LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Session 5:
Qualitätskriterien

„Magie“ verstehen: Klassische Schritte

1. Preprocessing

2. Analyse

3. Test auf Qualitätskriterien

Session 1
(Einführung & Preprocessing)

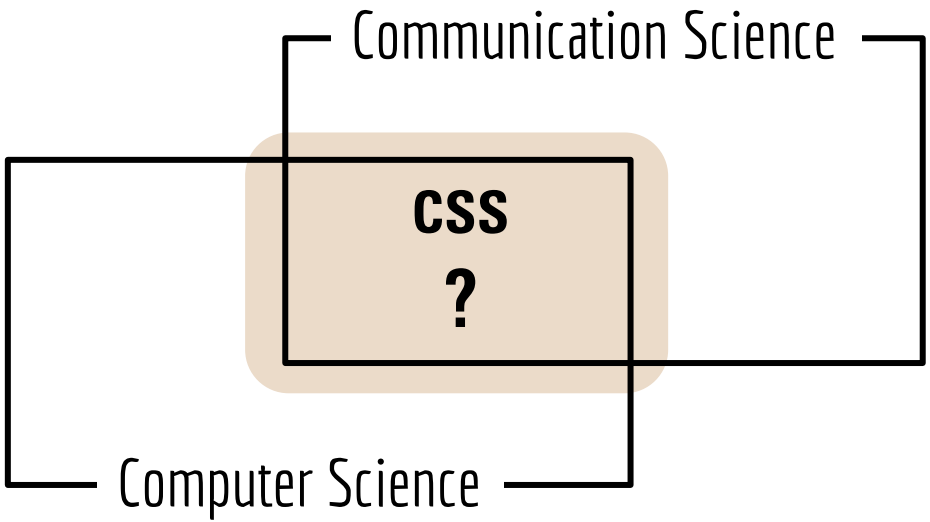
Session 2-4
(Co-Occurrence, Diktionäre, Topic Modeling)

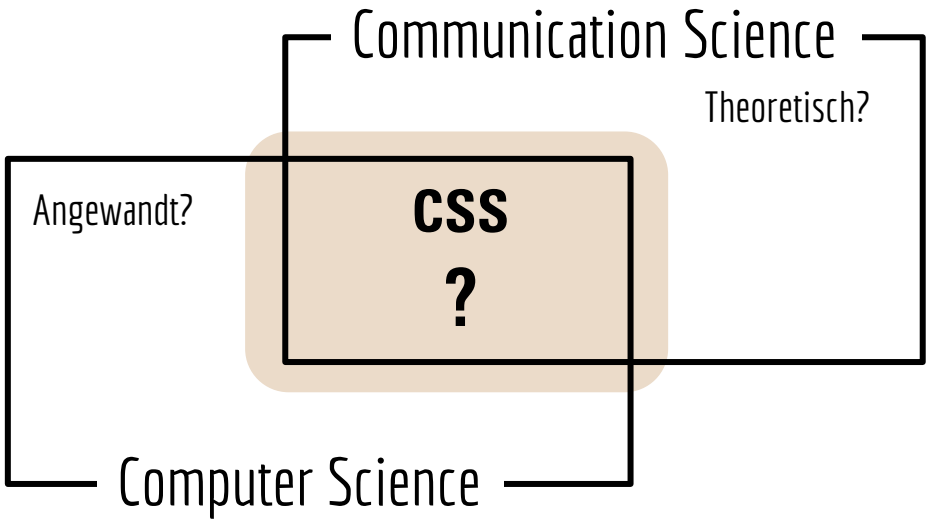
Session 5
(Qualitätskriterien)

Agenda

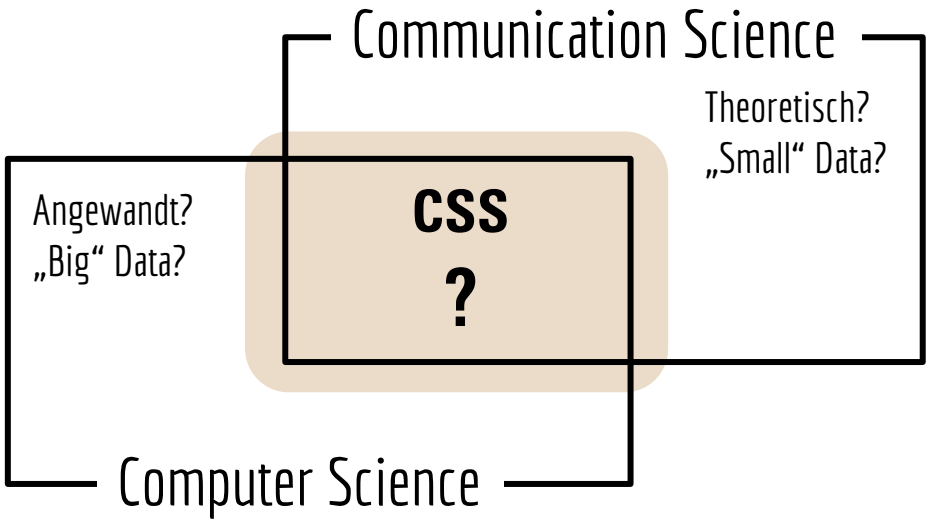
1. Einführung
2. Validierung
3. Die 4R’s
6. Outro

