gesis Leibniz-Institut für Sozialwissenschaften

Multi-class and Multi-label Text Classification in Python

Lukas Birkenmaier

Workshop for Ukraine, 22.02.2024

Leibniz Gemeinschaft

About me

RWTH AACHEN UNIVERSITY

Universität Konstanz

UNIVERSITAS WARWICENSIS

Maastricht University

gesis Leibniz Institute for the Social Sciences

pwc

Baden-Württemberg MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND WOHNUNGSBAU

UNIVERSITÄT DUISBURG ESSEN Offen im Denken

gesis Leibniz-Institut für Sozialwissenschaften

Agenda

- Lecture (30 min)
  - Introduction
    - **Multiclass** Classification
    - **Multilabel** Classification
  - Overview Methods
    - Traditional Approaches
    - Large Language Models (LLMs)
  - Annotation
  - Validation
- Break and Notebook Setup (20 min)
- Pratical Application (1h)
  - Live-Coding and individual exercises

Disclaimer

- This is an applied course!
  - Some knowledge of Python is useful. However, you can run the scripts and interpret the output without any python knowledge
  - No mathematical deep-dive, focus on applied setting
  - In the practical application: Focus on state-of-the-art LLMs
  - Especially in the second part, you are invited to go your own pace (e.g., adapting the scripts to a new dataset)
    - ***The materials are designed so that you can look at them later and copy / paste certain code snippets for your own work***

Introduction

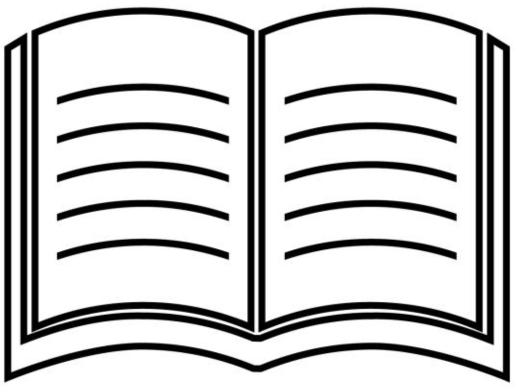
gesis Leibniz-Institut für Sozialwissenschaften

Relevance

- Text classification is an extremely popular task
- Subfield of Natural Language Processing (NLP), one of the subfields of artificial intelligence
- The term "text classification" usually refers to methods of (supervised) machine learning
- Many practical applications
  - Spam filtering
  - Hate Speech Detection
  - Language Identification
  - Meta Data Generation
  - Sentiment Analysis

https://lena-voita.github.io/nlp_course/text_classification.html#main_content

Text Classification



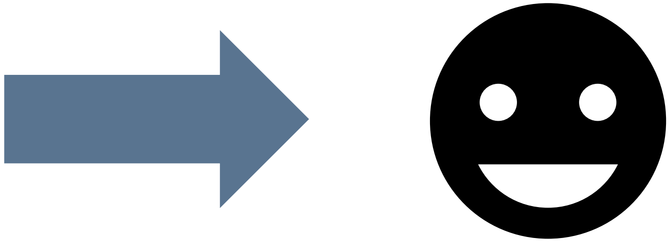
Text Classification

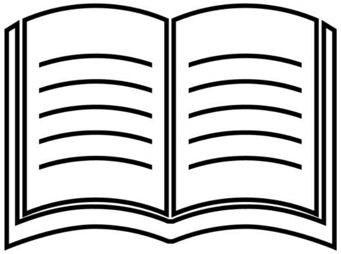
Text Classification

"Classification is the process of accurately classifying previously undiscovered data"

Text Classification

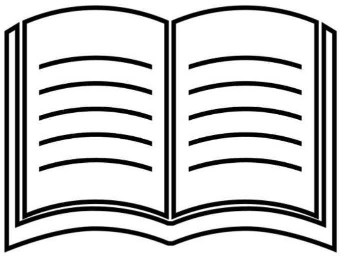
"I am so happy about my life"





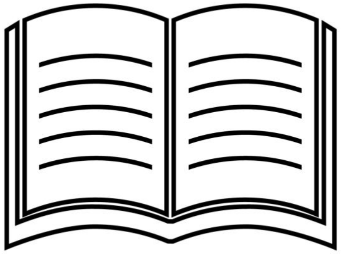
Text Classification

“I am so sad about my life”



Text Classification

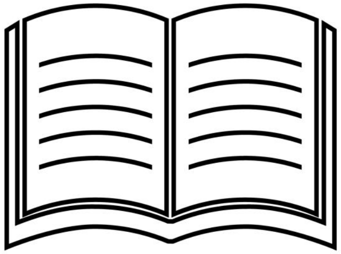
“We should protect our ecosystem”



Text Classification

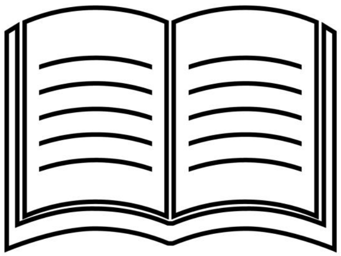
Topic: Environment

"We should protect our ecosystem"



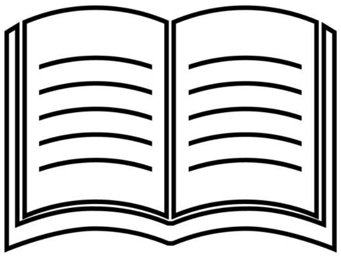
Text Classification

“Digital Technologies can help children to understand our ecosystem better”



Text Classification

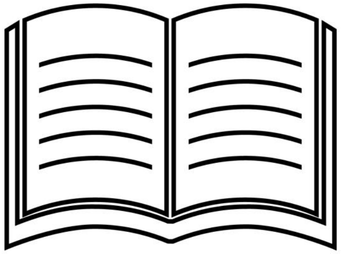
"Digital Technologies can help children to understand our ecosystem better"



Topic: Environment?

Text Classification

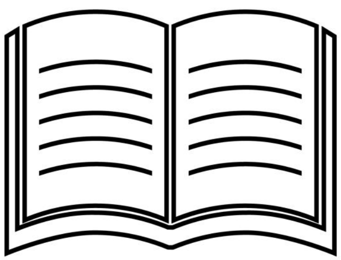
"Digital Technologies can help children to understand our ecosystem better"



Topic: Digitization?

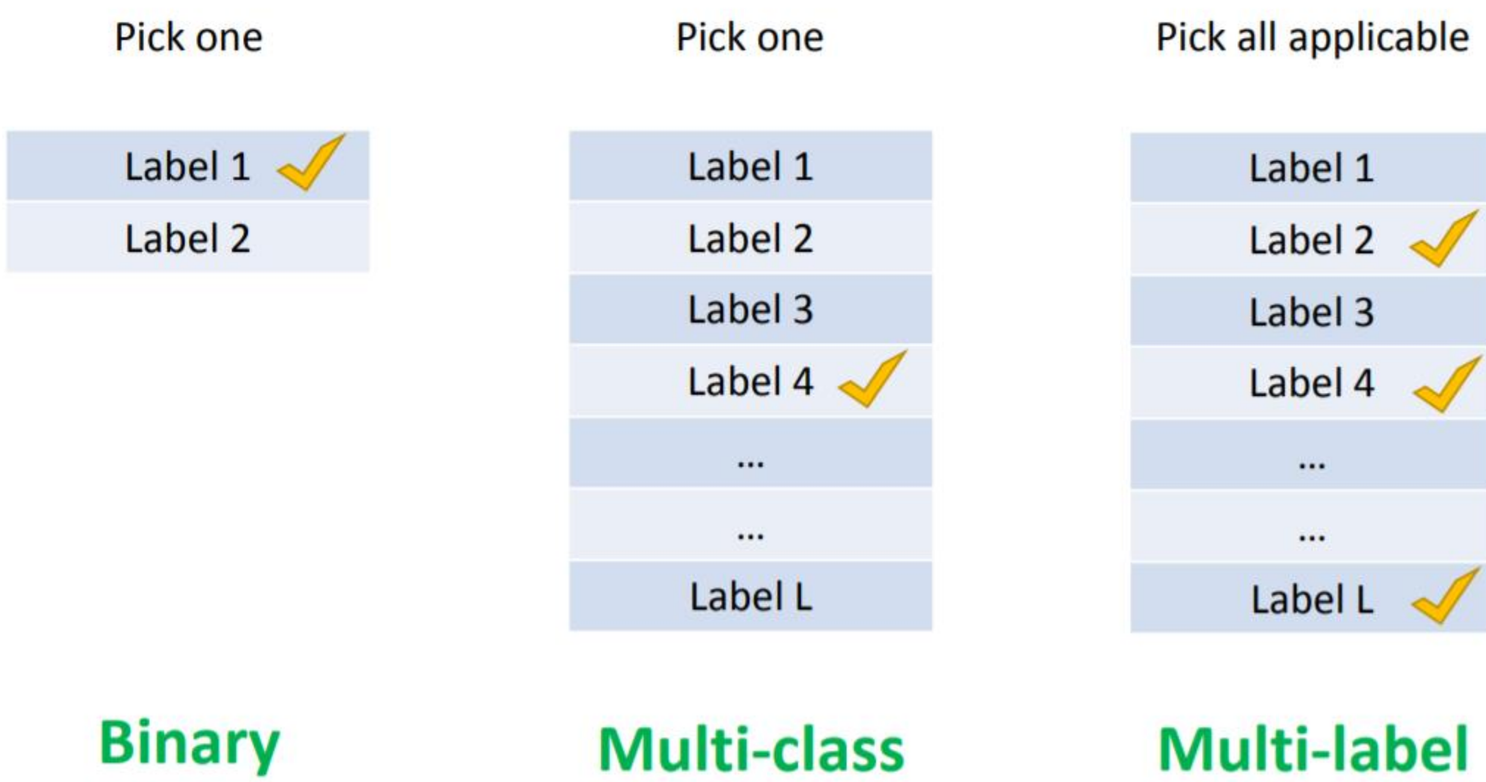
Text Classification

"Digital Technologies can help children to understand our ecosystem better"



Topic: Education?

