DATA CHALLENGE: OPTIMIERUNG UND TEXT-MINING MIT NACHHALTIGKEITSASPEKTEN

INHALT

- Projektbeschreibung und Ziele
- Rechtliche Aspekte Webscraping
- Ansatz: Datenvorverarbeitung & Modelle
- Ergebnisse
- Nachhaltigkeitsdiskussion
- Fazit & Ausblick

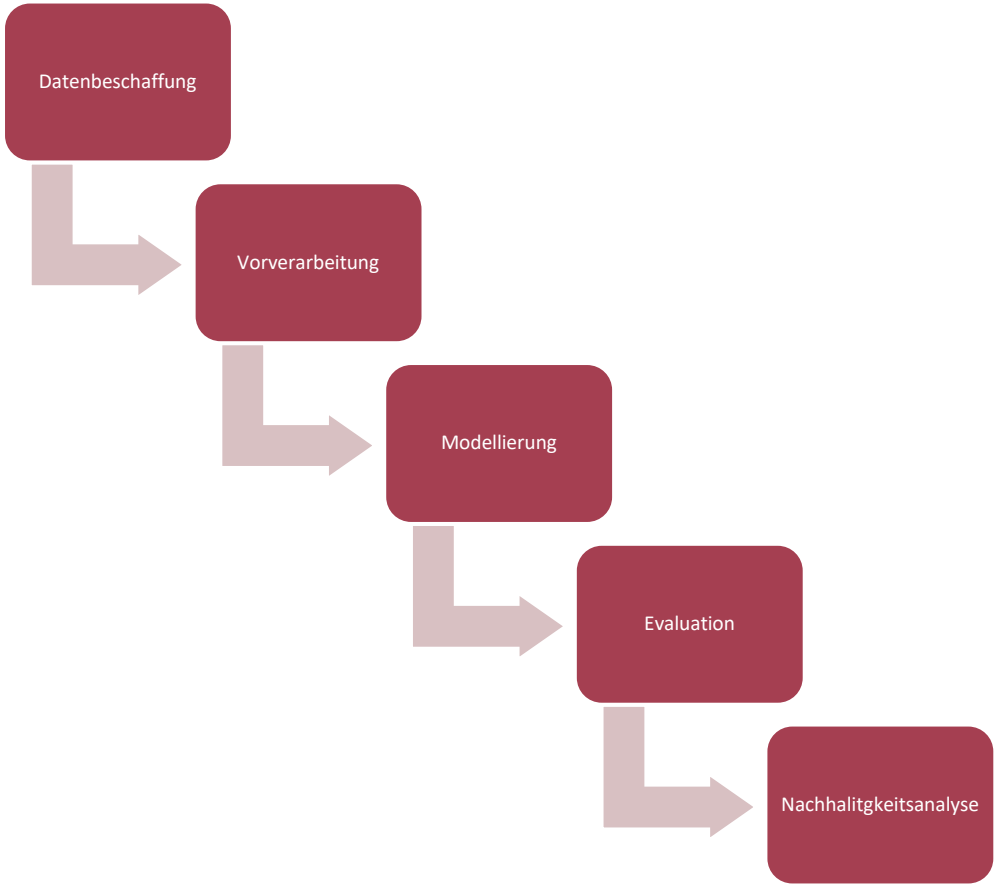
PROJEKTBESCHREIBUNG UND ZIELE

•Analyse von Nachrichtenartikeln zur Vorhersage der Shares

•Kombination von klassischen ML-Methoden und modernen Text-Embedding-Ansätzen

•Fokussierung auf Ressourceneffizienz und nachhaltige Datenverarbeitung

•Datenquelle: https://www.kaggle.com/datasets/thehapyone/uci-online-news-popularity-data-set



RECHTLICHE ASPEKTE & PSEUDOCODE

- Automatisiertes Webscraping kann ohne Erlaubnis rechtliche Probleme verursachen
- Verwendung von Pseudocode, um methodische Ansätze zu demonstrieren, ohne tatsächliche Scraping-Aktivitäten durchzuführen
- Hinweis: Für Scraping müssen Lizenzen oder Genehmigungen eingeholt werden