1. Einführung

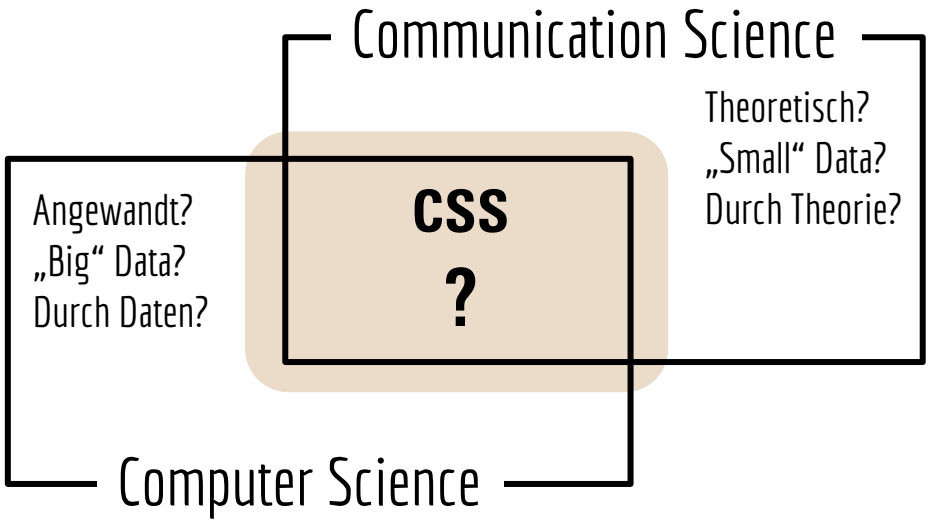

Communication Science
CSS
?
Computer Science



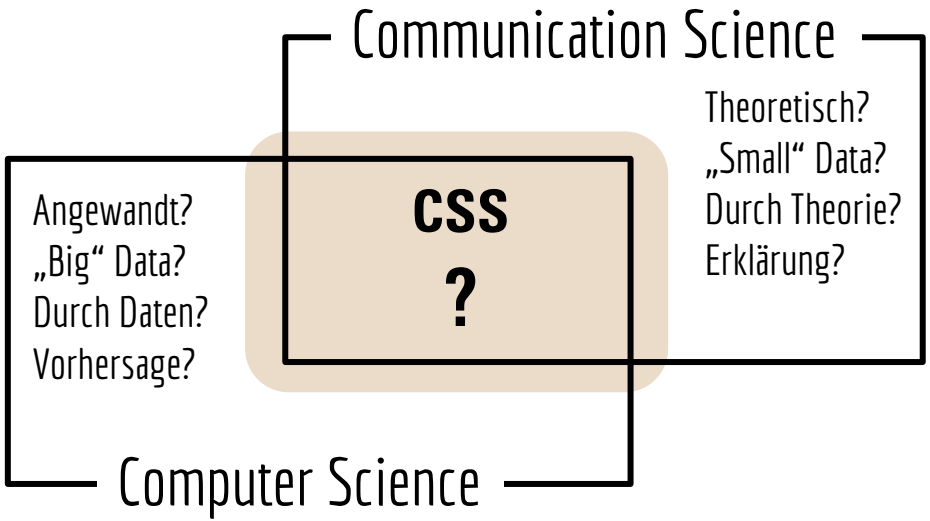
- Ziele & Fragestellungen?



- Ziele & Fragestellungen?
- Daten(-sammlung)?



- Ziele & Fragestellungen?
- Daten(-sammlung)?
- Variablen?



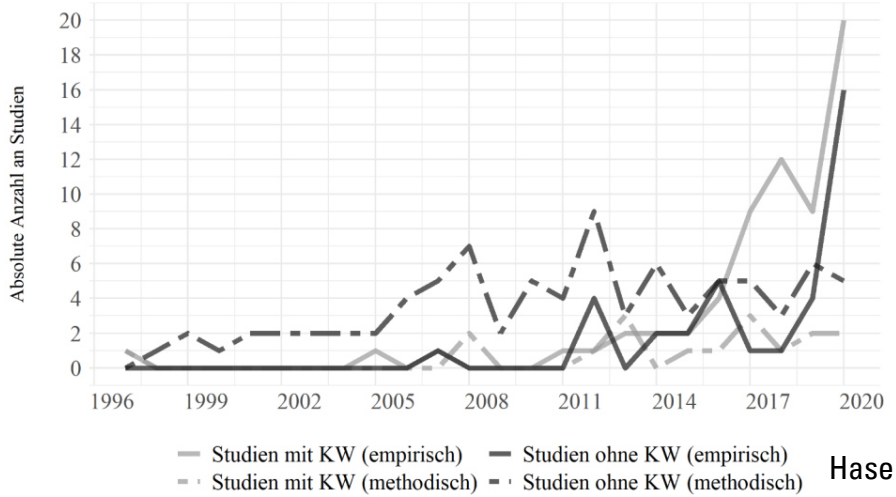
- Ziele & Fragestellungen?
- Daten(-sammlung)?
- Variablen?
- Datenanalyse?

Qualität: Ein Problem in der CSS?