General Overview Classification



https://harshjadhav100.medium.com/multiclass-classification-vs-multi-label-classification-a37cbe1627e

Definitions

- **Multiclass classification** is a classification task with more than two classes where **each sample can only be labeled** as one class.
- Labels are mutually exclusive
- E.g., classification of the dominant topic in a text

gesis Leibniz-Institut für Sozialwissenschaften

Definitions

- **Multiclass classification** is a classification task with more than two classes where **each sample can only be labeled** as one class.
- Labels are mutually exclusive
- E.g., classification of the dominant topic in a text

- **Multilabel classification** is a classification task labeling each sample **with m labels from $n_{classes}$, with $0 \leq m \leq n_{classes}$**
- Labels are not mutually exclusive
- E.g., classification of all the topics that appear in a text

https://scikit-learn.org/stable/modules/multiclass.html#multilabel-classification
https://www.analyticsvidhya.com/blog/2021/07/demystifying-the-difference-between-multi-class-and-multi-label-classification-problem-statements-in-deep-learning/

Definitions

- **Multiclass classification** can be seen as an exten-sion of binary classification
- Requires (usually) no problem transformation
- Probabilities for each class add up to 100%
- Evaluation via accuracy, precision, recall, F1 score (and some general internal validation/error analysis, see here)

https://scikit-learn.org/stable/modules/multiclass.html#multilabel-classification

gesis Leibniz-Institut für Sozialwissenschaften

Definitions

- **Multiclass classification** can be seen as an extension of binary classification
- Requires (usually) no problem transformation
- Probabilities for each class add up to 100%
- Evaluation via (average) accuracy, precision, recall, F1 score (and some general internal validation/error analysis, see [here](#))

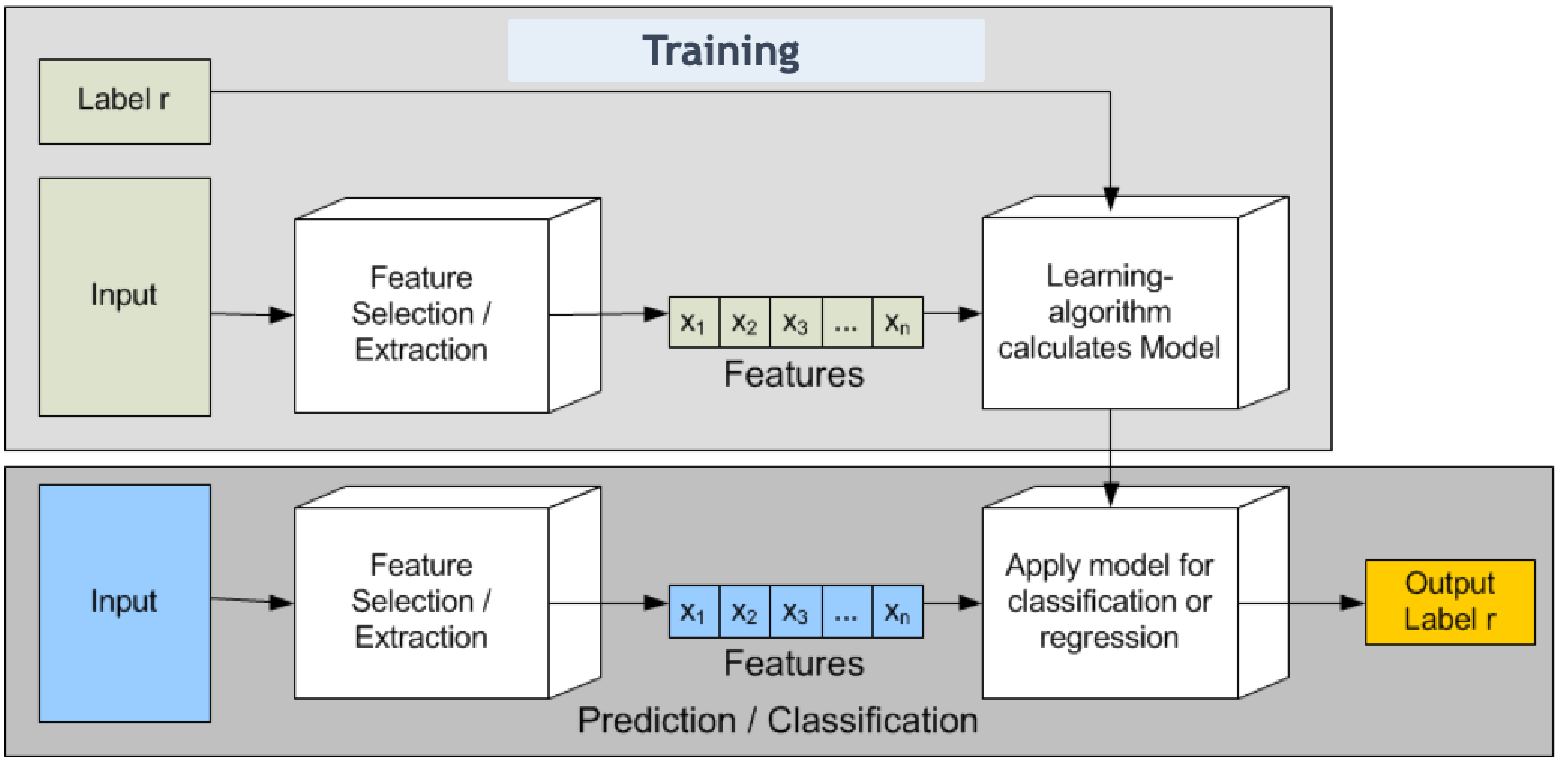
- **Multilabel classification** can be seen as an extension of multiclass classification
- Can require problem transformation
- Separate probabilities for each output class
- Evaluation via (average) accuracy, precision, recall, F1 score, or specific metrics such as Hamming Loss

https://scikit-learn.org/stable/modules/multiclass.html#multilabel-classification

Overview Methods

gesis Leibniz-Institut für Sozialwissenschaften

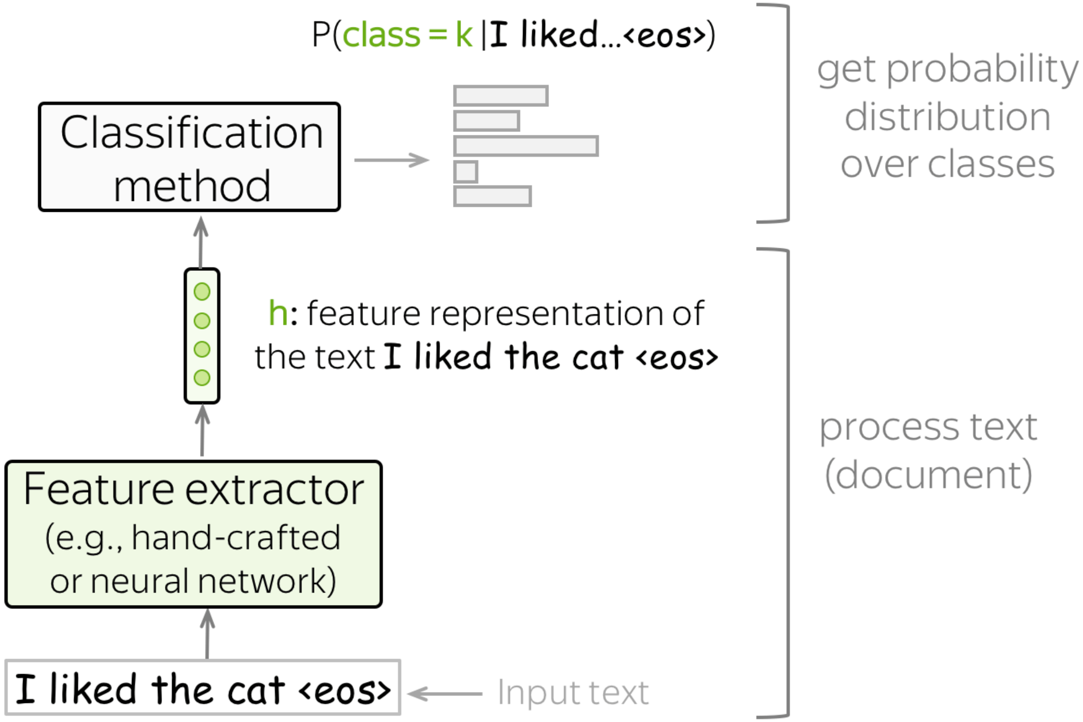
Basic Idea behind Machine Learning



https://griesshaber.pages.mi.hdm-stuttgart.de/nlp/06classification/06classification.html