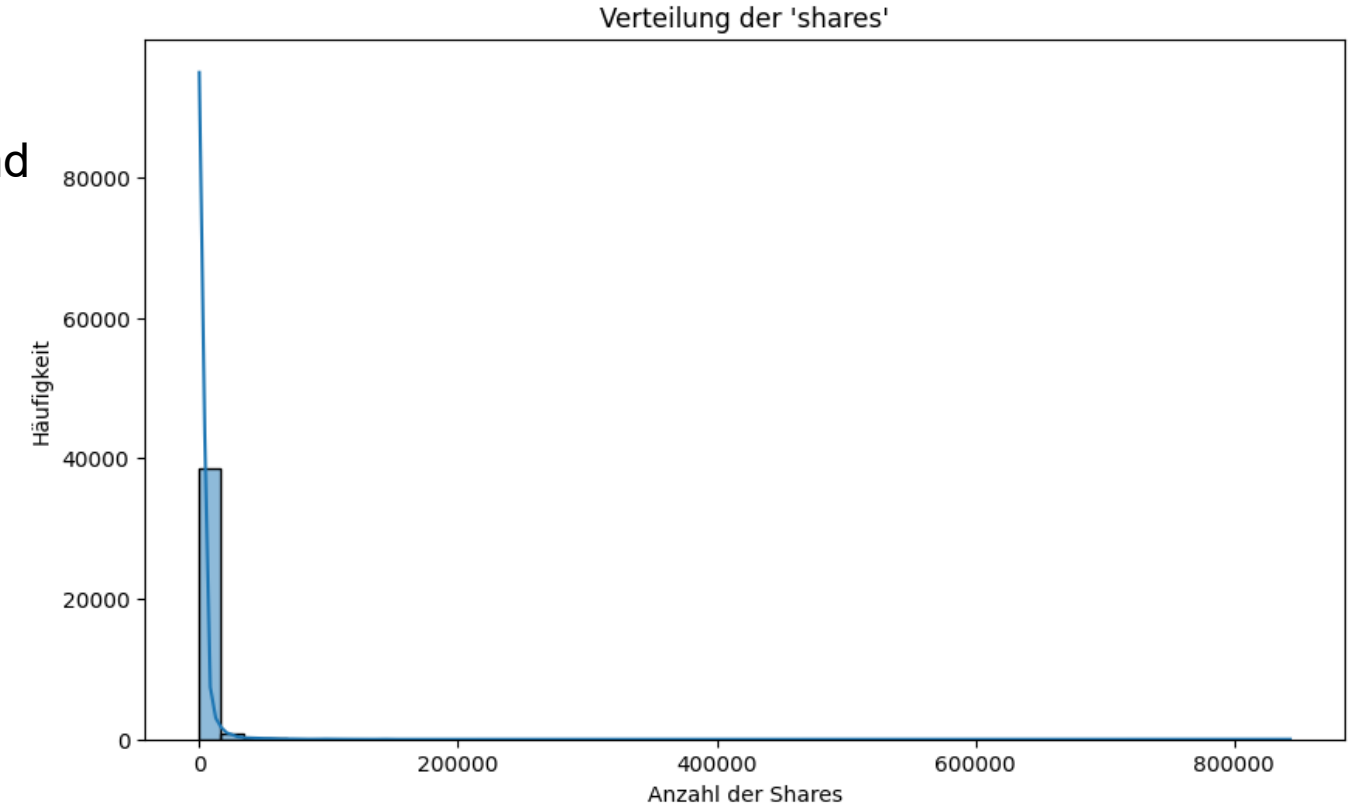
Beispiel: robots.txt von mashable.com

Ziff Davis content is made available for your non-commercial use subject to our # Terms of Use here: https://www.ziffdavis.com/terms-of-use # Use of any robot, crawler, or other tool to scrape, harvest, extract, or retrieve any content on # this website using automated means is prohibited without written permission from Ziff Davis. # Prohibited uses include but are not limited to: # (1) text and data mining under Art. 4 of the EU Directive on Copyright in the Digital Single # Market; # (2) development or operation of artificial intelligence or machine learning software or # databases, including by training, fine-tuning, embedding, and retrieval-augmented generation; # (3) creating data sets containing our content or sharing it with others; and # (4) any commercial purposes. # Contact licensing@ziffdavis.com for assistance. User-agent: * Disallow: /search Disallow: /archive/ Disallow: /cdn-cgi/ Allow: /*?page=[0-9] User-agent: "008" User-agent: Amazonbot User-agent: anthropic-ai User-agent: Applebot User-agent: Applebot-extended User-agent: Bytespider User-agent: CCBot User-agent: ClaudeBot User-agent: Claude-Web User-agent: cohere-ai User-agent: Diffbot User-agent: FacebookBot User-agent: GPTBot User-agent: HTTrack User-agent: Nutch User-agent: Offline Explorer User-agent: omgili User-agent: Scrapy User-agent: YouBot Disallow: / Sitemap: https://mashable.com/sitemap-index.xml Sitemap: https://mashable.com/sitemap-news-0.xml

DATENVORBEREITUNG & EXPLORATIVE DATA ANALYSIS (EDA)