Das Problem: “Everyone brings their practices and standards from their original field.”

(Theocharis & Jungherr, 2021: S. 12; Baden et al., 2022; Geise & Waldherr, 2021)

Abbildung 4: Disziplinäre Unterschiede: Nutzung der automatisierten Inhaltsanalyse



Hase et al., 2022, S. 69

Qualität: Ein Problem in der CSS?

- Das Problem: “Everyone brings their practices and standards from their original field to […] CSS.” (Theocharis & Jungherr, 2021: S. 12; Baden et al., 2022; Geise & Waldherr, 2021)
- Die Konsequenz: Kritik und Zweifel an der Qualität von CSS-Methoden und entsprechenden Ergebnissen
  - z.B. Messen wir theoretisch relevante Konstrukte? (Hase et al., 2022)
  - z.B. Lassen sich Ergebnisse reproduzieren? (Chan et al., 2024)

Qualität – Was ist das eigentlich?

?

Anhand welcher Kriterien beurteilt ihr die Qualität manueller Inhaltsanalysen?

Qualität – Was ist das eigentlich?

- Keine eindeutige Definition in der CSS (Haim et al., 2023)
- Vorgehen 1: Beurteilung dieser über Kriterien z.B.
  - **Validität**, u.a.: *Messen wir, was wir messen wollen?*
  - **Reliabilität/Robustheit**, u.a.: *Kommen wir mit anderen Instrumenten zu ähnlichen Ergebnissen?*
  - **Reproduzierbarkeit**, u.a.: *Können wir mit den gleichen Daten & Instrumenten die Ergebnisse reproduzieren?*
  - **Replizierbarkeit**, u.a. *Lassen sich unsere Ergebnisse für andere Daten reproduzieren?*
- Vorgehen 2: Beurteilung über Abwesenheit, v.a. (systematische) Fehler/„Bias“

Lösungsvorschläge

1. Qualitätskriterien entwickeln:

- z.B. Validität, etc. definieren

  (Haim et al., 2023)

- z.B. Error-Frameworks entwickeln

  (Daikeler et al., 2024; Sen et al., 2021)

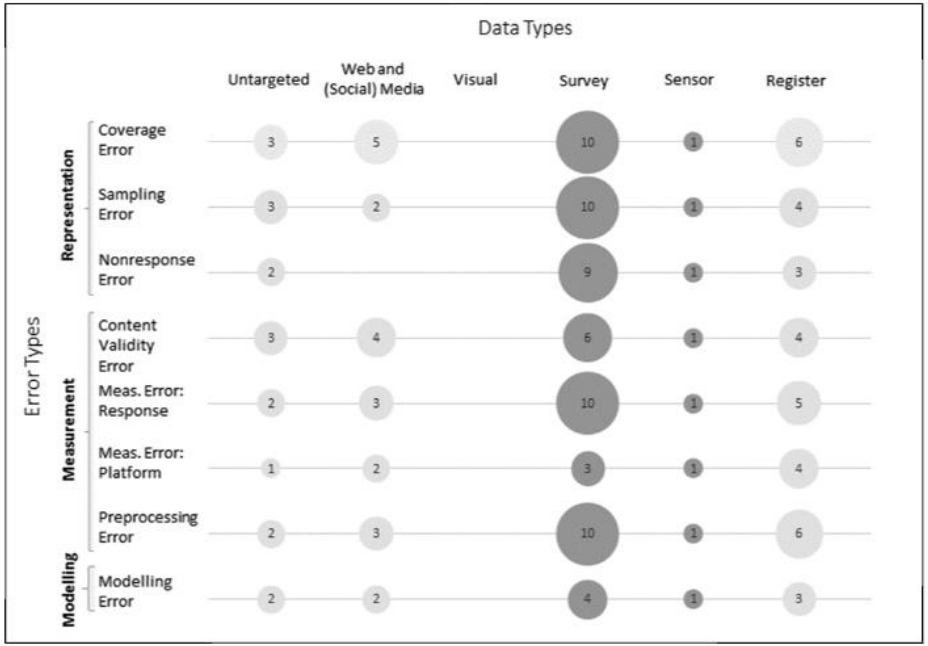


Figure 4. Evidence gap map for data types by error sources.

Quelle: Daikeler et al., 2024, S. 11

Lösungsvorschläge

2. Qualität bzw. Bias definieren

- z.B. Fehler in Stichproben

  (Knöpfle & Schatto-Eckrodt, 2024)

- z.B. Fehler in Messungen

  (TeBlunthuis et al., 2024)

