datascience2: Prädiktive Modelle auf Basis von Textdaten

Sebastian Sauer

9/13/2022

Inhaltsverzeichnis

Hinweise

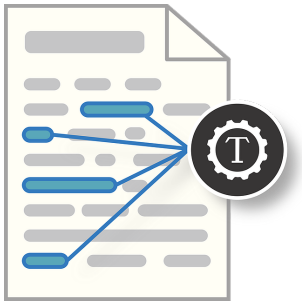

Bild von mcmurryjulie auf Pixabay

WORK IN PROGRESS

Lernziele

Nach diesem Kurs sollten Sie …

- Daten aus Sozialen Netzwerken wie Twitter automatisiert in großer Menge auslesen können
- Gängige Methoden des Textsminings mit R anwenden können (z.B. Tokenizing, Stemming, Regex)
- Verfahren des Maschinenlernens auf Textdaten anwenden können
- Den Forschungsstand zum Thema Erkennung von Hate Speech in Ausschnitten kennen

Voraussetzungen

Um von diesem Kurs am besten zu profitieren, sollten Sie folgendes Wissen mitbringen:

- fortgeschrittene Kenntnisse im Umgang mit R, möglichst auch mit dem tidyverse

- fortgeschrittene Kenntnisse der deskriptiven Statistik
- fortgeschrittene Kenntnis der Regressionsanalyse
- grundlegende Kenntnisse des Maschinenlernens

Software

- Installieren Sie R und seine Freunde.
- Installieren Sie die folgende R-Pakete[1]:
  - tidyverse
  - tidymodels
  - easystats
  - weitere Pakete werden im Unterricht bekannt gegeben (es schadet aber nichts, jetzt schon Pakete nach eigenem Ermessen zu installieren)
- R Syntax aus dem Unterricht findet sich im Github-Repo bzw. Ordner zum jeweiligen Semester.

Lernhilfen

Videos

- Auf dem YouTube-Kanal des Autors finden sich eine Reihe von Videos mit Bezug zum Inhalt dieses Buches.

Online-Zusammenarbeit

Hier finden Sie einige Werkzeuge, die das Online-Zusammenarbeiten vereinfachen:

- Frag-Jetzt-Raum zum anonymen Fragen stellen während des Unterrichts. Der Keycode wird Ihnen bei Bedarf vom Dozenten bereitgestellt.
- Padlet zum einfachen (und anonymen) Hochladen von Arbeitsergebnissen der Studentis im Unterricht. Wir nutzen es als eine Art Pinwand zum Sammeln von Arbeitsbeiträgen. Die Zugangsdaten stellt Ihnen der Dozent bereit.

[1] falls Sie die Pakete schon installiert haben, könnten Sie mal in RStudio auf "update.packages" klicken

Modulzeitplan

Nr	Thema	Datum	Kommentar
1	Twitter Mining	3. - 7. Okt. 2022	Die erste Unterrichtsstunde fällt auf de
2	Text Mining Grundlagen	10. - 14. Okt. 22	NA
3	Fallstudie Populismus	17. - 21. Okt. 22	NA
4	Word Embeddings	24. - 28. Okt. 22	NA
5	Projektwoche Twitter Hate Speech	31. Okt. - 4. Nov. 22	Ab diese Woche benötigen wir rstanarr
6	Hate Speech - Stand der Forschung	7. - 11. Nov. 22	NA
NA	NA	14. - 18. Nov. 22	Kein regulärer Unterricht
7	Regression	21. - 25. Nov. 22	NA
8	Klassifikation	28. Nov. - 2. Dez. 22	NA
9	Projektwoche Twitter Hate Speech 2	5. Dez. - 9. Dez. 22	NA
10	Quarto Blog	12. - 16. Dez. 22	NA
11	Coaching	19. - 23. Dez. 22	NA
NA	WEIHNACHTSFERIEN	NA	Kein Unterricht
12	Abschluss	9. Jan. 23 - 13. Jan. 23	NA

Literatur

Zentrale Begleitliteratur ist (**smltar_2021?**); der Volltext ist hier verfügbar.

Pro Thema wird ggf. weitere Literatur ausgewiesen.

