Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

# Text Classification

**Text, Web and Social Media Analytics Lab**

**Prof. Dr. Diana Hristova**

# Exercise 4: Representation (2)

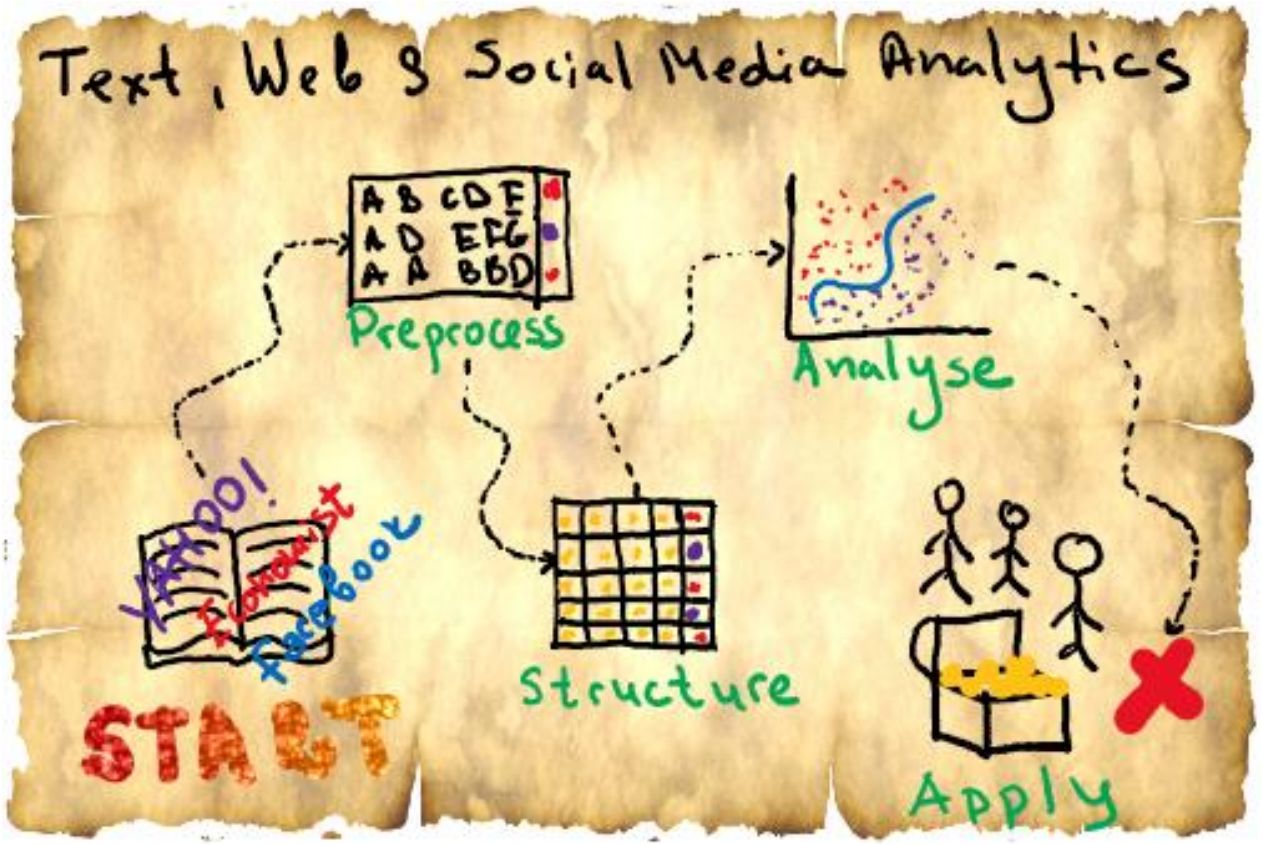
Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

**Can one group present please?**

# What did we learn last week?

# Treasury map: Text Classification

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law



# Course structure

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

| Date | Lecture | Exercise |
|---|---|---|
| 12.04.2021 | Introduction | Technical Installation |
| 19.04.2021 | Text Preprocessing | Projects kick-off |
| 26.04.2021 | Text Representation | Preprocessing Newsgroups |
| 03.05.2021 | Text Representation (2) | Text Representation Newsgroups |
| **10.05.2021** | **Text Classification** | **Text Representation Newsgroups (2)** |
| 17.05.2021 | Text Clustering | **Newsgroups Topic Classification** |
| 31.05.2021 | Text Mining in Social Media | Newsgroups Topic Clustering |
| 07.06.2021 | Mining Social Graphs | Sentiment Analysis and Time Series in Twitter |
| 14.06.2021 | Projects Status Update | Projects Status Update |
| 21.06.2021 | Web Analytics | Mining Social Graphs in Twitter |
| 28.06.2021 | Mock Exam | Web Analytics in E-commerce |
| 05.07.2021 | Final Presentation | Final Presentation |
| 19.07.2021 | Submit Code & Written report | |
| t.b.a. | Exam | |