Text classifiers have the following structure

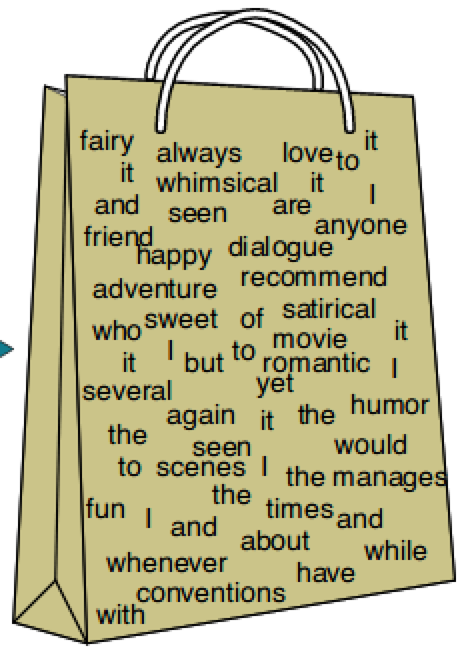
- **Feature extractor**
  - Makes the text machine-readable
  - Either manually defined or learned (e.g., with neural networks)
  - **Same** for **multiclass** and **multilabel** classification
- **Classifier**
  - Assigns class probabilities given feature representation of a text
  - **Different** for **multiclass** and **multilabel** classification



https://lena-voita.github.io/nlp_course/text_classification.html#general_view

Feature Extraction: Bag of Words

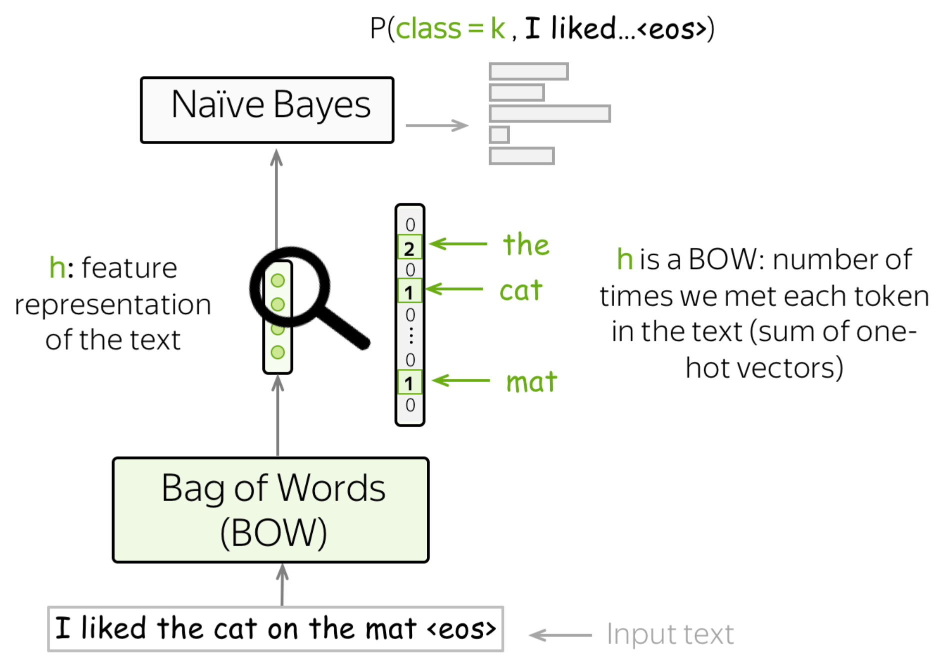
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Feature Extraction: Bag of Words

- One-Hot Encoding
- Assumption: word order does not matter
- Limitations
  - Discarding word context
  - Discarding grammatical structure
  - Vocabulary inconsistencies (e.g., grammatical errors, conjunctions)
  - Computationally inefficient (sparse matrix with most elements being 0)
- …



gesis Leibniz-Institut für Sozialwissenschaften

Classification Method

Feature Extraction

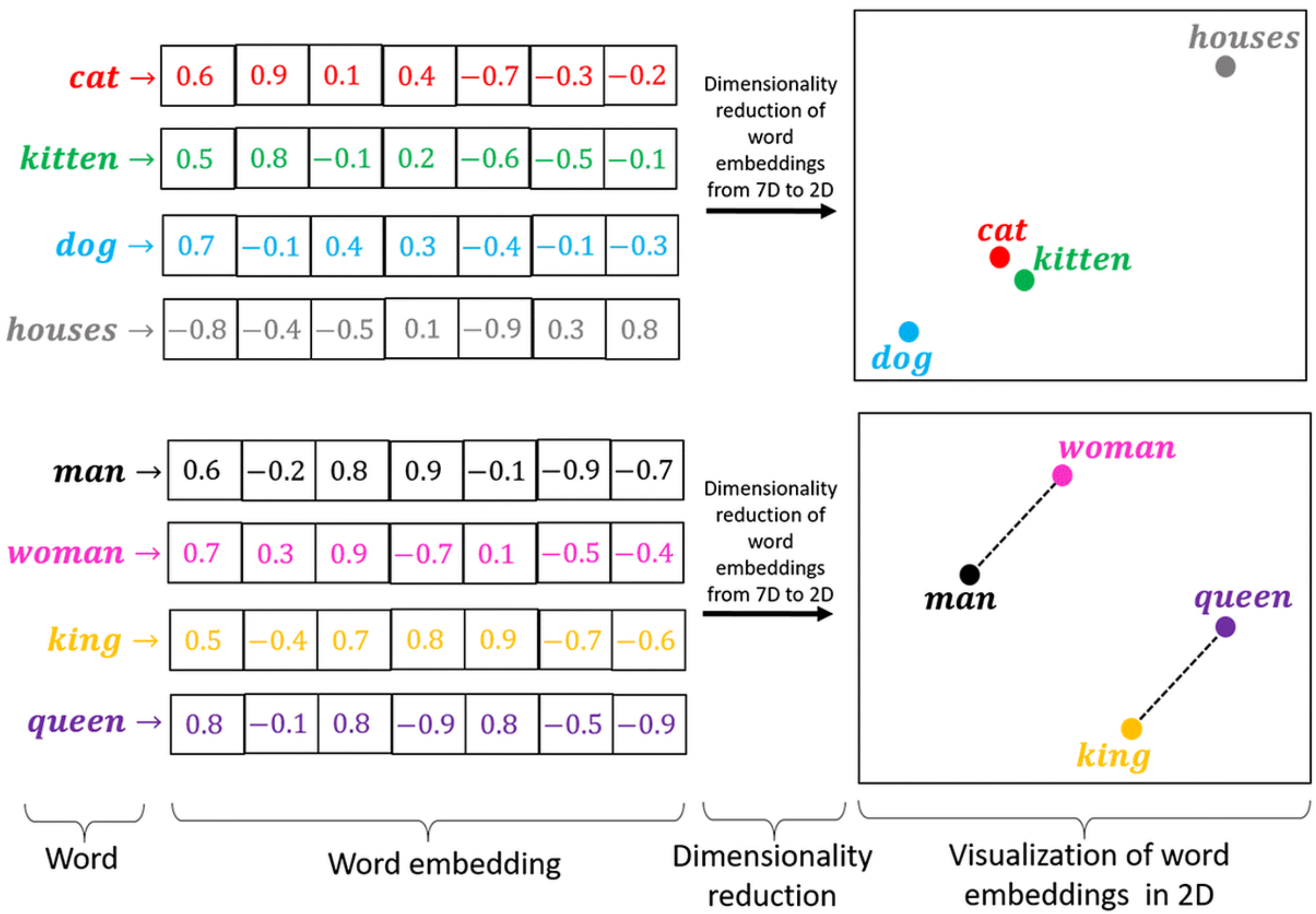
Document-term matrix

```
In [114]: df2
Out[114]:
```

	aa	aabb	aahl	aaptiv	aaron	aavitsland	ab	ababa	abaca	abad	...
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
...	...	...	...	...	...	...	...	...	...	...	...
564	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
565	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
566	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
567	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
568	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

