SENTIMENT ANALYSIS IN ONLINE REVIEWS

OVERVIEW

- WHAT IS THE TASK?

- SCRAPE REVIEWS

    - How is the data?

- TEXT PREPROCESSING

    - Cleaning

    - Normalisation

- The VADER Lexicon

- SENTIMENT ANALYSIS MODELS

    - Features (Bag-of-words)

    - Approaches (SVM, NAIVE BAYES, RFC AND MLP)

    - Evaluation (Metrics, Comparison, Insights)

- BUSINESS IMPLICATIONS

The task

- Customer satisfaction in REBTEL

- Predict sentiment of reviews

- Classify rating 1-5 stars
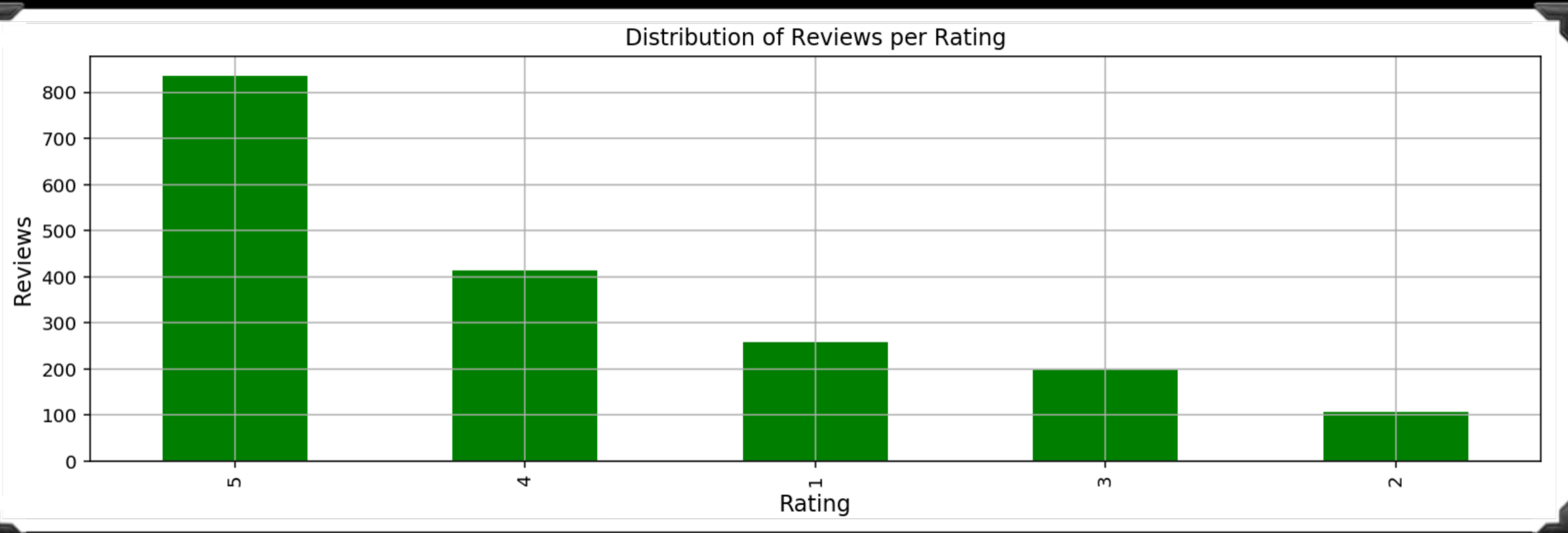
Scrape Reviews

- +2K reviews since 2011

- Information contained:

  - Review

  - Title

  - Rating

  - Time & Date

  - Country

  - Reviews

  - Reads

  - Useful
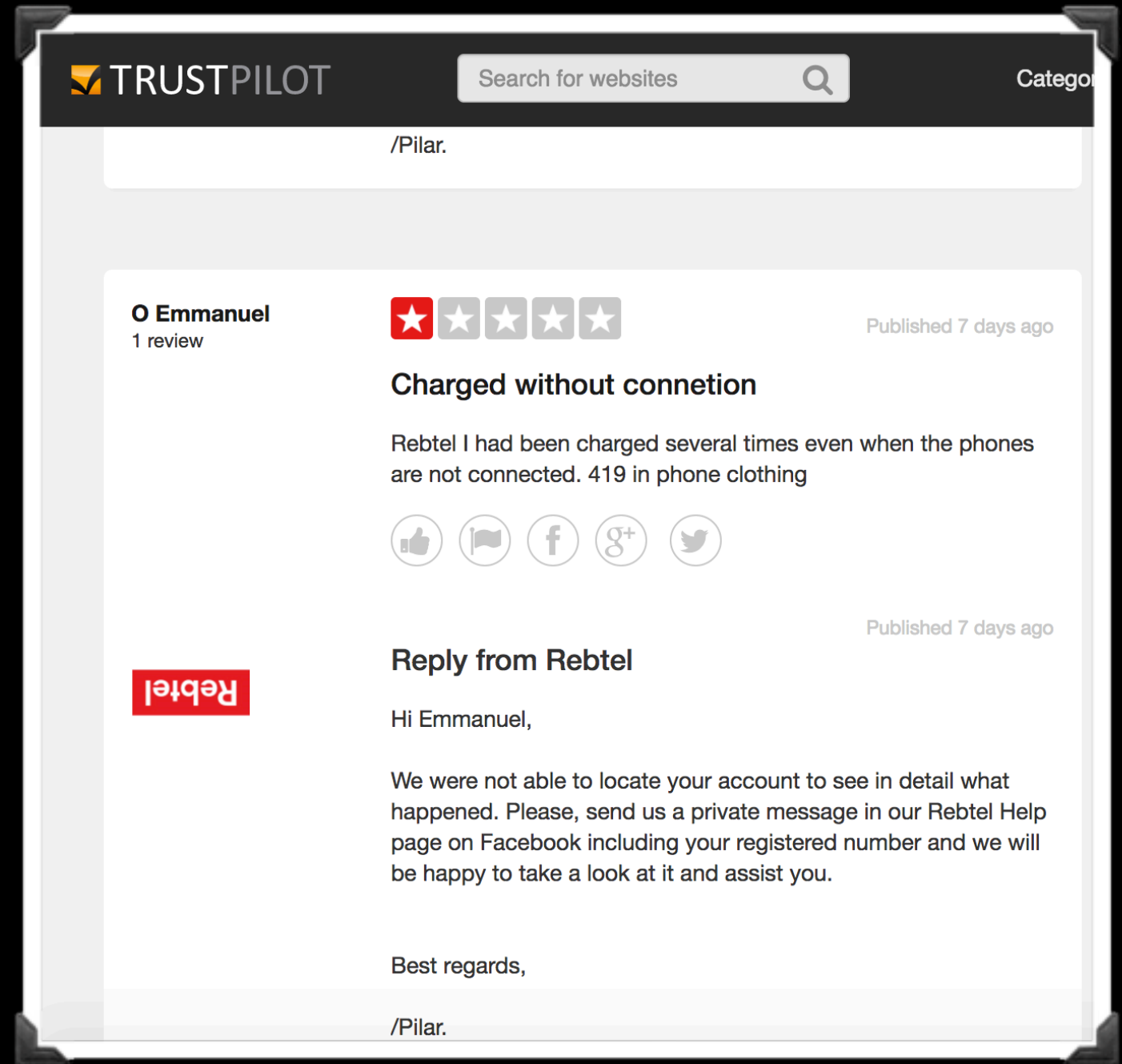
How is the data?