# What will we learn today?

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

**At the end of this lecture, you will:**

1. Learn the motivation behind applying classification and clustering approaches in text analytics
2. Know the mechanics and pros and cons of important classification approaches for text analytics
3. Understand how to split the dataset to effectively train a classification model

# Motivation: Why do we need text classification and clustering?

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

https://www.wallpaperflare.com/search?wallpaper=review

- After cleaning and structuring text data, we can analyse it to derive knowledge from it.
- **Important questions:**
  - Is an e-mail a spam?
  - Is a movie review positive or negative?
  - What is the credit rating of a company based on its annual report?
  - Which are the most discussed topics on Twitter this week?
  - Which documents match best a search term?
  - How can I group Yelp reviews based on their topics?

# Motivation: Why do we need text classification and clustering?



## For which tasks is the set of values of the target variable known in advance?

a. **E-mail spam, movie review sentiment, annual report credit rating**
b. **Document search, hot topics on Twitter, topics in reviews**

# Introduction: Text Classification and Clustering

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

## Supervised approaches

- **Idea:** the set of values of the target variable is known and the text data is annotated with them.
- **Example:**
  - "I require your financial assistance to transform you the heritage of the late Sir Johnson." ➔ spam
  - "We are meeting tomorrow at 5 o'clock." ➔ not spam
  - Aim: build a model that can classify each document in one of the categories.

➔ **Text classification**

## Unsupervised approaches

- **Idea:** the set of values of the target variable are not known, but need to be derived from the data.
- **Example:**
  - A: "The restaurant has terrible service."
  - B: "The meals were very satisfactory.
  - C: "The food was fantastic."
  - D: "The waitress took forever."

➔ A and D are similar as well as B and C

- Aim: build a model that can derive the unknown categories in each document.

➔ **Text clustering**

# Introduction: Text Classification and Clustering

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

**Focus today**

## Supervised approaches

- **Idea:** the set of values of the target variable is known and the text data is annotated with them.
- **Example:**
  - "I require your financial assistance to transform you the heritage of the late Sir Johnson." ➔ spam
  - "We are meeting tomorrow at 5 o'clock." ➔ not spam
  - Aim: build a model that can classify each document in one of the categories.

➔ **Text classification**

## Unsupervised approaches

- **Idea:** the set of values of the target variable are not known, but need to be derived from the data.
- **Example:**
  - A: "The restaurant has terrible service."
  - B: "The meals were very satisfactory.
  - C: "The food was fantastic."
  - D: "The waitress took forever."

➔ A and D are similar as well as B and C

- Aim: build a model that can derive the unknown categories in each document.

➔ **Text clustering**

# Text Classification: Main Idea

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

## 1. Annotated corpus for training

| | content | target | target_names |
|---|---|---|---|
| 0 | From: lerxst@wam.umd.edu (where's my thing)\nS... | 7 | rec.autos |
| 1 | From: guykuo@carson.u.washington.edu (Guy Kuo)... | 4 | comp.sys.mac.hardware |
| 2 | From: twillis@ec.ecn.purdue.edu (Thomas E Will... | 4 | comp.sys.mac.hardware |
| 3 | From: jgreen@amber (Joe Green)\nSubject: Re: W... | 1 | comp.graphics |
| 4 | From: jcm@head-cfa.harvard.edu (Jonathan McDow... | 14 | sci.space |

## 2. Preprocessed corpus for training

| target | target_names | preprocessed |
|---|---|---|
| 7 | rec.autos | lerxst wam umd edu thing subject car nntp post... |
| 4 | comp.sys.mac.hardware | guykuo carson washington edu gui kuo subject c... |

# Text Classification: Main Idea (2)

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

## 3. Structure corpus for training

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 55273 | 55274 | 55275 | 55276 | 55277 | 55278 | 55279 | 55280 | 55281 | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4 |

X

y

Classification approach

Model

$y = f(X)$

The Fiat 500 EV has been revealed and, despite being on a whole new platform, the car looks very similar to before. There are some neat new styling touches, though, and the 500 EV will offer a range of 199 miles.