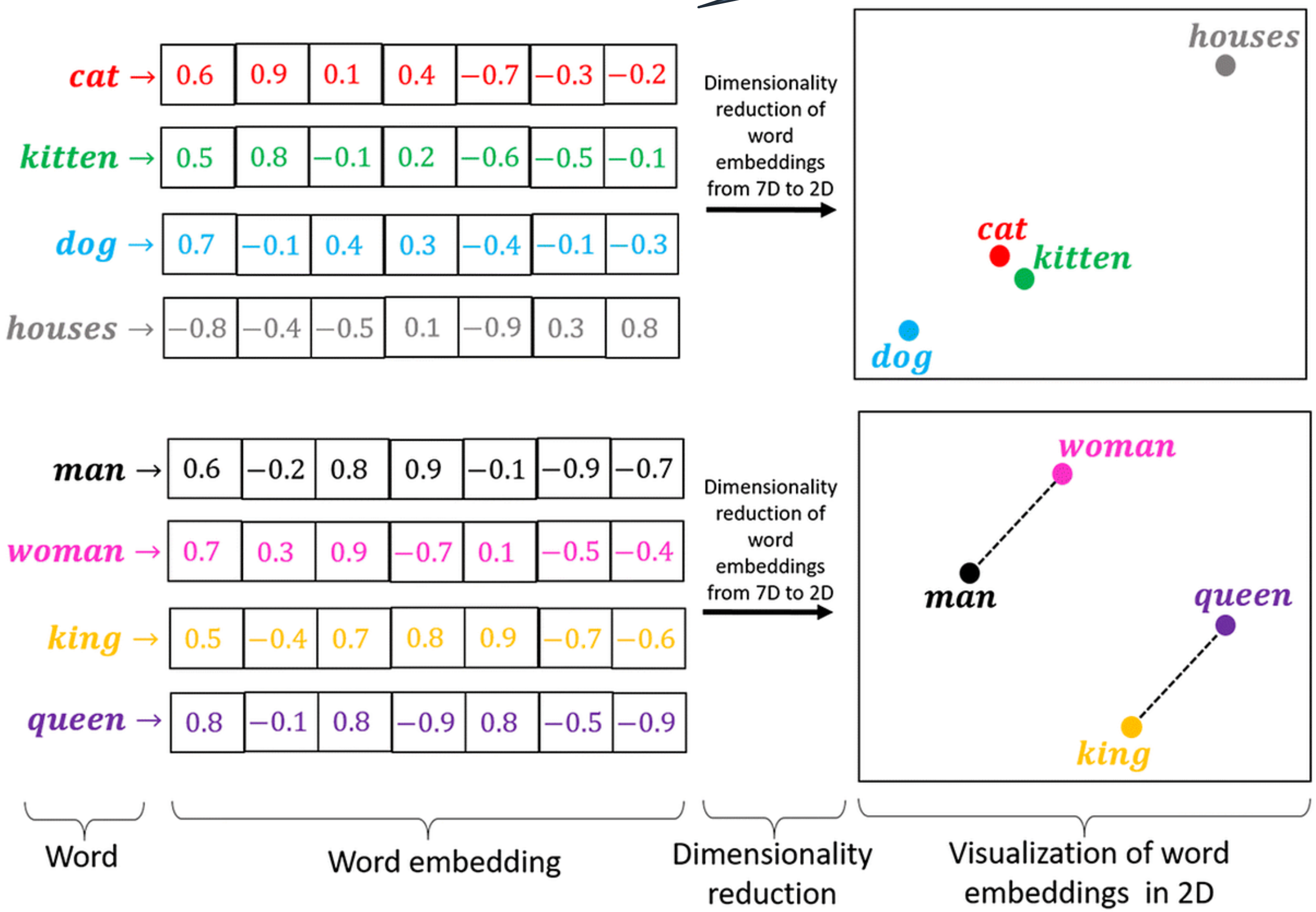
569 rows × 13794 columns

Feature Extraction: Word-embeddings



Feature Extraction: Word-embeddings

Usually, we do not know what the dimensions stand for (n_{dim}>700)



Feature Extraction: Word-embeddings

- Word embeddings capture similarities in words' meaning and function
  - fixed-length & low-dimensional
  - real-valued ("dense") ⇒ word vectors have no zero entries
  - distributed: information about words semantic properties and syntactic functions distributed across dimensions
- Static vs contextual word embeddings

	Static	Contextualized
Representation	static	dynamic
Context-	agnostic	aware
Models	pre-trained, non-adaptable	finetuning
Examples	Word2Vec, GloVe	BERT, GPT-[X]

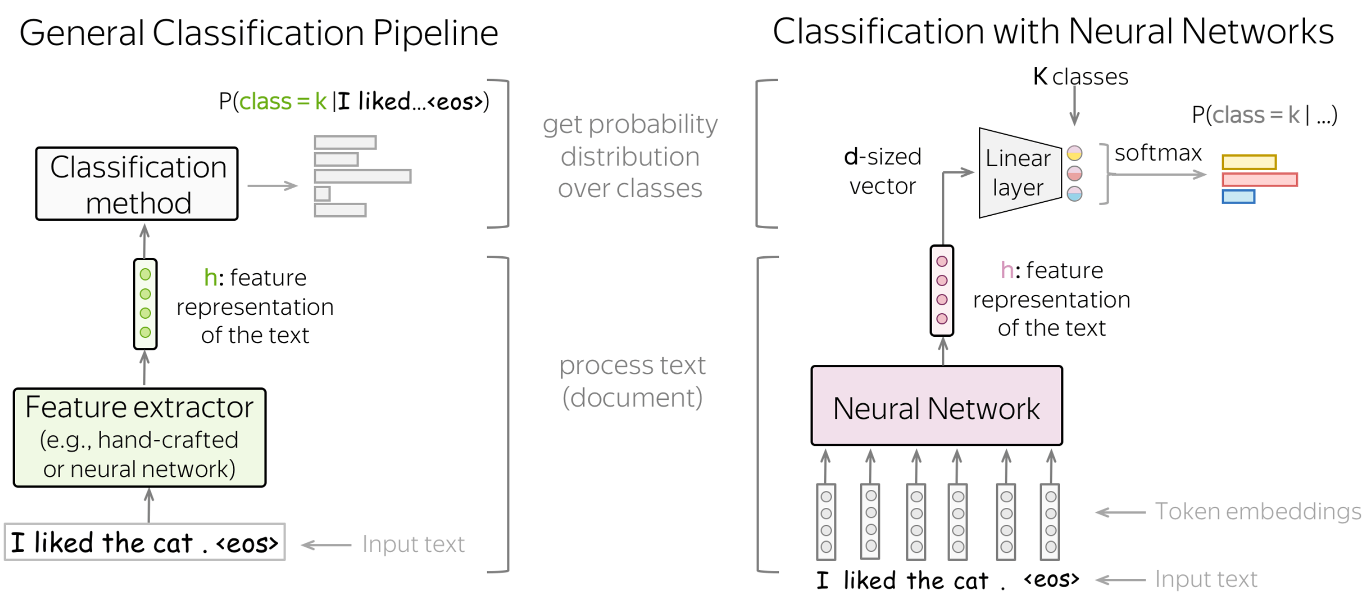
gesis Leibniz-Institut für Sozialwissenschaften