- REVIEW = TITLE + REVIEW + RATING

- Filter ENGLISH written reviews

- Rating 1-5 stars

- IMBALANCED CLASSES

Distribution of Reviews per Rating

Text preprocessing

- Tokenisation

- Stop words

- NOT = negative sentiment

- !!! = negative sentiment

- Placeholders: NUM, PRICE

- Check spelling

- Lemmatisation & Stemming

```python
print("This is an instance of the raw data: \n","Review : ",raw_data["Review"][0],
      "\n Title : ", raw_data["Title"][0])
```

```
This is an instance of the raw data:
 Review :   Rebtel I had been charged several times even when the phones are not connected. 419 in phone clothing
 Title :   Charged without connetion
```
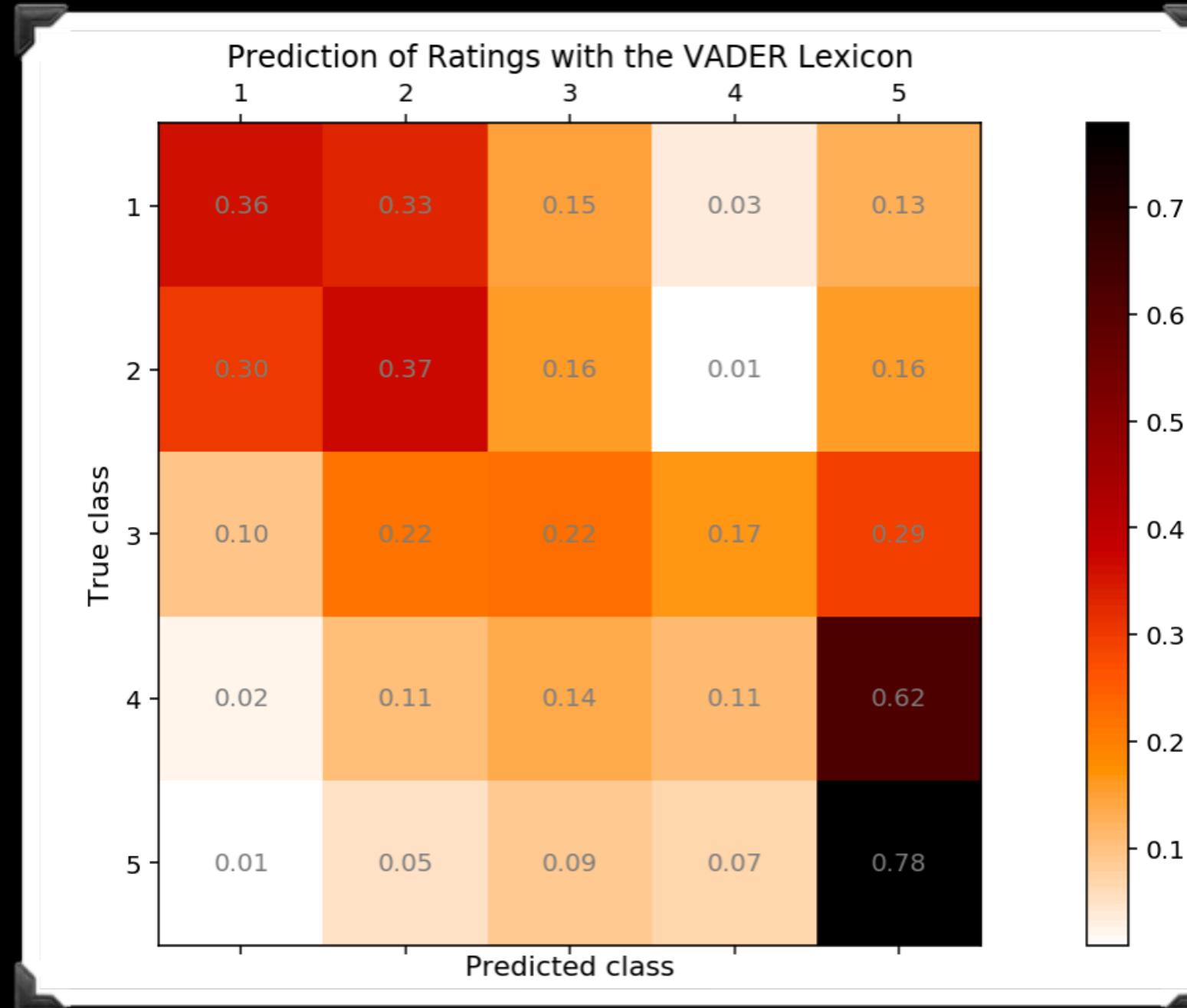
```python
print("This is the same instance preprocessed: \n","Review preprocessed: ",p_data["Rel_w_review"][0],
      "\n Title preprocessed : ",p_data["Rel_w_title"][0],
      "\n Title + Review with Stemming : ", p_data["Snow_title_review"][0])
```

```
This is the same instance preprocessed:
 Review preprocessed:   COMPANY charged several times even phones not !connected  NUM phone clothing
 Title preprocessed :   charged without connection
 Title + Review with Stemming :   charg without connect compani charg sever time even phone not ! connect num phone cl
oth
```