Technische Details

Dieses Dokument wurde erzeut am/um 2022-09-14 09:22:41.

```
- Session info ---------------------------------------------------------------
 setting  value
 version  R version 4.2.1 (2022-06-23)
 os       macOS Big Sur ... 10.16
 system   x86_64, darwin17.0
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       Europe/Berlin
 date     2022-09-14
```

```
 pandoc   2.19.2 @ /usr/local/bin/ (via rmarkdown)

- Packages -------------------------------------------------------------------
 package     * version    date (UTC) lib source
 assertthat    0.2.1      2019-03-21 [1] CRAN (R 4.2.0)
 cellranger    1.1.0      2016-07-27 [1] CRAN (R 4.2.0)
 cli           3.3.0      2022-04-25 [1] CRAN (R 4.2.0)
 colorout    * 1.2-2      2022-06-13 [1] local
 colorspace    2.0-3      2022-02-21 [1] CRAN (R 4.2.0)
 DBI           1.1.3      2022-06-18 [1] CRAN (R 4.2.0)
 digest        0.6.29     2021-12-01 [1] CRAN (R 4.2.0)
 dplyr         1.0.10     2022-09-01 [1] CRAN (R 4.2.0)
 ellipsis      0.3.2      2021-04-29 [1] CRAN (R 4.2.0)
 evaluate      0.16       2022-08-09 [1] CRAN (R 4.2.0)
 fansi         1.0.3      2022-03-24 [1] CRAN (R 4.2.0)
 fastmap       1.1.0      2021-01-25 [1] CRAN (R 4.2.0)
 generics      0.1.3      2022-07-05 [1] CRAN (R 4.2.0)
 ggplot2       3.3.6.9000 2022-09-05 [1] Github (tidyverse/ggplot2@a58b48c)
 glue          1.6.2      2022-02-24 [1] CRAN (R 4.2.0)
 gt            0.7.0      2022-08-25 [1] CRAN (R 4.2.0)
 gtable        0.3.1      2022-09-01 [1] CRAN (R 4.2.0)
 htmltools     0.5.3      2022-07-18 [1] CRAN (R 4.2.0)
 jsonlite      1.8.0      2022-02-22 [1] CRAN (R 4.2.0)
 knitr         1.40       2022-08-24 [1] CRAN (R 4.2.0)
 lifecycle     1.0.2      2022-09-05 [1] Github (r-lib/lifecycle@f92faf7)
 magrittr      2.0.3      2022-03-30 [1] CRAN (R 4.2.0)
 munsell       0.5.0      2018-06-12 [1] CRAN (R 4.2.0)
 pillar        1.8.1      2022-08-19 [1] CRAN (R 4.2.0)
 pkgconfig     2.0.3      2019-09-22 [1] CRAN (R 4.2.0)
 purrr         0.3.4      2020-04-17 [1] CRAN (R 4.2.0)
 R6            2.5.1      2021-08-19 [1] CRAN (R 4.2.0)
 readxl        1.4.1      2022-08-17 [1] CRAN (R 4.2.0)
 rlang         1.0.5      2022-08-31 [1] CRAN (R 4.2.0)
 rmarkdown     2.16       2022-08-24 [1] CRAN (R 4.2.0)
 rstudioapi    0.14       2022-08-22 [1] CRAN (R 4.2.0)
 scales        1.2.1      2022-08-20 [1] CRAN (R 4.2.0)
 sessioninfo   1.2.2      2021-12-06 [1] CRAN (R 4.2.0)
 stringi       1.7.8      2022-07-11 [1] CRAN (R 4.2.0)
 stringr       1.4.1      2022-08-20 [1] CRAN (R 4.2.0)
 tibble        3.1.8      2022-07-22 [1] CRAN (R 4.2.0)
 tidyselect    1.1.2      2022-02-21 [1] CRAN (R 4.2.0)
 utf8          1.2.2      2021-07-24 [1] CRAN (R 4.2.0)
 vctrs         0.4.1      2022-04-13 [1] CRAN (R 4.2.0)
```

```
 xfun          0.32       2022-08-10 [1] CRAN (R 4.2.0)
 yaml          2.3.5      2022-02-21 [1] CRAN (R 4.2.0)

 [1] /Users/sebastiansaueruser/Rlibs
 [2] /Library/Frameworks/R.framework/Versions/4.2/Resources/library

------------------------------------------------------------------------------
```

1 Prüfung

Abbildung 1.1: Text als Datenbasis prädiktiver Modelle

Bild von mcmurryjulie auf Pixabay

Alle folgenden Hinweise gelten nur insoweit Ihre Lehrkraft Ihnen keine anders lautenden Hinweise gegeben hat (schriftlich).

1.1 Allgemeines

1. Gegenstand dieser Prüfungsform ist die *Analyse* eines Datensatzes nach einer *Forschungsfrage* und die *Dokumentation* dieser Analyse.
2. Schreiben Sie Ihre Datenanalyse in Form eines *Berichts*, der sich an den Gliederungspunkten wie unten dargestellt orientiert.
3. Wenden Sie die passenden, im Unterricht eingeführten, *statistischen Verfahren* an. Es steht Ihnen frei, andere (nicht im Unterricht behandelte) Verfahren zur Analyse der Daten anzuwenden, nach Maßgabe der fachlichen Angemessenheit.
4. Werten Sie die Daten mit *R* aus.
5. Die *R-Syntax* soll im Hauptteil des Berichts dokumentiert werden. R-Output darf ggf. gekürzt wiedergegeben werden.
6. Fügen Sie *keine Erklärungen* oder Definitionen von statistischen Verfahren an.
7. *Beschreiben* und *interpretieren* Sie jede Analyse bzw. jeden R-Code bzw. jedes Ergebnis (jede R-Ausgabe).
8. Von hoher Bedeutung ist die *Korrektheit* der Beschreibung und Interpretation der *statistischen Modellierung* (z.B. mit der Regressionsanalyse).
9. Es hat *keinen Einfluss* auf Ihre Note, ob sich ein (erwarteter) *Effekt* zeigt und wie stark dieser Effekt ggf ist.