Classification Method

Feature Extraction

Feature Extraction: Word-embeddings

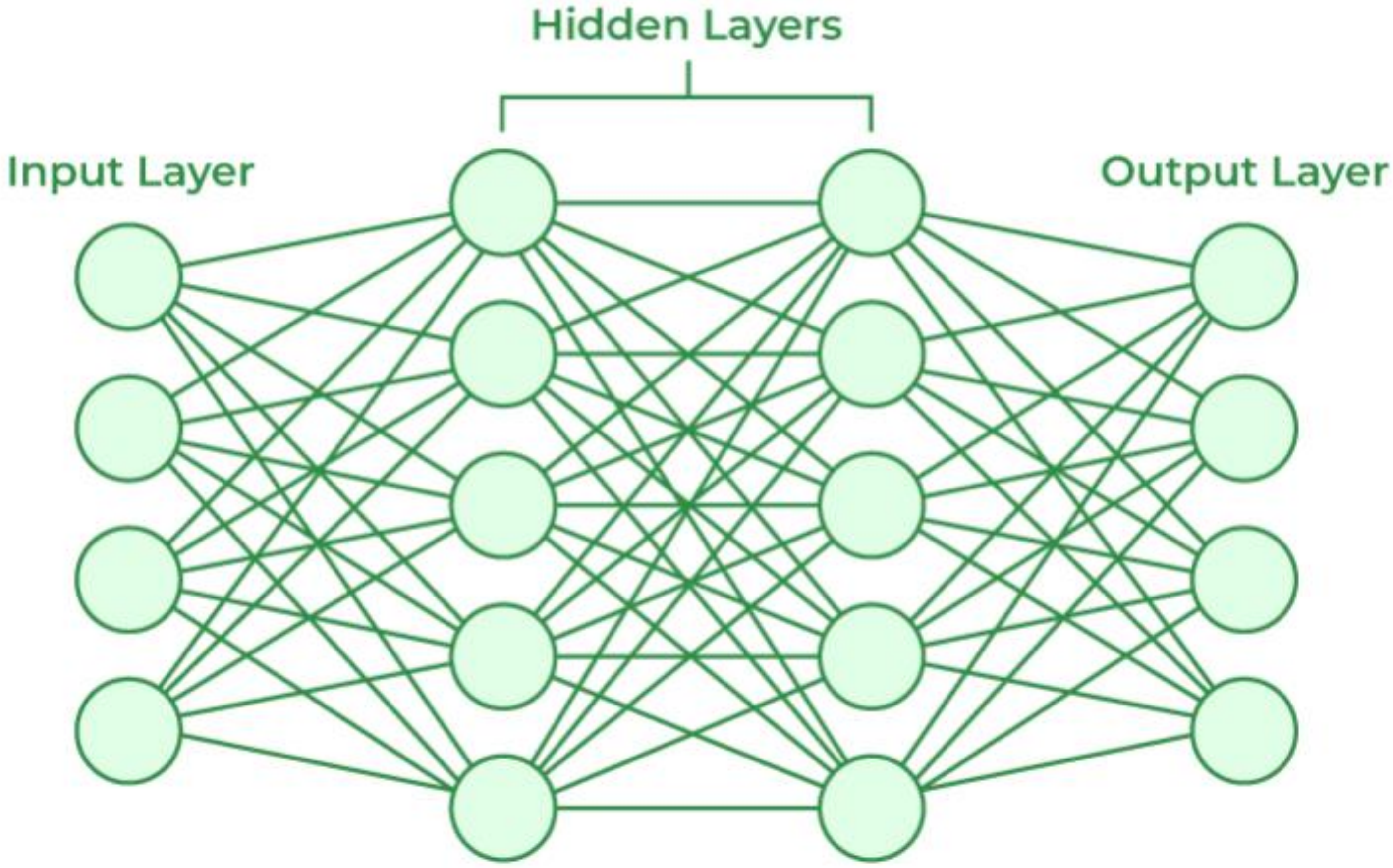
- For feature extraction, we feed the embeddings of the input tokens to a neural network
- The neural network gives us a vector representation of the input text
- Ultimately, this vector is used for classification.



Neural Networks

Classification Method

Feature Extraction

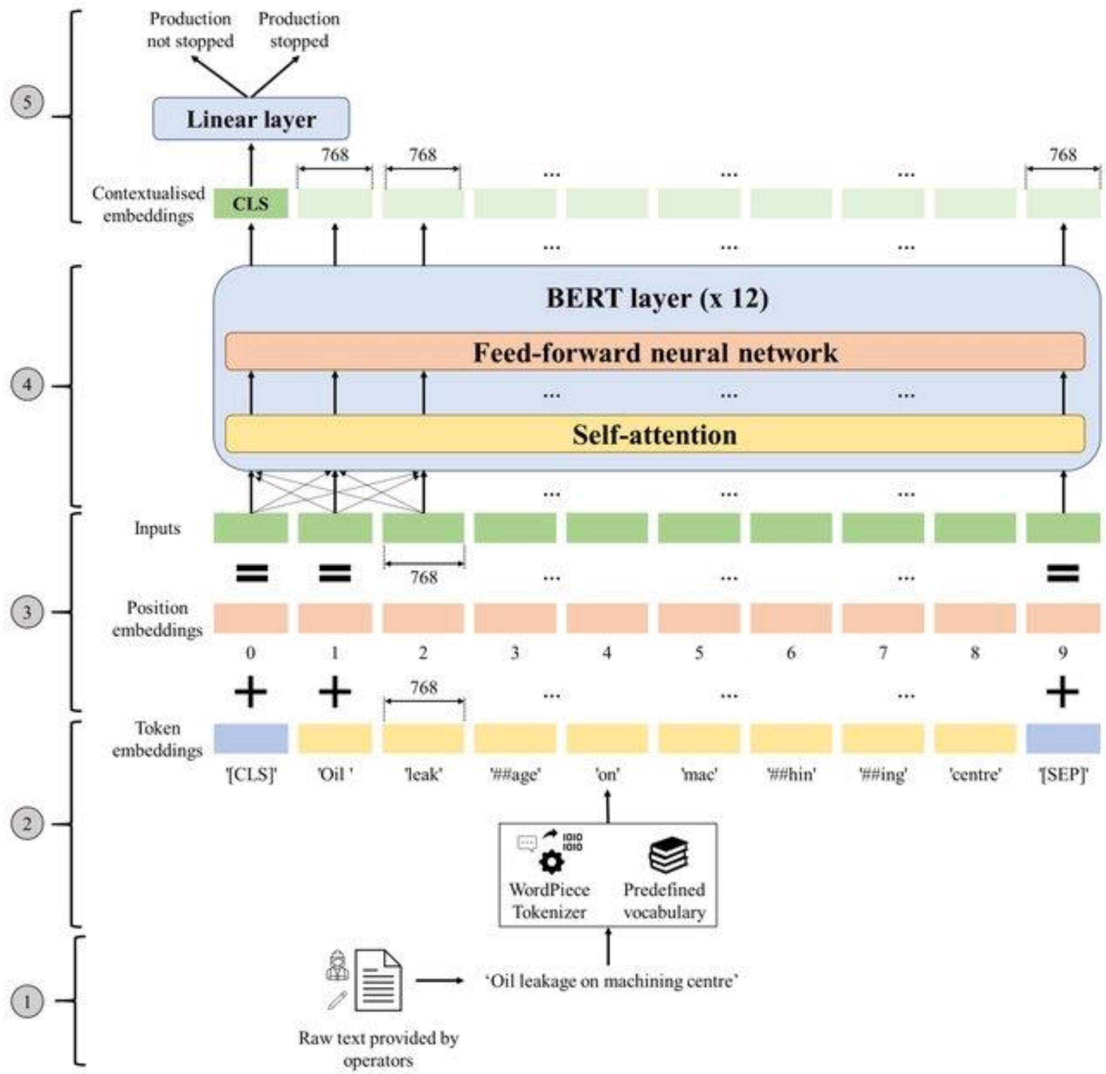


https://medium.com/@asifurrahmanaust/lesson-3-neural-network-is-nothing-but-a-linear-regression-e05a328a0f23

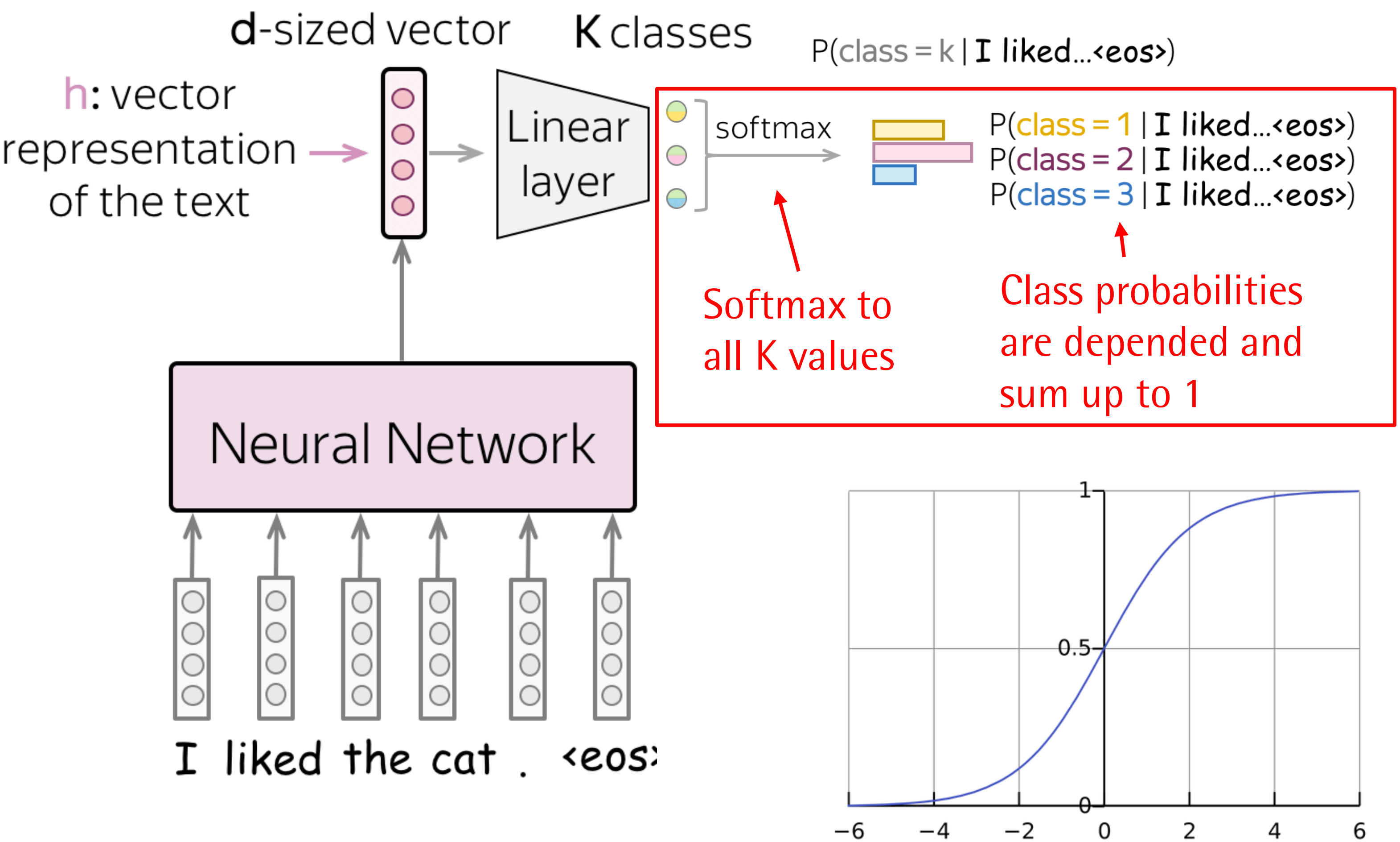
gesis Leibniz-Institut für Sozialwissenschaften

https://www.researchgate.net/publication/353419108_Using_deep_learning_to_value_free-form_text_data_for_predictive_maintenance/figures?lo=1