The VADER Lexicon

- Rule-based tool

- Sensitive: word-order relationships and punctuation marks

- Detects: SUX! and "Kinda"

- Measures:

  1. Compound

  2. POS, NEG, NEU

- Based on Lexical information

- Reviews with same original structure

- Limitations: Typos, Length Review, Unknown words

Prediction of Ratings with the VADER Lexicon

True class \ Predicted class	1	2	3	4	5
1	0.36	0.33	0.15	0.03	0.13
2	0.30	0.37	0.16	0.01	0.16
3	0.10	0.22	0.22	0.17	0.29
4	0.02	0.11	0.14	0.11	0.62
5	0.01	0.05	0.09	0.07	0.78

```
The misspelled title of the review is:  Theives!!!!
Its score is neutral:  {'neu': 1.0, 'compound': 0.0, 'neg': 0.0, 'pos': 0.0}
```

Sentiment Classifiers models

1. Features

- Bag-of-words representation

- Boolean, Term-frequency and Tf-idf

- N-gram models

- PCA: Most representative attributes

Sentiment Classifiers models

2. Imbalanced classes

- Weighted Samples

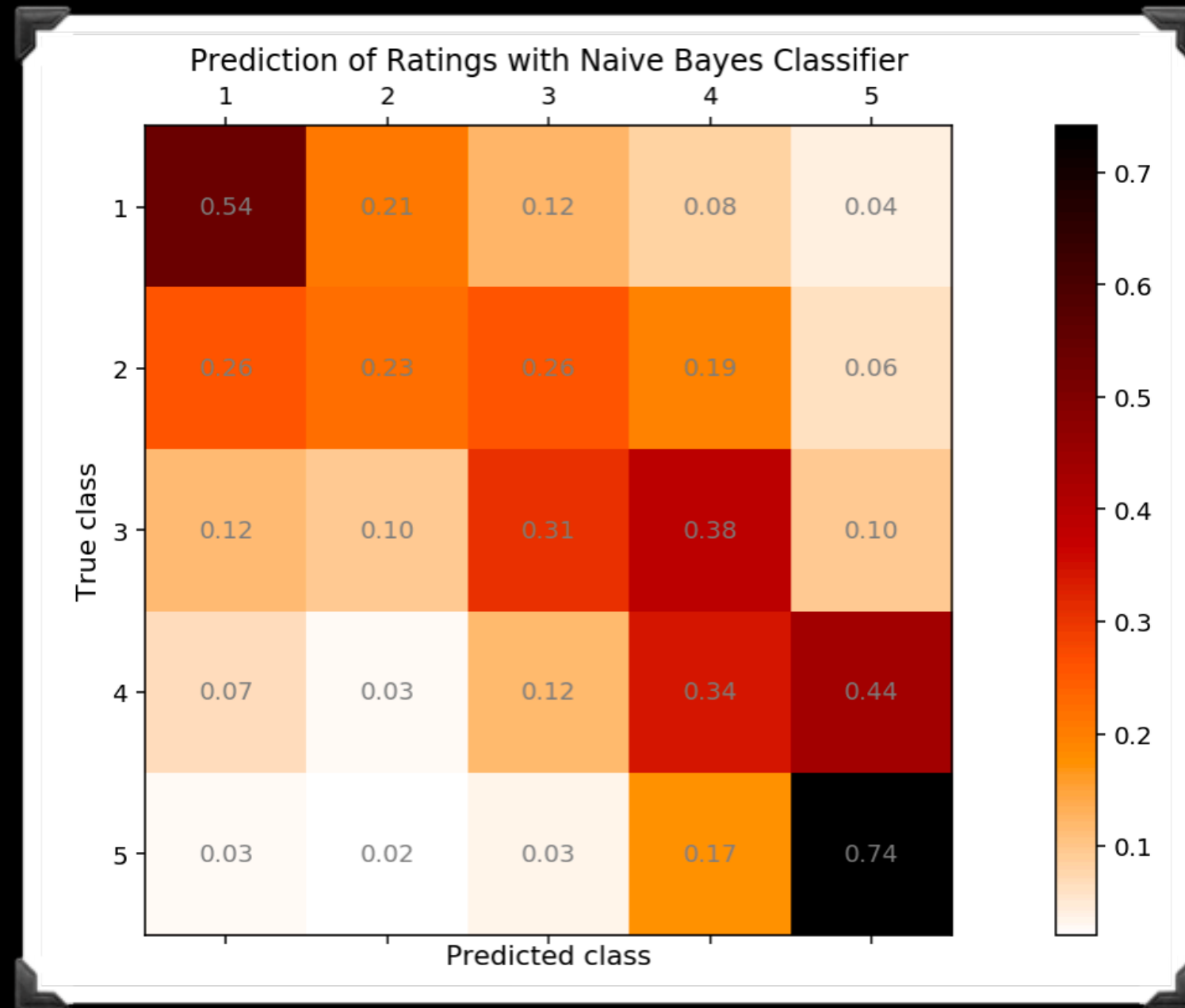- Oversampling (Random oversampling)