https://www.carbuyer.co.uk/news/167673/best-new-cars-coming-in-2020

rec.autos

# Text Classification: Main Idea (2)



**Which classification models have you used?**

# Classification approaches: Naïve Bayes

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

## Naive Bayes (NB)

- **Main idea:** use Bayes theorem to determine which words contribute to the probability of which class.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- For a given document representation $doc_1 = (v_1, v_2)$ and two possible classes $A$ and $B$, determine:

$$max_{x \epsilon A,B} P(class\ x|v_1, v_2) = max_{x \epsilon A,B} P(class\ x|v_1) P(class\ x|v_2)$$
$$= max_{x \epsilon A,B} (P(v_1|class\ x) P(v_2|class\ x) P(class\ x)^2) / P(v_1) P(v_2)$$
$$\Leftrightarrow max_{x \epsilon A,B} P(v_1|class\ x) P(v_2|class\ x) P(class\ x)^2$$

➔ Assign the document to the class with the higher probability, assuming independence between the features.

? How valid is this assumption?

# Classification approaches: Naïve Bayes



## Naive Bayes (NB)

- **Main idea:** use Bayes theorem to determine which words contribute to the probability of which class.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- For a given document representation $doc_1 = (v_1, v_2)$ and two possible classes $A$ and $B$, determine:

$$max_{x \epsilon A,B} P(class\ x|v_1, v_2) = max_{x \epsilon A,B} P(class\ x|v_1) P(class\ x|v_2)$$
$$= max_{x \epsilon A,B} (P(v_1|class\ x) P(v_2|class\ x) P(class\ x)^2) / P(v_1) P(v_2)$$
$$\Leftrightarrow max_{x \epsilon A,B} P(v_1|class\ x) P(v_2|class\ x) P(class\ x)^2$$

➔ Assign the document to the class with the higher probability, assuming independence between the features.

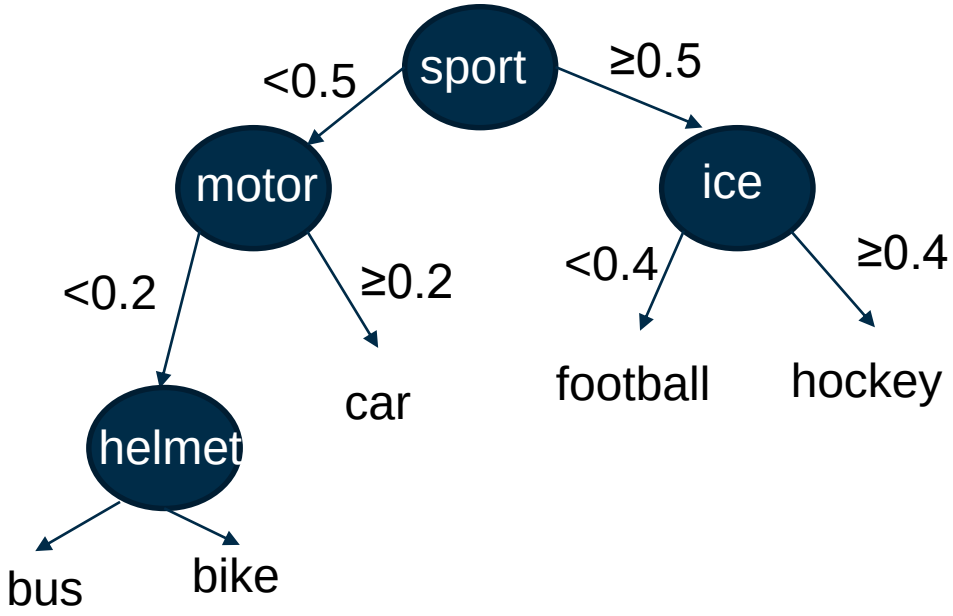
- **Advantage:** simple, quick and interpretable
- **Disadvantage:** strong assumptions, often not fulfilled

# Classification approaches: Decision Trees

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

## Decision Trees (DT)

- **Main idea:** build a tree which nodes are the features and branches are conditions on their values. The leaves represent the classes of the target.



