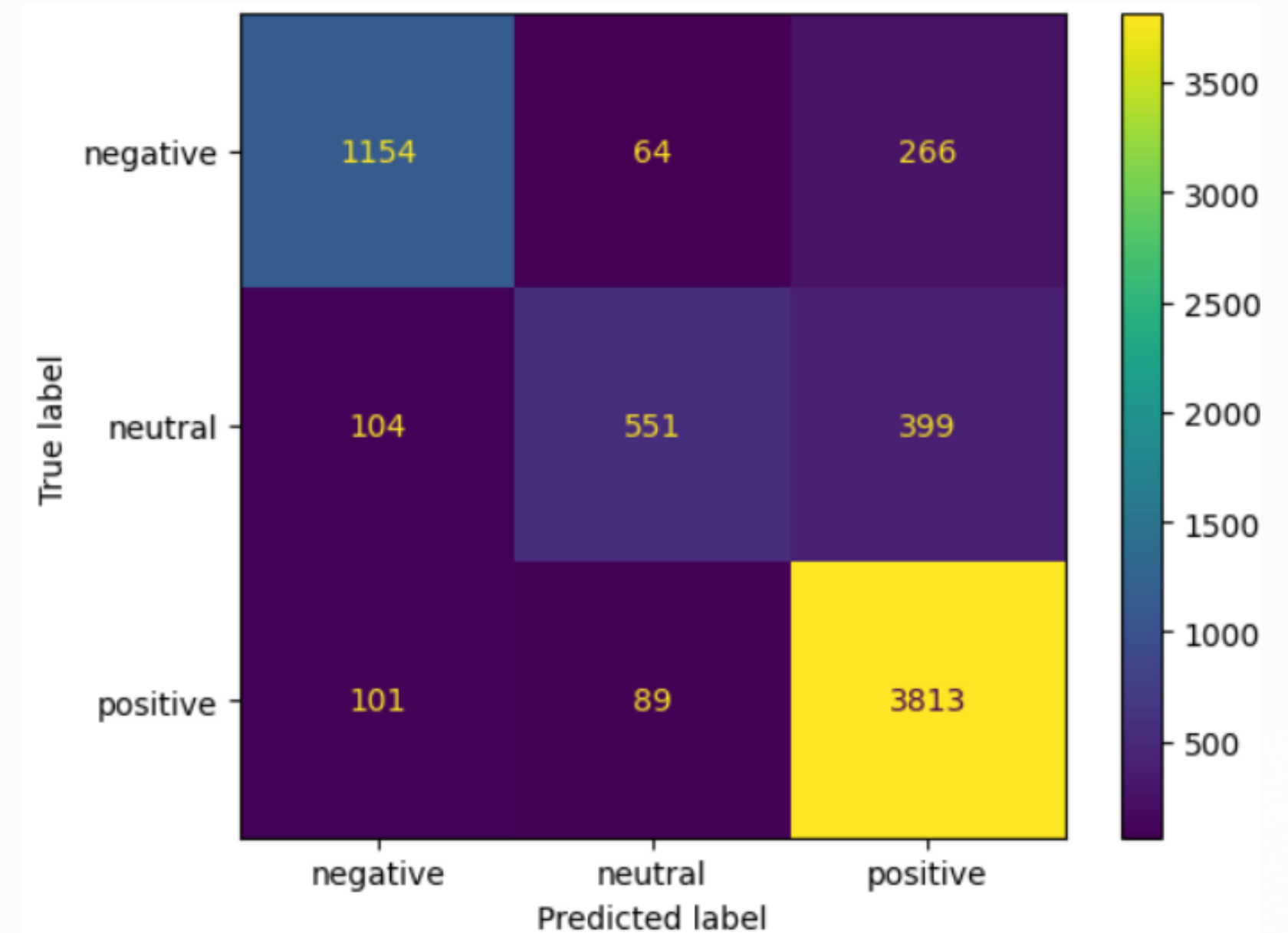
	precision	recall	f1-score
negative	1.00	0.99	0.99
neutral	1.00	0.96	0.98
positive	0.99	1.00	0.99
accuracy			0.99
macro avg	0.99	0.98	0.99
weighted avg	0.99	0.99	0.99