Transformer

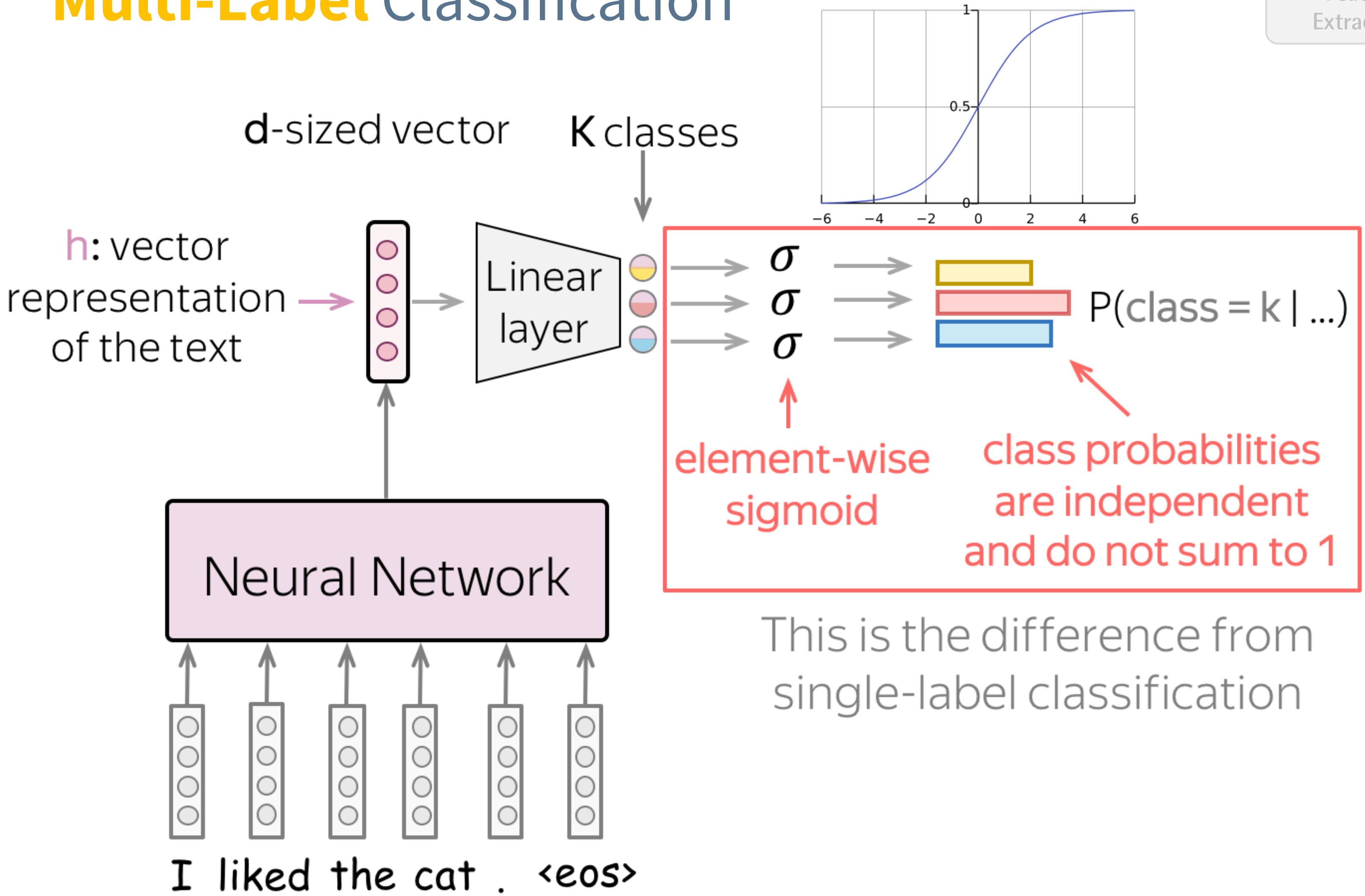


Multi-Class Classification



https://lena-voita.github.io/nlp_course/text_classification.html#multi_label

Multi-Label Classification



This is the difference from single-label classification

Annotation

gesis Leibniz-Institut für Sozialwissenschaften

Types of Annotation

Annotations are required for

- **Training / Finetuning**
- **Evaluation!**

Multiple ways to annotate text

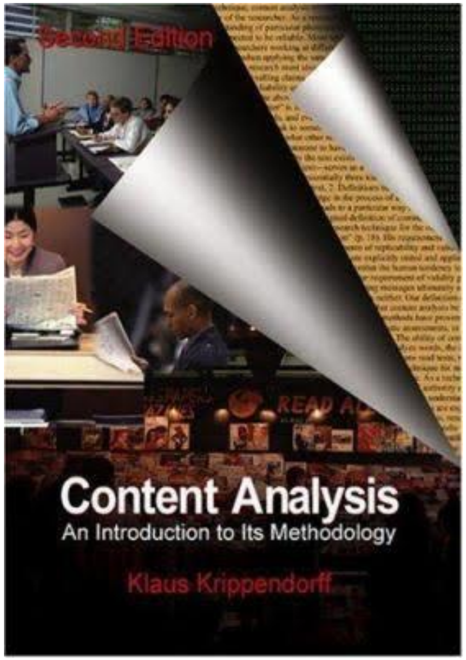
- **experts**
- **trained coders**
- **crowd workers** (since ~2010)
- **"Zero-Shot Classification of other LLMs/GPT" ?**
- (you already have labelled data, but this is often not the case 😭)

- see Krippendorff "Content Analysis: An Introduction to Its Methodology" on experts and trained coders
- see Benoit *et al.* (2016) for optimistic view on crowd coding
  - more opinions: here, here, here
- research on LLMs for annotation: here, here, here, here, and here (but still *many* open meth. questions)

Best Practices

- **Concept development**
- **Codebooks & instructions**
- **Coder training**
- **Quality assurance**

Read here for practical guidance

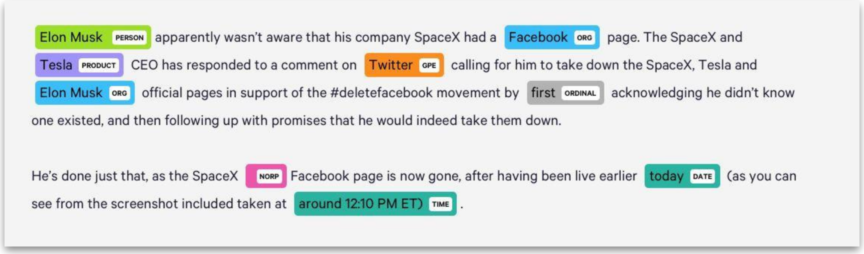


Text data are very context-dependent!
Always inspect your data critically to reflect how your constructs reflect themselves in the text!

Concept development

For instance, concept "populist/non-populist" https://doi.org/10.1017/pan.2022.32

- Level of annotation
  - **document** ⇒ "holistic grading"
  - **paragraph** ⇒ sequence classification (1+ label per para.)
  - **sentence** ⇒ sequence classification (1+ label per sent.)
  - **pairs of sentences** (see here)
  - **word** ⇒ "token classification" (1 label per word, see here)



Quality assurance & assessment

Annotation quality

- important for supervised learning
  - bad annotation result in “noisy” labels
  - noisy labels impair ability to learn the relevant signal
- related to replicability: if coders can agree, task should be replicable
- commonly quantified with inter-coder reliability metrics

Inter-coder reliability

- just % agreement is not enough (need to adjust for baseline)
- compute “chance-adjusted” agreement metrics
  - Krippendorff’s alpha
  - Cohen’s kappa
- read here and here
- https://github.com/Toloka/crowd-kit

gesis Leibniz-Institut für Sozialwissenschaften

Quality assurance & ass

Main Take away:
If humans are unsure how to classify texts, computational methods will fail as well!

