# Case Study Binary classification NLP problem for BRAIN ONE

Ronny Hein

19 Oktober 2019

#### Overview

#### Facts:

- ▶ 11 features
- ▶ 248252 observations
- ▶ 70 % test/train split
- ▶ 192426 observations in train set

#### Insights:

- overall is left-skewed
- ▶ target is imbalanced with 89.19 % positive

#### Visualisation

#### Postive

```
purchased fabric: Style work a style work aring a style work aring a style work aring a style work aring price ≥ large recommend big comfortable nice tight footleather comfortable nice tight footleather beautiful small day by a style work aring goodlight top wanted years boots Wear): 1-glove buy true product bought: 2-g pair super a black happy watch made bit #34 dress material shoes pretty
```

#### Visualisation

### Negative

```
buy amazun made wearing ordered #34 recommend love small disappointed price nice shoe of give pants? It has a looked order thin cheap fabric reviews $200 thing pair footwearlarge bad feet big fit black process feet big smaller back process for the smaller back 
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         uncomfortable
tight
```

#### Model

Bayes' theorem: 
$$p(C_k \mid \mathbf{x}) = \frac{p(C_k) \ p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Results on train set:

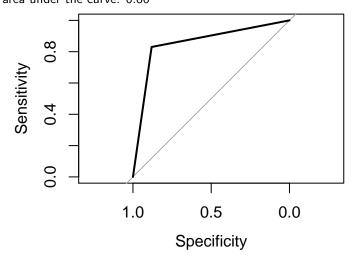
Table 1: actual (cols) versus predicted (rows)

_		
	-1	1
-1	78.97	1.13
1	10.22	9.67

#### **Evaluation**

accuracy of classifier on test: 87.55 %

true positive: 78.84 %
false negative: 10.67 %
area under the curve: 0.86



#### Considerations

- 1. more text preprocessing
- 2. use x-fold cross-valiation
- 3. apply more models and tune parameters
- 4. try more advanced NLP methods like BERT

## End

Thank you