SVC training on the review summary

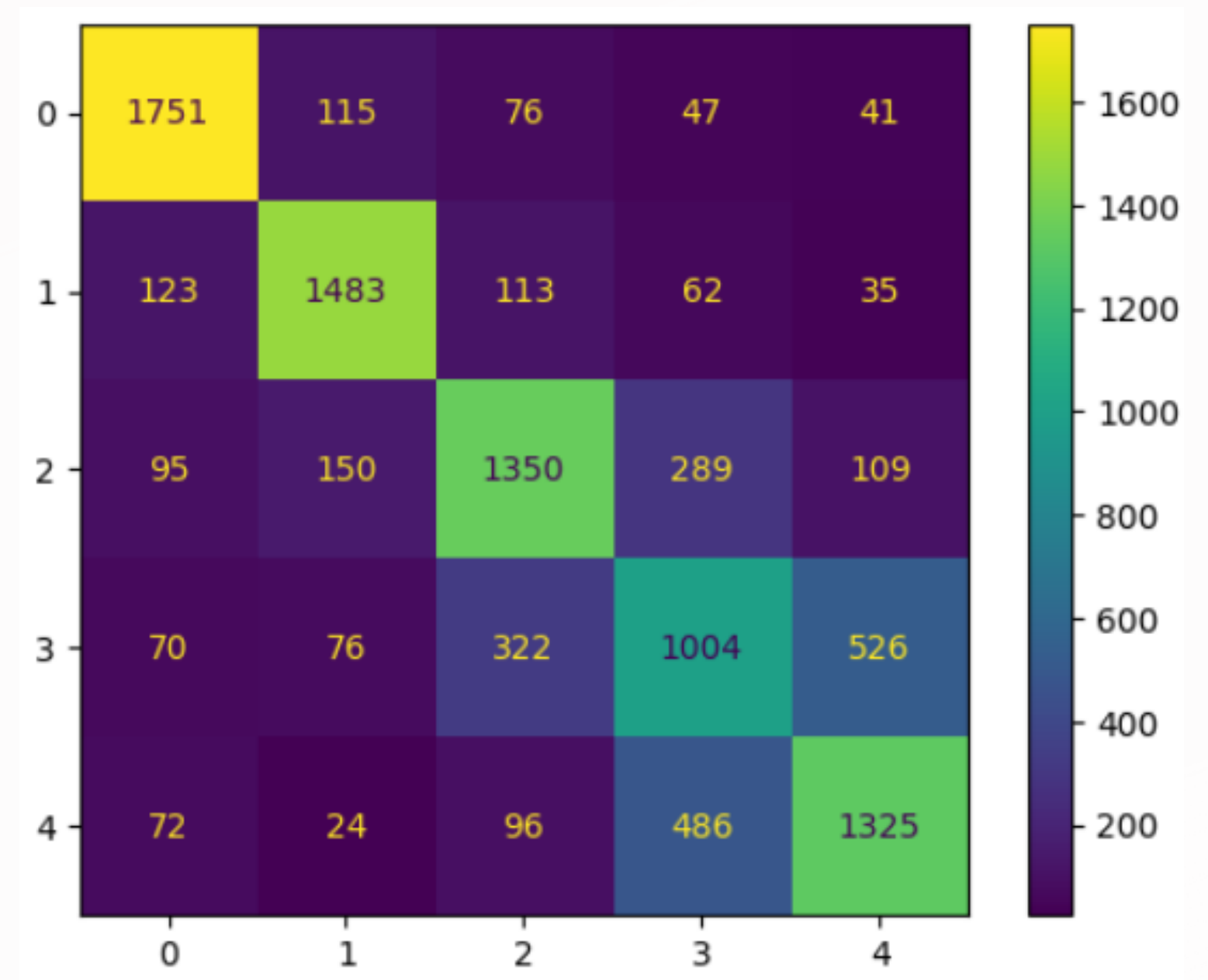
Classification Report for Training Data:

	precision	recall	f1-score
negative	0.95	0.93	0.94
neutral	0.94	0.81	0.87
positive	0.94	0.99	0.96
accuracy			0.95
macro avg	0.95	0.91	0.93
weighted avg	0.95	0.95	0.94



Rating classification with SVC

	precision	recall	f1-score
1.0	0.98	0.99	0.98
2.0	0.98	0.99	0.98
3.0	0.97	0.97	0.97
4.0	0.95	0.95	0.95
5.0	0.95	0.95	0.95
accuracy			0.97
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97



Discussion

Train on summary

- fast execution time
because less text
- good results
- **BUT :**

```
the review: Definitely Not For Women!  
the true rating: negative  
the model rating: neutral  
the review: Storyline ....  
the true rating: positive  
the model rating: neutral  
the review: Not a good study guide  
the true rating: negative  
the model rating: neutral
```

```
the review: disappointing read  
the true rating: neutral  
the model rating: negative  
the review: just lost interest 3/4 way through  
the true rating: neutral  
the model rating: negative
```

Discussion

Train on text review

- Longer execution time
but more accurate
results
- **BUT SOMETIMES:**

```
the review: I had to read this for a women's lit class. Book came very very quickly and in PERFECT shape and has large easy-to-read font. The actual book
the true rating: neutral
the model rating: positive
the review: Generally speaking a great book if you are not familiar with management accounting and turning heaps of information into valuable reports for
the true rating: negative
the model rating: positive
```

```
the review: Couldn't wait to finish but left me wondering what happened to so many of the characters. Thought I had missed something so went back and listened
the true rating: positive
the model rating: neutral
```

Conclusion

- sentiment analysis is complicated to guess because it's subjective
- we obtained good results , specially with SVC model
- training from the summary can be a good alternative in the event of a lack of time or data

**THANK
YOU**

any questions ?