Annotation quality

- important for supervised learning
  - bad annotation result in "noisy" labels
  - noisy labels impair ability to learn the relevant signal
- related to replicability: if coders can agree, task should be replicable
- commonly quantified with inter-coder reliability metrics

Int

- just % agreement is not enough (need to adjust for baseline)
- compute "chance-adjusted" agreement metrics
  - Krippendorff's alpha
  - Cohen's kappa
- read here and here
- https://github.com/Toloka/crowd-kit

https://github.com/haukelicht/advanced_text_analysis/blob/main/slides/05_annotation.pdf

gesis Leibniz-Institut für Sozialwissenschaften

Validation

Validation

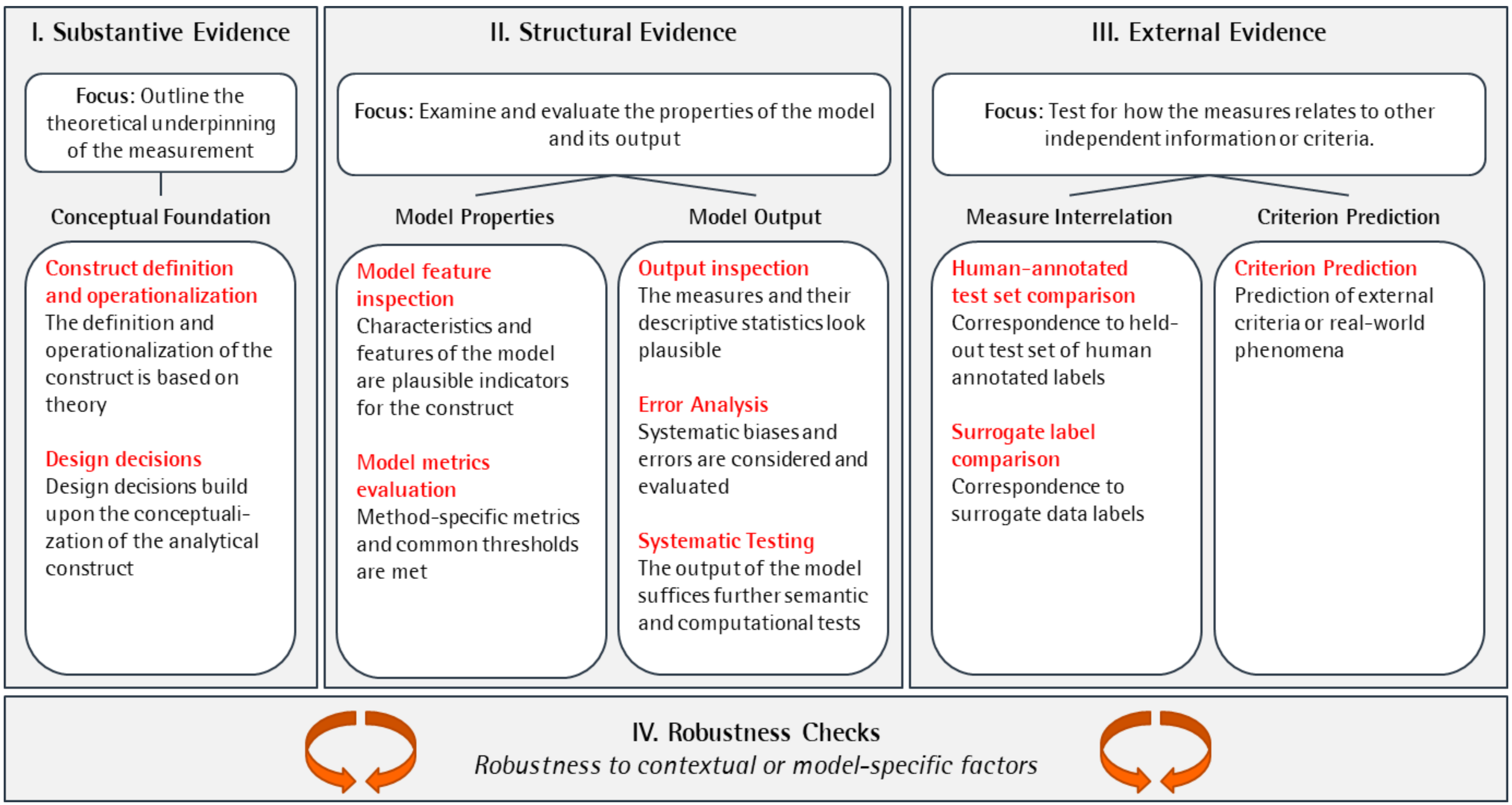
- Validation is critical task for any classification effort
- Making sure that the classification is
  - Free from Bias (systematic)
  - Little error (random)
- Two broad categories
  - Internal Validation (i.e., evaluating the measures and model features, error analysis etc.)
  - External Validation (i.e., comparing with gold-standard data)
- Especially for multi-dimensional social science constructs, validation should be taken seriously!

https://www.tandfonline.com/doi/full/10.1080/19312458.2023.2285765

gesis Leibniz-Institut für Sozialwissenschaften

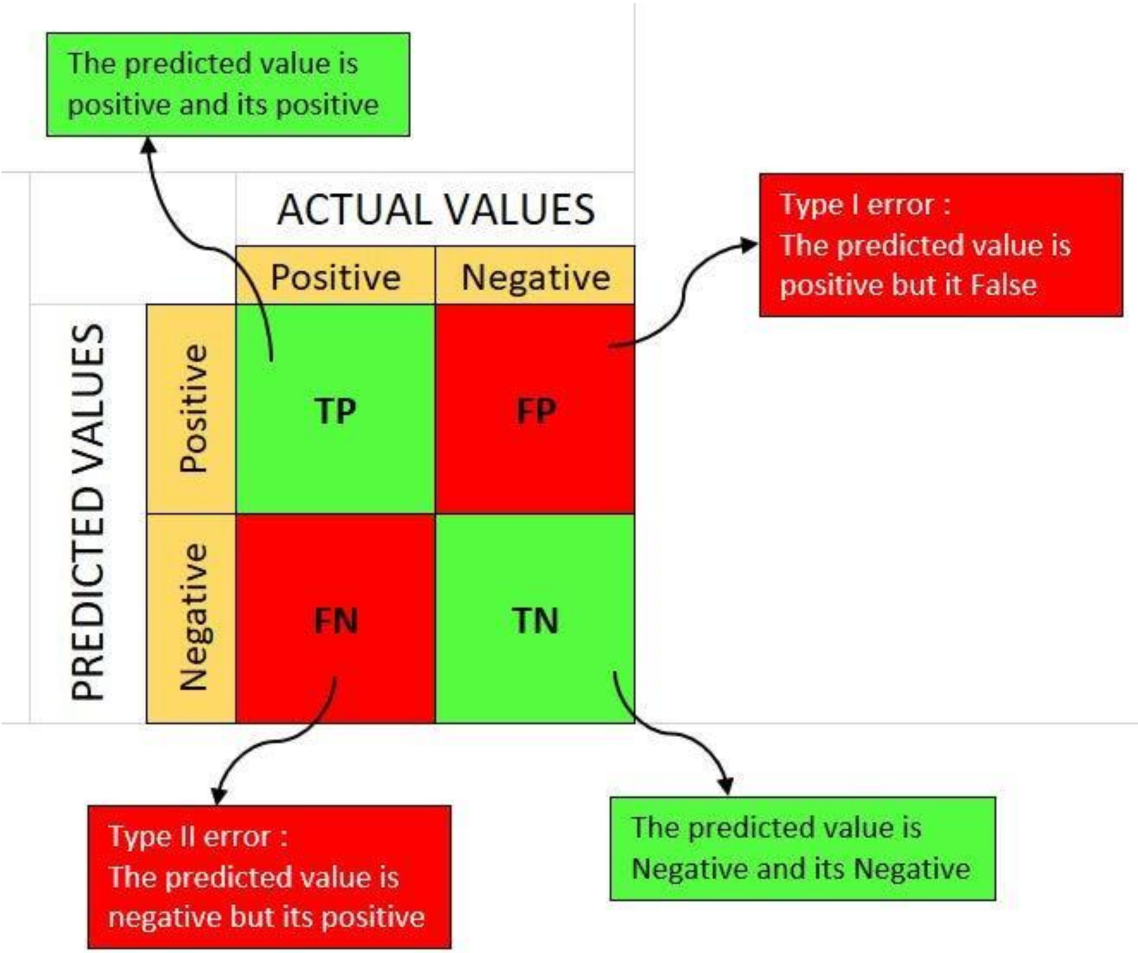
Validation

[For more information:](https://www.tandfonline.com/doi/full/10.1080/19312458.2023.2285765)
https://www.tandfonline.com/doi/full/10.1080/19312458.2023.2285765



https://www.valitex.info/
https://arxiv.org/abs/2307.02863

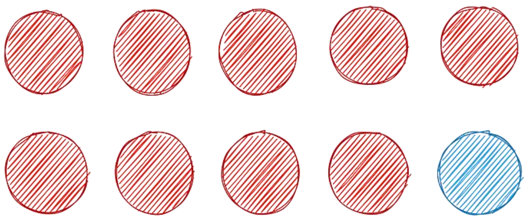
Validation: Comparison with Gold-Standard Labels



$$
\begin{aligned}
precision &= \frac{TP}{TP+FP} \\
recall &= \frac{TP}{TP+FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP+TN}{TP+FN+TN+FP} \\
specificity &= \frac{TN}{TN+FP}
\end{aligned}
$$