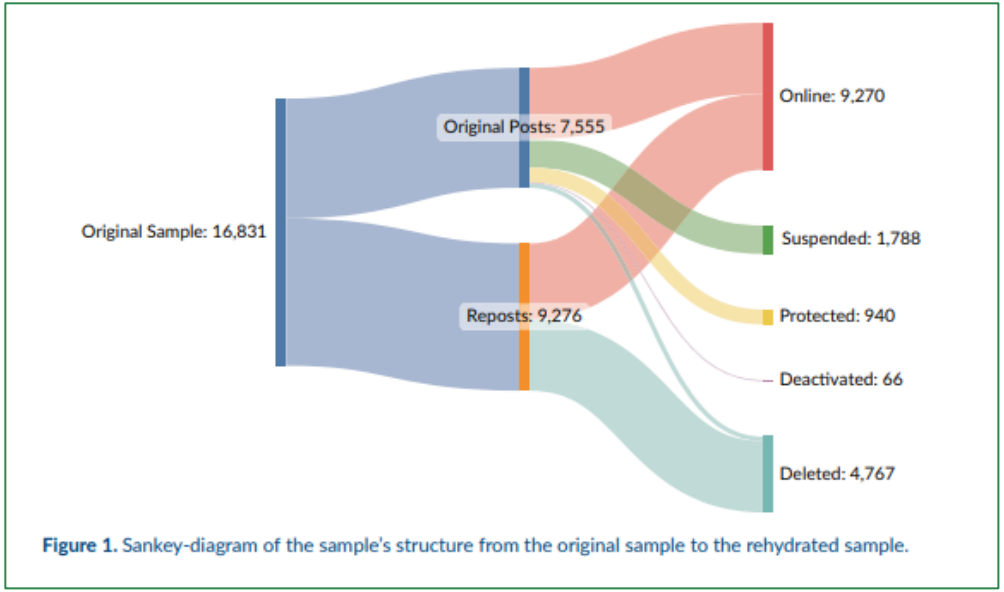


Figure 1. Sankey-diagram of the sample's structure from the original sample to the rehydrated sample.

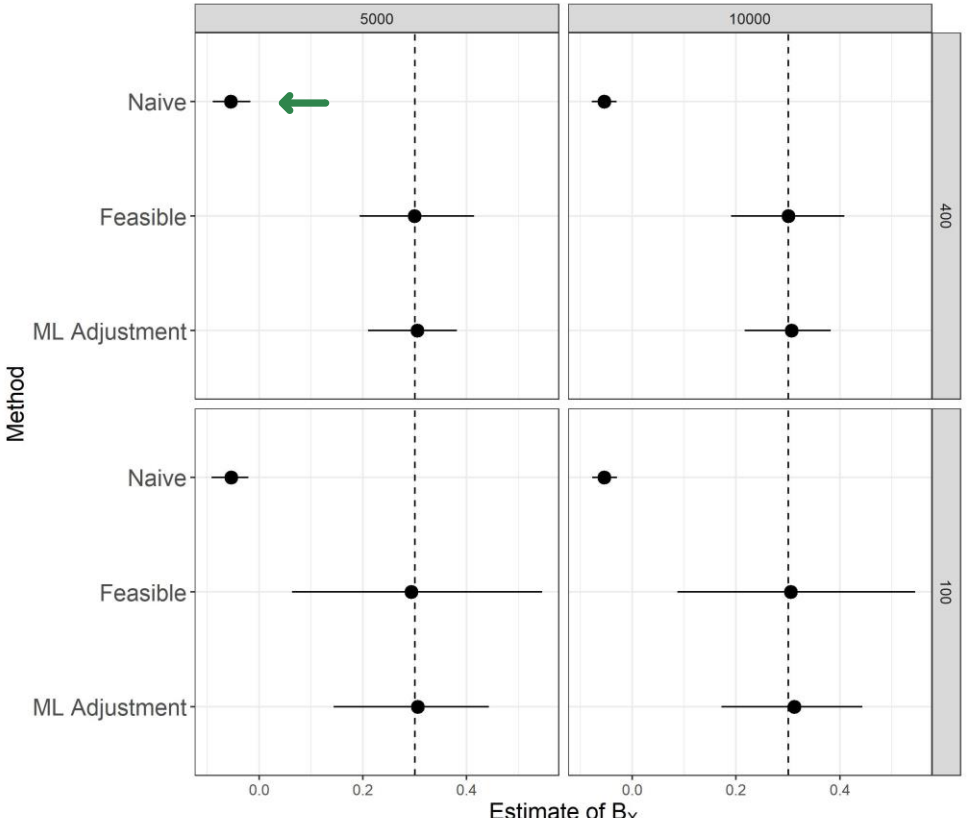
Quelle: Knöpfle & Schatto-Eckrodt, 2024, S. 11

Lösungsvorschläge

3. Methoden entwickeln, um Bias zu addressieren

- z.B. Fehler in Messungen

  (TeBlunthuis et al., 2024)



2. Validität

Validität

- Validität u.a.: *Messen wir, was wir messen wollen?*
- Misst mein Diktionär wirklich „Emotionen“?
- Misst mein Topic Model wirklich „Themen“?

Validität

- Validität u.a.: *Messen wir, was wir messen wollen?*
- Fehler können u.a. entstehen durch
  - Preprocessing
  - Operationalisierungen
  - Analysen
- Validierung hilft uns zu verstehen, **wo** wir falsch liegen und **wie falsch** wir liegen.

Validität: Messen wir, was wir messen wollen?

Qualitätssicherung z.B. via (Bernhard et al., 2023; Quinn et al., 2010) . . .