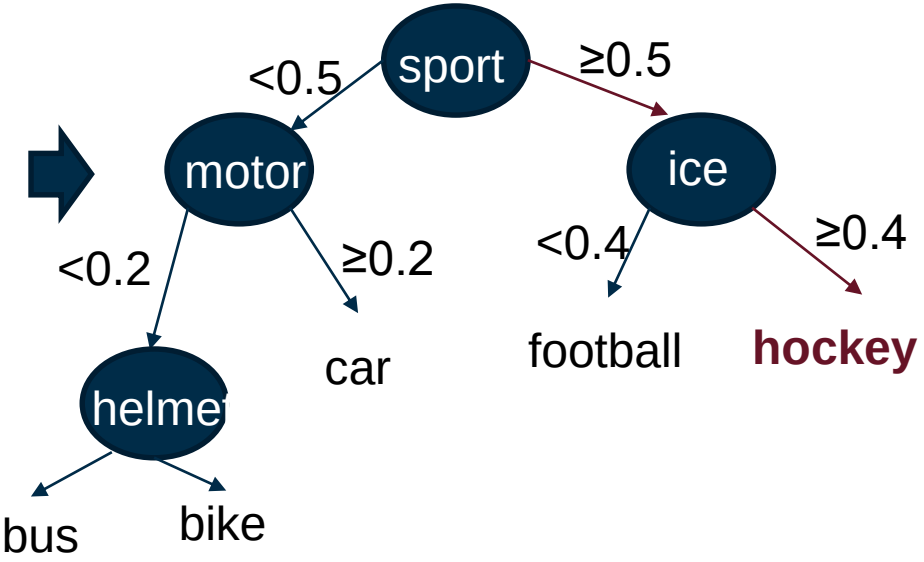
# Classification approaches: Decision Trees (2)



## Decision Trees (DT)

- A new document is classified by following the tree paths.

| sport | motor | ice | helmet |
|---|---|---|---|
| 0.6 | 0.1 | 0.5 | 0.3 |




- **Advantage:** Efficient and interpretable
- **Disadvantage:** Overfitting

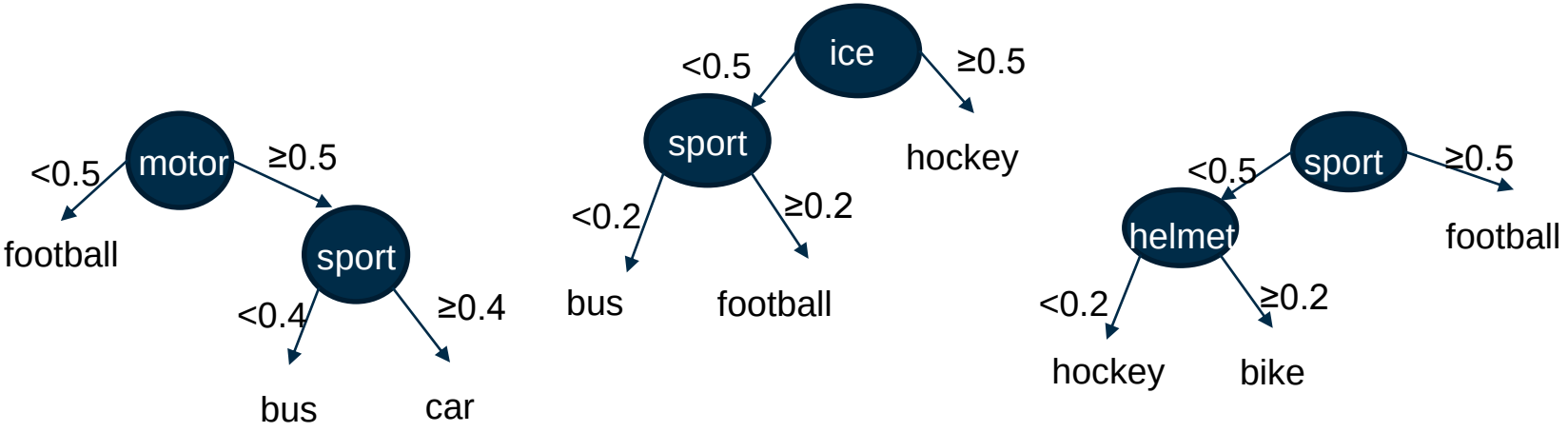
# Classification approaches: Random Forests

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

## Random Forest (RF)

- **Main idea:** combine multiple trees together and average their result.
- Each tree is build on a subsample of data and features and thus too deep trees are avoided.