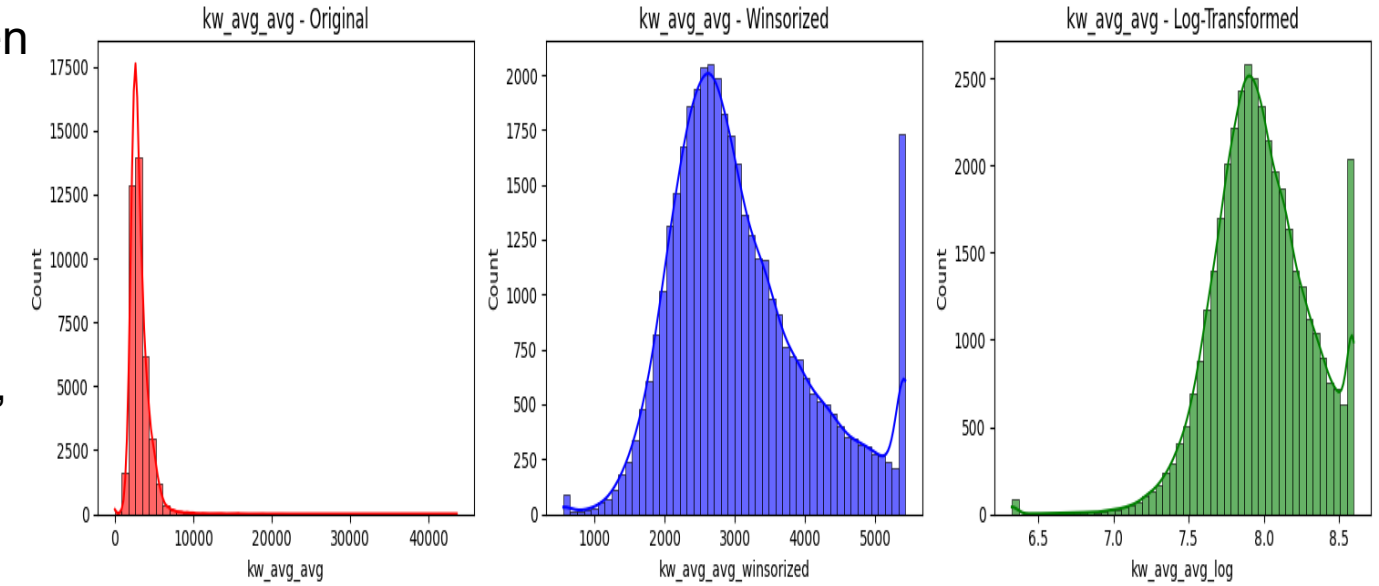
•Daten einlesen, Spaltennamen bereinigen und erste statistische Kennzahlen ermitteln

•Visualisierung der Zielvariable „shares“ (Histogramm)



FEATURE ENGINEERING & TRANSFORMATION

- Identifikation von kontinuierlichen und binären Variablen
- Anwendung von Winsorisierung und Log-Transformation zur Ausreißerbehandlung
- Extraktion textbasierter Features (Sentiment, Lesbarkeitsmetriken)
- Berechnung von Transformer-Embeddings