- Theoretischer (!) Ableitung von Messungen
- Vergleich mit manueller Codierung
- Vergleich mit externen Ereignissen

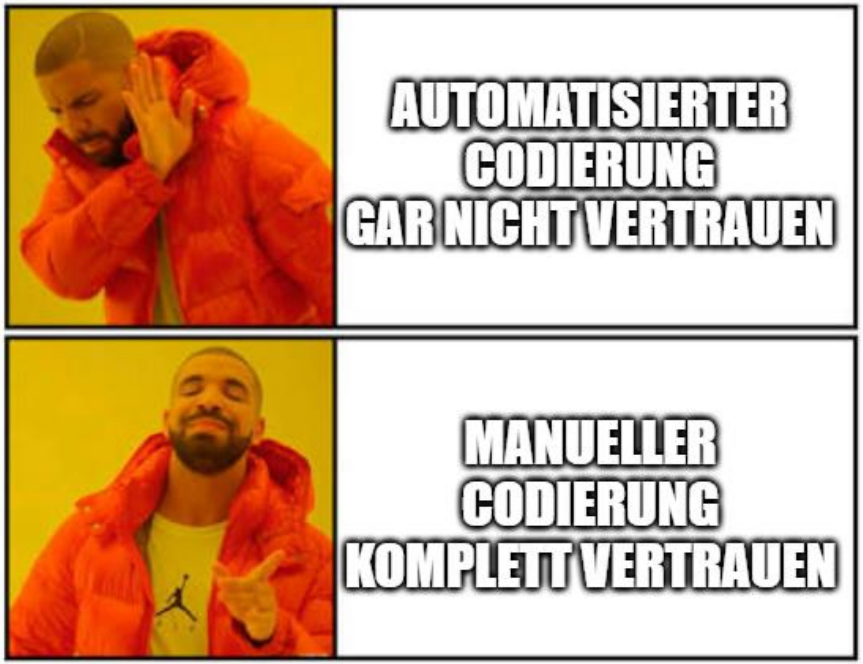
Validität: Messen wir, was wir messen wollen?

Qualitätssicherung z.B. via (Bernhard et al., 2023; Quinn et al., 2010) . . .

- Theoretischer (!) Ableitung von Messungen
- **Vergleich mit manueller Codierung**
- Vergleich mit externen Ereignissen

Vergleich mit manueller Codierung

- Wir kodieren denselben Text: a) automatisiert und b) manuell („Goldstandard“)
- Wir vergleichen Ähnlichkeiten und Unterschiede:
  - Precision
  - Recall
  - F1-Wert



Vergleich mit manueller Codierung

		Manueller «Goldstandard»	
		positive	negative
Automatisierte Inhaltsanalyse	positive	True positive	False positive (Type I error)
	negative	False negative (Type II error)	True negative

Vergleich mit manueller Codierung

		Manueller «Goldstandard»	
		positive	negative
Automatisierte Inhaltsanalyse	positive	True positive	False positive (Type I error)
	negative	False negative (Type II error)	True negative

True positive: Serien, die dem Genre Krimi angehören (gemäß Goldstandard), werden auf Basis der automatisierten Analyse korrekt als „Krimi“ klassifiziert.

Vergleich mit manueller Codierung

		Manueller «Goldstandard»	
		positive	negative
Automatisierte Inhaltsanalyse	positive	True positive	False positive (Type I error)
	negative	False negative (Type II error)	True negative

True negative: Serien, die **nicht** dem Genre Krimi angehören (gemäß Goldstandard), werden auf Basis der automatisierten Analyse korrekt als „kein Krimi“ klassifiziert.

Vergleich mit manueller Codierung

		Manueller «Goldstandard»	
		positive	negative
Automatisierte Inhaltsanalyse	positive	True positive	False positive (Type I error)
	negative	False negative (Type II error)	True negative

False negative: Serien, die dem Genre Krimi angehören (gemäß Goldstandard), werden auf Basis der automatisierten Analyse **fälschlicherweise** als „kein Krimi“ klassifiziert.

Vergleich mit manueller Codierung

		Manueller «Goldstandard»	
		positive	negative
Automatisierte Inhaltsanalyse	positive	True positive	False positive (Type I error)
	negative	False negative (Type II error)	True negative

False positive: Serien, die **nicht** dem Genre Krimi angehören (gemäß Goldstandard), werden auf Basis der automatisierten Analyse **fälschlicherweise** als „Krimi“ klassifiziert.

Vergleich mit manueller Codierung

- Precision: Inwieweit erfasst unsere Klassifizierung nur „True Positives“, d. h. nur relevante Fälle?

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Vergleich mit manueller Codierung

- Precision: Inwieweit erfasst unsere Klassifizierung nur „True Positives“, d. h. nur relevante Fälle?
- Recall: Inwieweit erfasst unsere Klassifizierung alle „True Positives“, d.h. alle relevanten Fälle?

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Vergleich mit manueller Codierung

- Precision: Inwieweit erfasst unsere Klassifizierung nur „True Positives“, d. h. nur relevante Fälle?
- Recall: Inwieweit erfasst unsere Klassifizierung alle „True Positives“, d.h. alle relevanten Fälle?
- F1: Harmonisches Mittel aus Precision und Recall
- **Wichtig**: Präzision und Recall können nicht unabhängig voneinander optimiert werden. Je besser eines von beiden wird, desto schlechter wird das andere.