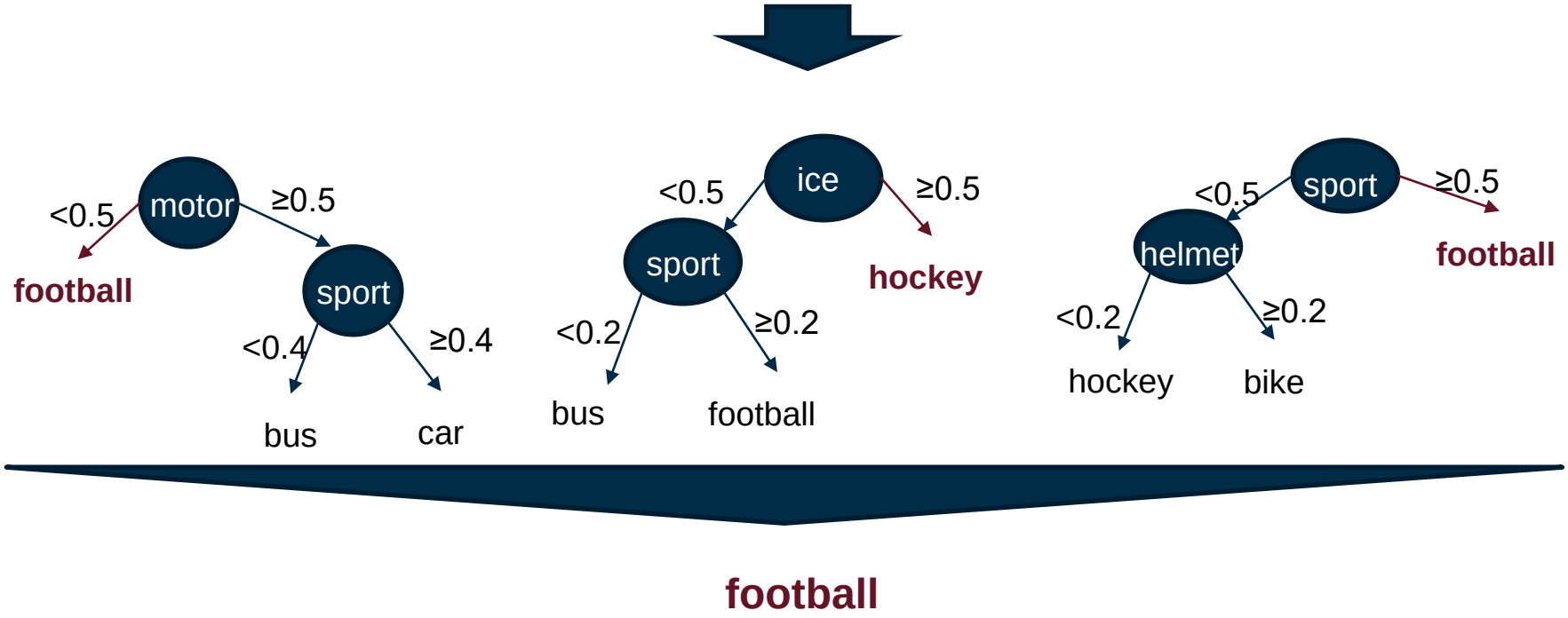
➔ Less overfitting



# Classification approaches: Random Forests (2)



| sport | motor | ice | helmet |
|---|---|---|---|
| 0.6 | 0.1 | 0.5 | 0.3 |



# Classification approaches: Random Forests (3)

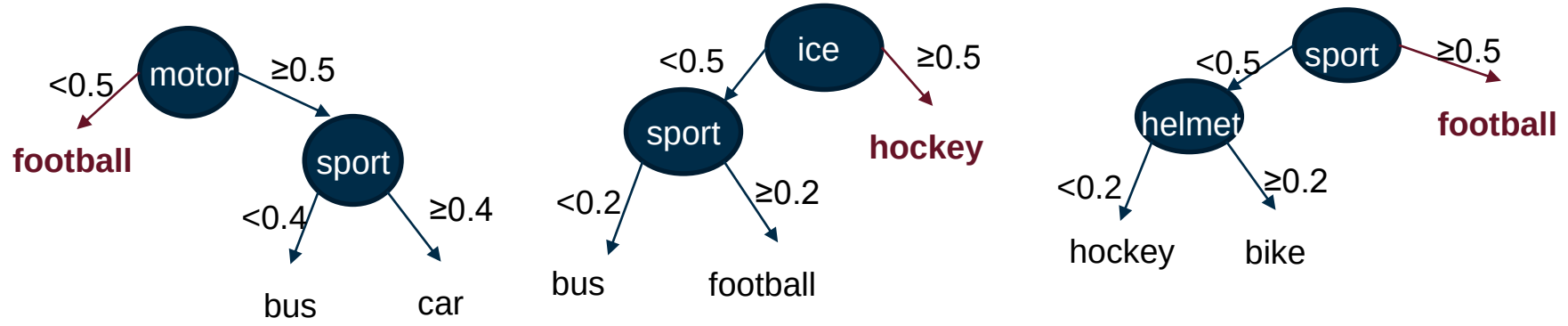


## What is a disadvantage of random forests?

A. Not efficient

B. Problems with complex non-linear relationships

# Classification approaches: Random Forests (3)




- **Advantage:** Efficient and (somehow) interpretable
- **Disadvantage:** Not able to model complex, non-linear relationships