Why we need different metrics

- E.g., for imbalanced data, accuracy does not give the full picture
- When everything is classified as red, our classifier would have an accuracy of 90%
  - True positive = 0 (we never predict the positive class)
  - True negative = 9 (we always predict the negative class)
  - False positive = 0 (we never predict the positive class)
  - False Negative = 1 (we labeled the positive class as neg)



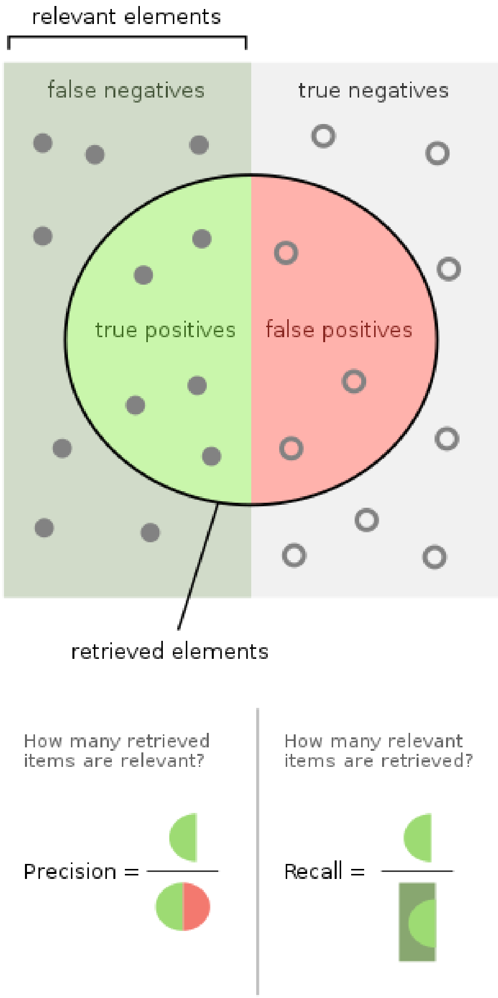
Accuracy = TP + TN / TP + TN + FP + FN
= 0+9/0+9+0+1
= 0.9

Precision = TP / (TP + FP)
= 0 / (0 + 0)
= undefined

Recall = TP / (TP + FN)
= 0 / (0 + 1)
= 0

https://towardsdatascience.com/evaluating-multi-label-classifiers-a31be83da6ea

Precision and Recall



- **Precision (Positive Predictive Value)**
  - Definition: The ratio of correctly predicted positive observations to the total predicted positive observations.
  - Importance: Critical in scenarios where the cost of false positives is high (e.g., pregnancy test)
- **Recall (Sensitivity, True Positive Rate)**
  - Definition: The ratio of correctly predicted positive observations to the all observations in actual class.
  - Importance: Essential in situations where missing a positive case has a significant consequence (e.g., COVID-test at the beginning of the pandemic)

Multi-Class Confusion Matrix

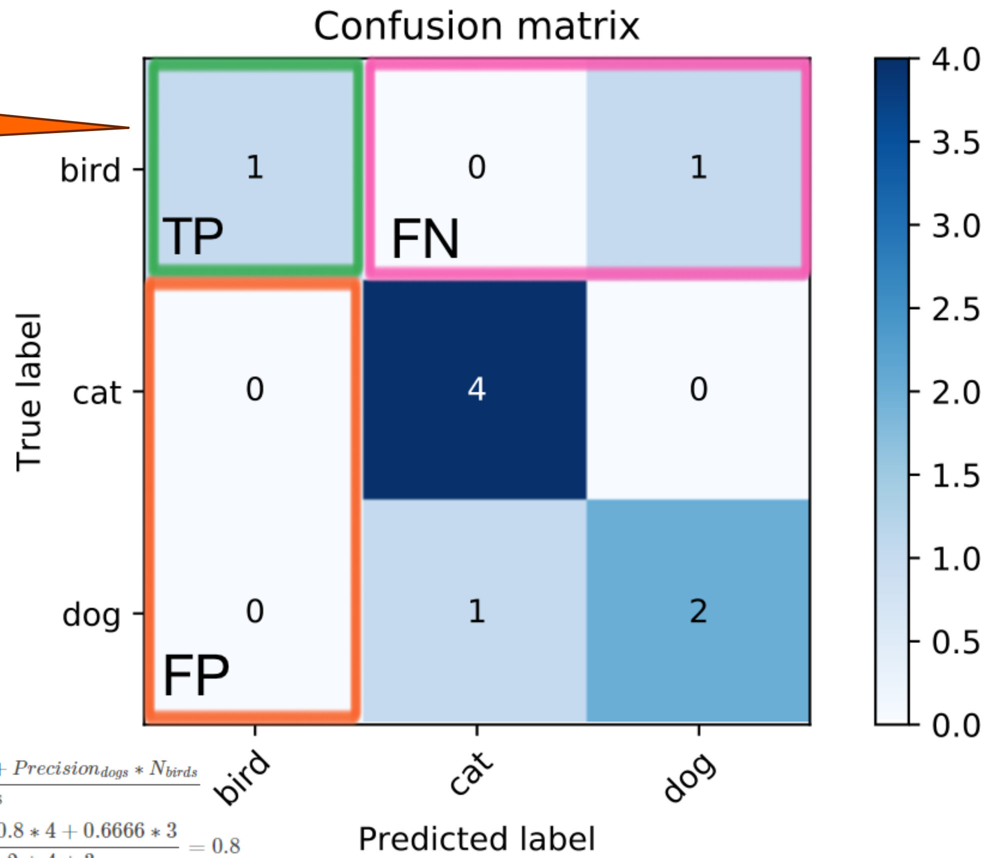
„bird“ is the reference category here

	TP	FP	FN	Precision	Number of samples
bird	1	0	1	1	2
cat	4	1	0	0.8	4
dog	2	1	1	0.667	3
TOTAL	7	2	2		

$$\text{Micro-averaged Precision} = \frac{TP_{total}}{TP_{total} + FP_{total}} = \frac{7}{7+2} = 0.7777$$

$$\text{Macro-averaged Precision} = \frac{1}{3} Precision_{birds} + Precision_{cats} + Precision_{dogs}$$
$$= \frac{1}{3}(1 + 0.8 + 0.6666) = 0.8222$$

$$\text{Weighted-averaged Precision} = \frac{Precision_{birds} * N_{birds} + Precision_{cats} * N_{birds} + Precision_{dogs} * N_{birds}}{\text{Total number of samples}}$$
$$= \frac{1 * 2 + 0.8 * 4 + 0.6666 * 3}{2 + 4 + 3} = 0.8$$



- **Micro-averaged**: all samples equally contribute to the final averaged metric
- **Macro-averaged**: all classes equally contribute to the final averaged metric
- **Weighted-averaged**: each classes's contribution to the average is weighted by its size

https://www.mariakhalusova.com/posts/2019-04-17-ml-model-evaluation-metrics-p2/

Multi-Label Confusion Matrix

```
expected    predicted
A, C         A, B
C            C
A, B, C      B, C
```

```
expected    predicted
1 0 1       1 1 0
0 0 1       0 0 1
1 1 1       0 1 1
```

$$\begin{bmatrix} TN & FP \\ TP & FN \end{bmatrix}$$

Class A: 1 0
1 1

```
Class B: 1   1
         0   1

Class C: 0   0
         1   2
```

Class A

Precision = TP / (TP + FP)
1 / (1 + 0) = 1

Recall = TP / (TP + FN)
1 / (1 + 1) = 0.5

F1-Score = 0.667

Class B

```
Precision = 0.5
Recall = 1.0
F1-score = 0.667
```

Class C

```
Precision = 1.0
Recall = 0.667
F1-score = 0.8
```

https://towardsdatascience.com/evaluating-multi-label-classifiers-a31be83da6ea

Take-aways Validation

- Calculating (average) accuracy, precision, recall, and F1-score is possible for both **multi-class** and **multi-label** classification
- Provide immediate metrics of model performance
- Software automates calculation
- More validation is required if the quality (truthfulness) of your predictions is important

Tutorials

- **Multiclass** and **Multilabel** classification
  - https://colab.research.google.com/drive/1h75O5iS9fKxHHVUJfu-u0ZOnvkmGI9RL?usp=sharing

Thank you!

gesis Leibniz-Institut für Sozialwissenschaften

Leibniz Gemeinschaft