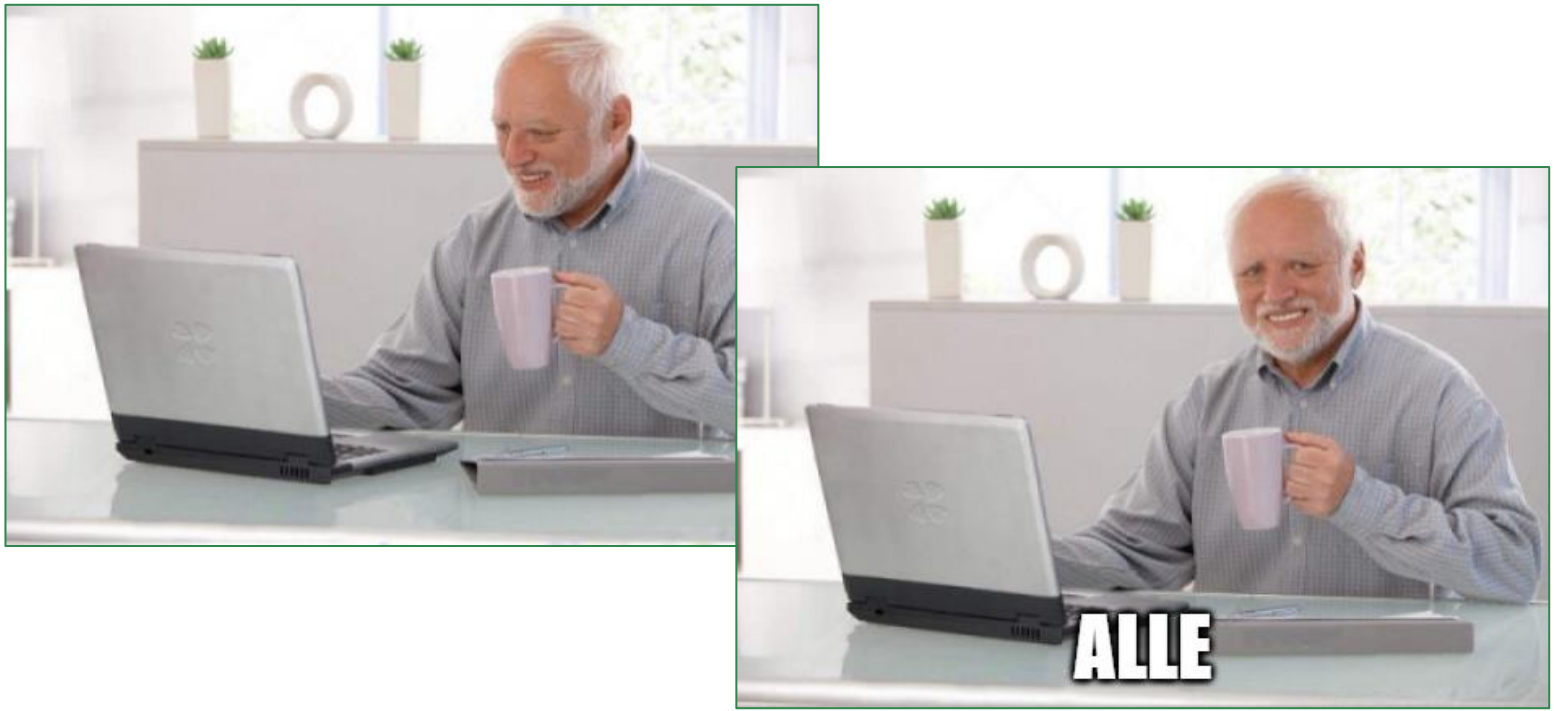
Vergleich mit manueller Codierung

- **Forschungsdesign**: Wir wollen Serien als „Krimi“ (1) oder „kein Krimi“ (0) klassifizieren.
- **Precision**: Inwieweit erfasst unsere Klassifizierung nur „True Positives“, d.h. nur Krimis werden als solche klassifiziert.
- **Recall**: Inwieweit erfasst unsere Klassifizierung alle „True Positives“, d.h. alle Krimis werden gefunden.

Wie kann ich Ergebnisse in R validieren?

- Kleine Auswahl möglicher R-Pakete
  - „`oolong`“ (z.B. für Diktionäre, Topic Models)
  - „`caret`“(z.B. für Diktionäre)

Für welche Fragestellungen eignet sich die Validierung?




Zeit für R!

Pakete installieren & aktivieren

```
#install.packages("tidyverse)
#install.packages("quanteda")
#install.packages("RCurl")
#install.packages("caret")

library("tidyverse")
library("quanteda")
library("RCurl")
library("caret")
```

Preprocessing

```
# Daten laden
url <-  getURL("https://raw.githubusercontent.com/valeriehase/textasdata-ms/main/da
data <- read.csv2(text = url)

# Preprocessing
tokens <- tokens(data$Description,
                 what = "word", #Tokenisierung, hier zu Wörtern als Analyseeinheit
                 remove_punct = TRUE, #Entfernung von Satzzeichen
                 remove_numbers = TRUE) %>% #Entfernung von Zahlen

  # Kleinschreibung
  tokens_tolower()

# Text-as-Data Repräsentation als Document-Feature-Matrix
dfm <- tokens %>%
  dfm()
```