3. Model optimisation

- Greed Search + Cross-validation

- Bagging

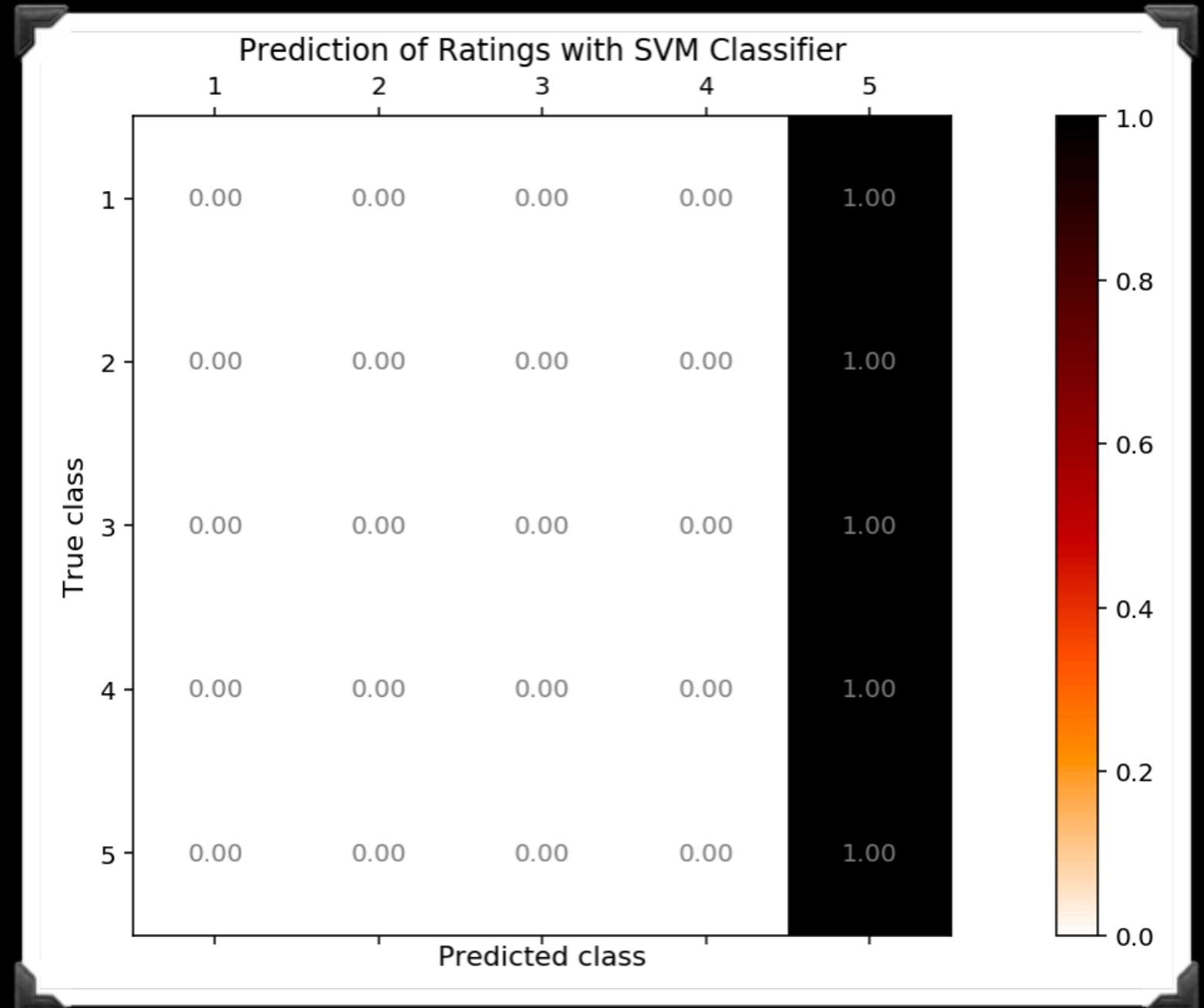## 4. Metrics

- Precision, Recall and F-score

- NOT Accuracy

Naive Bayes

- Multinomial NB

- Precision: 60%