# Text Classification: Training a model



- The classification model should be chosen such that it:
  1. **Performs best** among all alternatives
  2. Does so on **unseen** data

? How do you achieve 1. and 2.?

# Text Classification: Training a model



- The classification model should be chosen such that it:
  1. **Performs best** among all alternatives
  2. Does so on **unseen** data

? How do you achieve 1. and 2.?

- ✓ Train the model on a training set, test it on an independent test set
- ✓ Use a validation set to determine the best model and parameters

# Training a model: Validation set



## Validation set

- Most of the classification models have different parameters that can be set.
- For example, in a Random Forest the following (hyper)parameters can vary:
  - Number of trees
  - Splitting criterion
  - Maximum tree depth
  - Minimum samples for splitting
  - Minimum samples in a leaf
  - ….
- Using the default setting rarely provides the best performance.
- How do you choose the best hyperparameters?

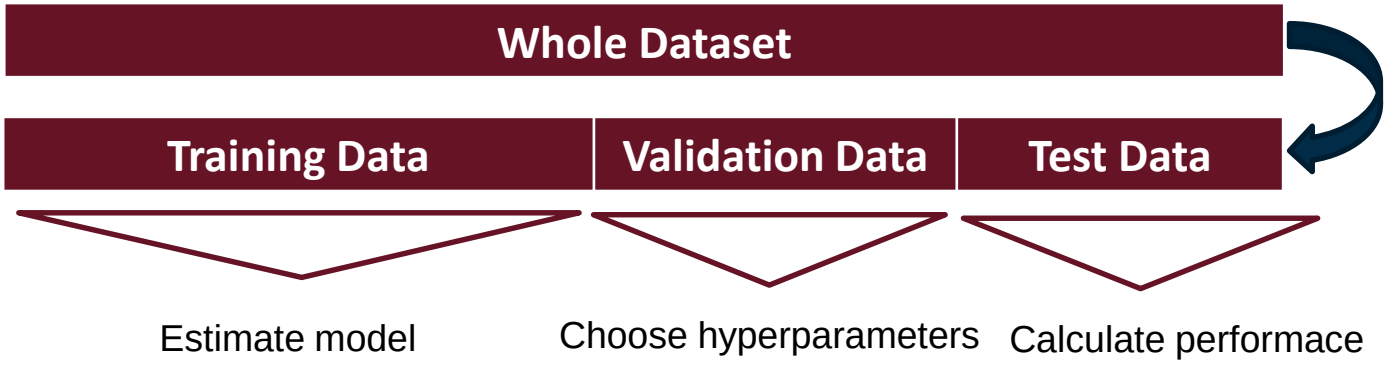
➔ Validation set

# Training a model: Validation set (2)



## Validation set

- One possibility is to randomly split the training set:



# Training a model: Validation set (3)



**What is a disadvantage of this approach?**

# Training a model: Cross validation

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

## Cross validation

- Sometimes the data set is too small to be able to conduct a Training/Validation split i.e. the Training set is too small and the model cannot be adequately estimated.

➔ Cross validation

- K-fold cross validation splits randomly the Training set into k parts.
- First k-1 parts are used for training, last part for hyper parameter estimation
- This is repeated until each part was used once for hyper parameter estimation

➔ k accuracies

- The best hyperparameters are the ones with the highest average performance

# Summary and Outlook



## Summary:

- Classification approaches can be used to conduct supervised learning on text representation
- The Naive Bayes approach is based on the Bayes‘ Theorem
- Decision Trees and Random Forests use tree representation to split the solution space
- Model training should consider the performance on the validation and test set
- Cross validation can be applied for small datasets
- **Outlook:** many text datasets are not labelled ➔ clustering

# Questions?

# Exercise 5



In a minute, six break-out rooms will be created. Choose the room that corresponds to your group in Moodle e.g. Room 1= Group 1. In your project group discuss and document the solution for Exercise 5 (in Moodle).