Analyse

```
diktionär_crime <- dictionary(list(crime = c("crim*", "police*", "gun*",
                                             "shot*", "dead*", "murder*",
                                             "kill*", "court*", "suspect*",
                                             "witness*", "arrest*", "officer*",
                                             "verdict*")))
```

Analyse

```
#Diktionär anwenden
crime_tvshows <- dfm %>%
  dfm_weight(scheme = "prop") %>%
  dfm_lookup(dictionary = diktionär_crime)

# Ergebnis für die weitere Analyse in e
crime_tvshows <- convert(crime_tvshows,
                         to = "data.fra

  # Umwandlung in tibble-Format
  as_tibble %>%

  # Wir ergänzen zunächst wieder die Se
  mutate(Title = data$Title) %>%
  # Wir erstellen eine Variable, die Texte als
  # "1" (Krimi) oder "0" (kein Krimi) identifiziert
  mutate(crime_binary = 1,
         crime_binary = replace(crime_binary,
                                crime == 0,
                                0)) %>%

# Sortierung der Variablen
select(Title, crime, crime_binary)

#Ausgabe der Ergebnisse
head(crime_tvshows)
```

Analyse

```
#Ausgabe der Crime vs. Non-Crime Serien
crime_tvshows %>%

  # absolute Anzahl jeder Sentiment-Art (n)
  count(crime_binary) %>%

  # Ausgabe in Prozent (perc)
  mutate(perc = prop.table(n)*100,
         perc = round(perc, 2))
```

```
# A tibble: 2 × 3
  crime_binary     n  perc
         <dbl> <int> <dbl>
1            0   730  81.1
2            1   170  18.9
```

Analyse

```r
sample <- data %>%

  #Erstellung der Variable ID
  mutate(ID = paste0("ID", 1:nrow(data))) %>%

  # Stichprobe ziehen
  slice_sample(n = 30) %>%

  # Variable Manual Coding hinzufügen
  mutate(Manual.Coding = NA) %>%

  # Reduktion auf die drei relevanten Variablen
  select(ID, Description, Manual.Coding)
```

```r
write.csv2(sample, "validation_dictionary.csv")
```

Analyse

```
sample_coded <- read.csv2("validation_dictionary_coded.csv")
```

```
confusion <- crime_tvshows %>%

  # Erstellung der ID Variable für das Matching
  mutate(ID = paste0("ID", 1:nrow(data))) %>%

  # Match mit den codierten Daten
  right_join(sample_coded) %>%

  # Reduktion auf die relevanten Variablen
  select(ID, crime_binary, Manual.Coding) %>%
  mutate(crime_binary = as.factor(crime_binary),
         Manual.Coding = as.factor(Manual.Coding)) %>%

  # Anpassung der Variablennamen
  rename(automated = crime_binary,
         manual = Manual.Coding)

#Ausgabe der Ergebnisse
head(confusion)
```

Analyse

```
#Ausgabe der Ergebnisse
head(confusion)
```

```
# A tibble: 6 × 3
  ID    automated manual
  <chr> <fct>     <fct>
1 ID7   0         0
2 ID26  0         0
3 ID55  0         0
4 ID81  0         0
5 ID142 0         0
6 ID153 0         0
```

Analyse

```
# Berechnung der Validität
confusionMatrix(data = confusion$automated,
                reference = confusion$manual,
                mode = "prec_recall",
                positive = "1")
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 23  1
         1  2  4
```

Analyse

```
               Accuracy : 0.9
                 95% CI : (0.7347, 0.9789)
    No Information Rate : 0.8333
    P-Value [Acc > NIR] : 0.2396

                  Kappa : 0.6667

 Mcnemar's Test P-Value : 1.0000

              Precision : 0.6667
                 Recall : 0.8000
                     F1 : 0.7273
             Prevalence : 0.1667
         Detection Rate : 0.1333
   Detection Prevalence : 0.2000
      Balanced Accuracy : 0.8600

       'Positive' Class : 1
```