- Recall: 58%

- F-Score: 59%

- Accuracy: 58%

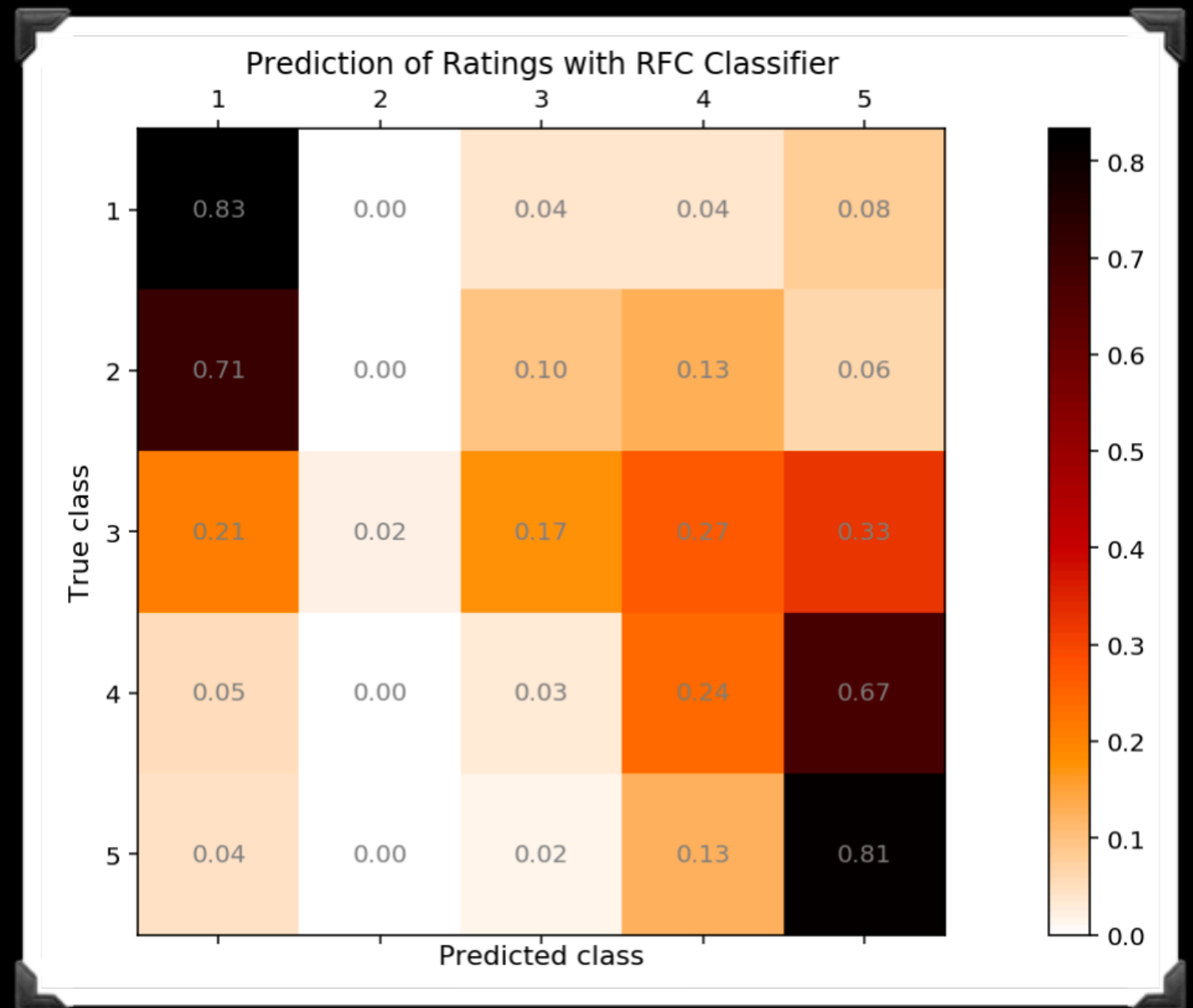Prediction of Ratings with Naive Bayes Classifier

SVM

- RBF Kernel

- Precision: 38%

- Recall: 62%

- F-Score: 47%

- Accuracy: 61%!