1.2 Beurteilungskriterien

Die Arbeit wird im Hinblick auf *drei Kriterien* bewertet:

1. *Formalia* (z. B. Vollständigkeit der Abarbeitung, Angemessenheit der äußeren Gestaltung, Fokus auf Wesentliches)
2. *Methodik* (z. B. Richtige Auswahl und Anwendung der Verfahren)
3. *Inhalt* (z. B. Verständlichkeit, Breite und Tiefe der Problemlösung, Korrektheit der Interpretation)

Sie erhalten für jedes der drei Kriterien eine Teilnote sowie eine Gesamtnote. Außerdem erhalten Sie ggf. für die Kriterien noch ausformulierte Hinweise.

Die Gesamtnote muss sich nicht als Mittelwert der Teilnoten ergeben.

Insbesondere kann eine Fünf in einem Kriterium zum Durchfallen führen, auch wenn die anderen beiden Kriterien gut oder sehr gut beurteilt wurden.

1.3 Beispiele für Aspekte der Beurteilungskriterien

1. Wurden deskriptive Statistiken (an angemessenen Ort) berichtet?
2. Wurden Diagramme und Tabellen angemessen eingesetzt?
3. Wurde Inferenzstatistik (angemessen) eingesetzt?
4. Wurden Effektstärkemaße (idealerweise mit Konfidenzintervallen dazu) berichtet?
5. Wurden alle relevanten Informationen für ein statistisches Verfahren angegeben (z.B. zum gewählten Prior)?
6. Wurde die Aussagekraft von Modellergebnissen richtig eingeschätzt?
7. Waren die Schlussfolgerungen, die aus den statistischen Ergebnissen gezogen wurden, angemessen (z. B. wurde erkannt, dass ein Nicht-Verwerfeen einer Hypothese nicht automatisch ein Bestätigen derselben bedeutet)?
8. Wurde angemessen gerundet (inkl. konsistente Anzahl von Nachkommastellen)?
9. Passen die statistischen Verfahren zu den Hypothesen?
10. Wurden die Voraussetzungen der statistischen Verfahren geprüft?
11. Sind die Ergebnisse reproduzierbar (Daten und Syntax eingereicht)?

1.4 Beispiele für Fehler

Schwere Fehler, die zum Durchfallen oder deutlichem Abwerten der Note führen können, sind z.B.:

- fehlende Inferenzstatistik (oder adäquatem Ersatz)
- falsche Interpretation von Posteriori-Verteilungen oder p-Werten
- keine Angabe von Konfidenzintervallen
- falsche Interpretation von Konfidenzintervallen
- Wahl des falschen Intervalls (Vorhersageintervall vs. Perzentilintervall vs. HDI)
- falsche Entscheidung zum Hypothesentest auf Basis entsprechender Kennwerte (wie ROPE-Wahrscheinlichkeit oder p-Wert)
- falsche Wahl des statistischen Verfahrens
- fehlende Deskriptivstatistik

Häufige kleinere Mängel sind z. B.

- pixelige Abbildungen
- R-Ausgaben oder R-Syntax als Screenshot
- fehlende Seitenzahlen (nur bei paginierten Formaten, nicht bei HTML)
- unübersichtliche Diagramme
- kein (verlinktes) Inhaltsverzeichnis
- fehlende oder unverständliche Achsenbeschriftung bei Diagrammen
- fehlende oder falsche Beschreibung der/des Skalenniveau(s) der untersuchten Variablen

2 Twitter Mining

Abbildung 2.1: Text als Datenbasis prädiktiver Modelle

Bild von mcmurryjulie auf Pixabay

2.1 Vorab

2.1.1 Lernziele

- Twitterdaten via API von Twitter auslesen

2.1.2 Vorbereitung

- Lesen Sie in Hvitfeldt und Silge (2022) Kap. 1 und 2.
- Legen Sie sich ein Konto bei Github an.

2.1.3 R-Pakete

```
library(tidyverse)
library(twitteR)
```

References

Hvitfeldt, Emil, und Julia Silge. 2022. *Supervised Machine Learning for Text Analysis in R*. 1. Aufl. Boca Raton: Chapman; Hall/CRC. https://doi.org/10.1201/9781003093459.