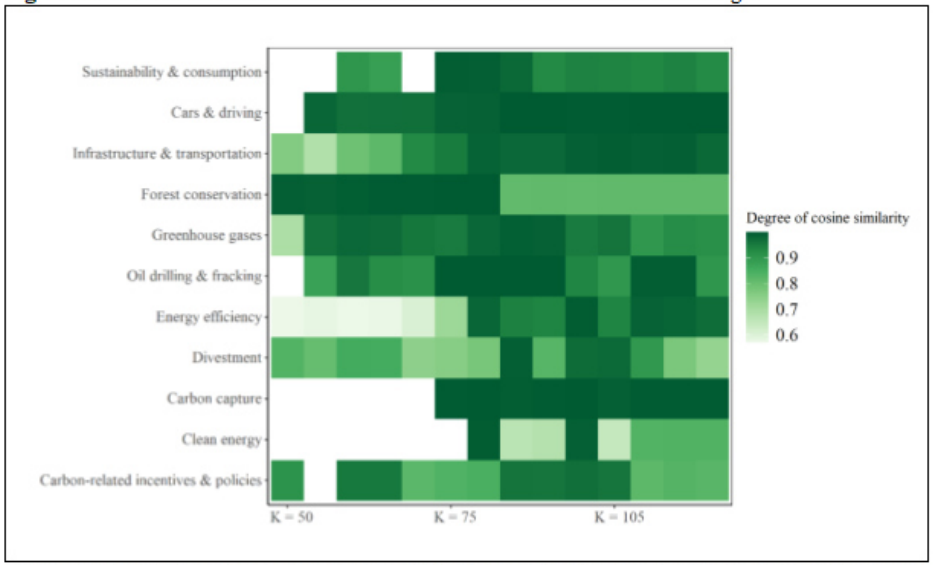
3. Die 4 R‘s

Reliabilität/Robustheit

- Reliabilität/Robustheit u.a.: *Kommen wir mit anderen Instrumenten zu ähnlichen Ergebnissen?* (Roberts et al., 2016; Wilkerson & Casas, 2017)
- Probleme, z.B.
  - **Sitzung 2 – Co-Occurrence**: Unterscheiden sich Ergebnisse je nachdem, welches „Window“ (Wörter vor/nach) Schlüsselwort ich nutze?
  - **Sitzung 3 – Diktionäre**: Unterscheiden sich Ergebnisse je nachdem, welches Diktionär ich nutze?
  - **Sitzung 4 – Topic Models**: Unterscheiden sich Ergebnisse je nach Code-Iteration?

Reliabilität/Robustheit

Figure D1. Robustness of Theme: Causes of & Solutions to Climate Change



Note: Green spaces indicate that the topic in our reference model with $K = 85$ was reproduced in models with other K. Y-axis identifies topic in our reference model, x-axis identifies robustness model with different K. The darker the green, the higher the cosine similarity between top terms of topic in the reference model and the robustness models.

(Hase et al., 2021 – Supplement)

Reproduzierbarkeit

- Reproduzierbarkeit u.a.: *Können wir mit den gleichen Daten & Instrumenten die Ergebnisse reproduzieren?*
- Lösungsvorschläge u.a. von Chan et al., 2024.:
  - Open Source Software nutzen
  - Mit z.B. „Quarto“ arbeiten (sequenzielle Reihenfolge der Coe-Ausführung garantieren!)
  - Kompendium (Code & Daten in einheitlicher Struktur; Docker)
  - Abhängigkeiten, z.B. von Paket-Versionen, reduzieren

Replizierbarkeit

- Replizierbarkeit u.a: *Lassen sich unsere Ergebnisse für andere Daten reproduzieren?*
- Lösungsvorschläge (Breuer & Haim, 2024; Long, 2021):
  - Präregistrierung
  - Auf statistische Power achten (Poweranalyse, z. B. mit Simulationen?)
  - Selbst exakte/konzeptuelle Replikationen durchführen

4. Outro

Wie berichte ich Tests auf Gütekriterien?

3.3.3. Validity & replicability

Scholars have pointed out important limitations of topic modeling (Brookes and McEnery, 2019; Grundmann, 2021; Maier et al., 2018), for instance a lack of linguistic sensitivity. To reassure linguistic sensitivity, we followed recent recommendations (Brookes and McEnery, 2019; Song et al., 2020). At least ten articles related to each topic were read by every member of the research team before labeling and interpretation. Moreover, results were validated manually based on two validation sets ($F_1 = 0.74$ and $F_1 = 0.76$ for classification of dimensions). Results showed not overly high, but sufficient validity scores except for the theme *Economic Impacts*, which should thus be interpreted with caution. Another limitation relates to the replicability and robustness of results, for instance models converging to different solutions. To reassure replicability, we employed spectral learning as a deterministic method for initialization (Roberts et al., 2016). We also checked the robustness of results independent of parameter settings, here topics being reproduced for other choices for K (Wilkerson and Casas, 2017). Detailed information on these tests can be found in the Supplementary Material (Appendix D). We agree that a final limitation – the theoretical underpinnings of topics – still applies (Brookes and McEnery, 2019; Grundmann, 2021; Maier et al., 2018) as is discussed later.

← Validität

← Reliabilität/Robustheit

Und dazu: 140 Seiten Appendix ☺

Beispiel aus Hase et al. (2021)

Wie berichte ich Tests auf Gütekriterien?

- **Immer**: Validieren!
  - Z.B. durch Abgleich mit manuellen Codierungen
  - „Goldstandard“ sollte entsprechende Reli-Werte aufweisen
- **Noch besser**: s. Vorschläge auf vorherigen Folien ☺

Wie geht es weiter?

ZEITPLAN

Mi, 24. Juli

- 09:00 - 12:00: 1 *Einführung & Preprocessing*
- 12:00 - 13:00: *Mittagspause*
- 13:00 - 15:00: 2 *Co-Occurence-Analysen*
- 15:00 - 17:00: 3 *Diktionäre*

Do, 25. Juli

- 09:00 - 12:00: 4 *Topic Modeling*
- 12:00 - 13:00: *Mittagspause*
- 13:00 - 15:00: 5 *Qualitätskriterien*
- 15:00 - 16:00: 6 *Ausblick*

Danke! Fragen?

Dr. Valerie Hase
IfKW, LMU Munich

valeriehase

valerie-hase.com

Luisa Kutlar
IfKW, LMU Munich

luisakutlar