RFC

- 37 Trees, 245 words

- Precision: 55%

- Recall: 59%

- F-Score: 56%

- Accuracy: 58%

Prediction of Ratings with RFC Classifier

Multi-Layer Perceptron

- 3 Neurons, 64 layers,

- Optimiser: Categorical cross-entropy

- Precision: 62%

- Recall: 60%

- F-Score: 60%

- Accuracy: 59%

Prediction of Ratings with MLP Classifier
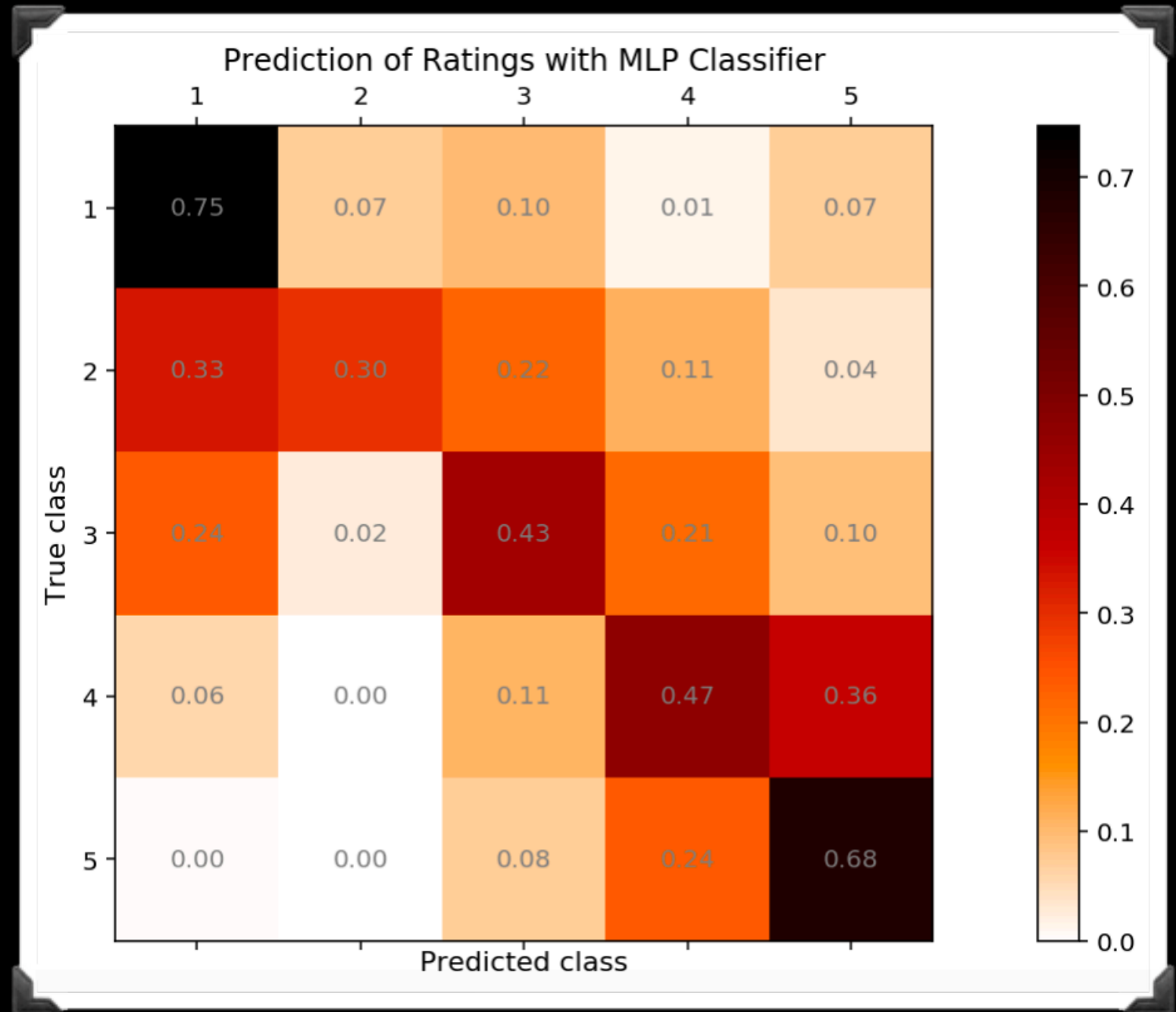
Sentiment Classifiers models

5. Evaluation

Model	Precision	Recall	F-Score	Accuracy
NB	60%	58%	59%	58%
SVM	38%	62%	47%	61%
RFC	55%	59%	56%	58%
MLP	62%	60%	60%	59%

6. Next Steps

- Reviews (2,3 and 4 stars) high level ambiguity even for humans!

- Manual data curation = better data quality

Business Implications

- Impact

- Tool direct contact with customers

- Holistic understanding of customers

- Customers post reviews voluntarily. Avoid predisposition to more synthetic surveys.

- Complement metrics for better business decisions

- Application

- Models updated constantly:

  - Detect Market dynamics (regions customers face problems)

- Extended for semantic analysis: (trends social media, compare competitors)

- Challenge: Understand evolution and expectation of customers.

Q&A

Thank you!