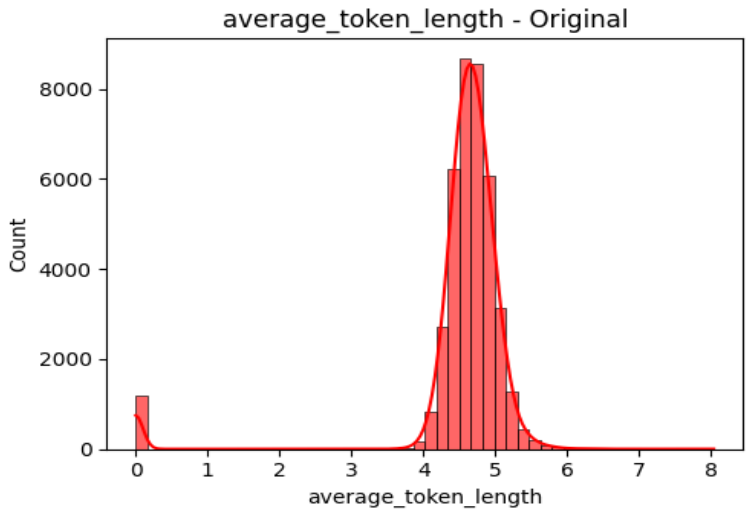
FEATURE ENGINEERING & TRANSFORMATION

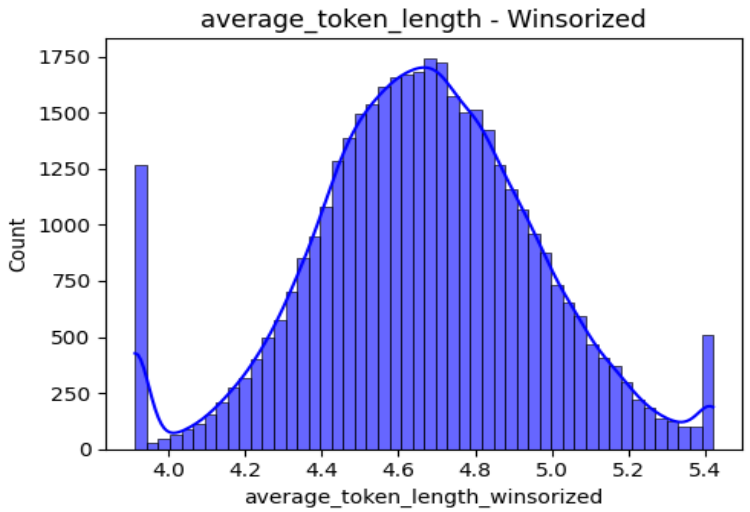
Winonisierung

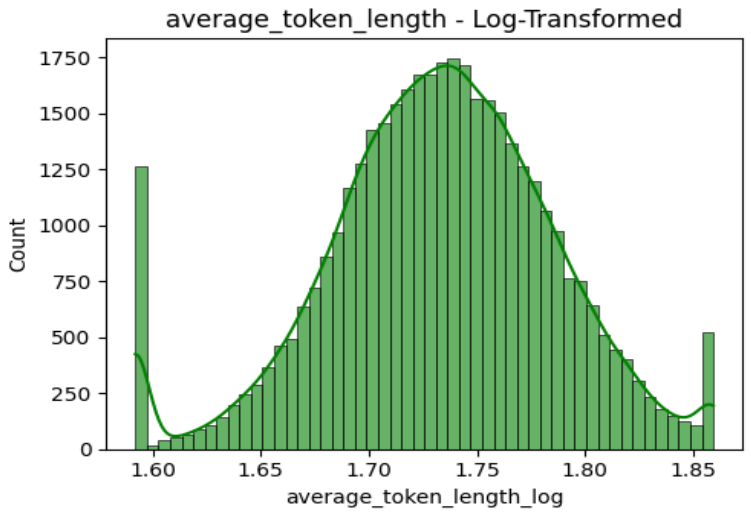
Winsorisierung begrenzt extreme Ausreißer, indem Werte oberhalb bzw. unterhalb eines bestimmten Quantils auf feste Grenzen gesetzt werden. So werden extreme Schwankungen reduziert, ohne Daten zu entfernen, was die Robustheit der Daten erhöht.

Log-Transformation

Die Log-Transformation (log(1+x)) komprimiert die Skala der winsorized Daten und reduziert die Schiefe, wodurch die Verteilung symmetrischer und die Modellierung stabiler wird.







MODELLIERUNG-BASELINE MODELLE

- **Lineare Regression:**
Dieses Modell ermittelt eine lineare Beziehung zwischen den Eingangsvariablen und der Zielgröße, was es sehr einfach zu implementieren und gut interpretierbar macht – ideal als Ausgangspunkt, um erste Hypothesen zu testen.

- **Random Forest (Baseline):**
Random Forest kombiniert mehrere Entscheidungsbäume zu einem Ensemble, was die Robustheit erhöht und Überanpassung reduziert; es dient als solider Referenzwert, da es trotz seiner Komplexität relativ leicht zu verstehen ist.

- **Gradient Boosting:**
Bei Gradient Boosting wird ein sequentielles Ensemble aus schwachen Lernmodellen aufgebaut, wobei jedes Modell die Fehler des vorherigen korrigiert, um so eine höhere Vorhersagegenauigkeit zu erreichen; es bietet eine interessante Alternative, da es häufig sehr präzise Ergebnisse liefert.

TRANSFORMER-ANSATZ IM NOTEBOOK

•**Zielsetzung:**
Vortrainierte Transformer-Modelle werden genutzt, um semantische Textrepräsentationen (Embeddings) aus Nachrichtenartikeln zu generieren.

•**Was wird demonstriert?**
•Im Notebook wird das vortrainierte Modell **all-MiniLM-L6-v2** aus der SentenceTransformer-Bibliothek verwendet.
•Für eine Liste von Beispieltexten werden Embeddings berechnet – diese Vektoren fassen komplexe linguistische und semantische Informationen kompakt zusammen.
•Die Berechnung der Embeddings dauert ca. 2 Sekunden, was zeigt, dass die Methode effizient in der Feature-Generierung ist.

Vorteile und Integration:

•**Semantisches Verständnis:** Der Transformer erfasst Kontext und Bedeutung von Wörtern, was klassischen Bag-of-Words-Ansätzen überlegen ist.

•**Ergänzende Features:** Die generierten Embeddings können als zusätzliche Features in klassische ML-Modelle (z. B. Gradient Boosting) integriert werden, um die Vorhersagekraft zu erhöhen.

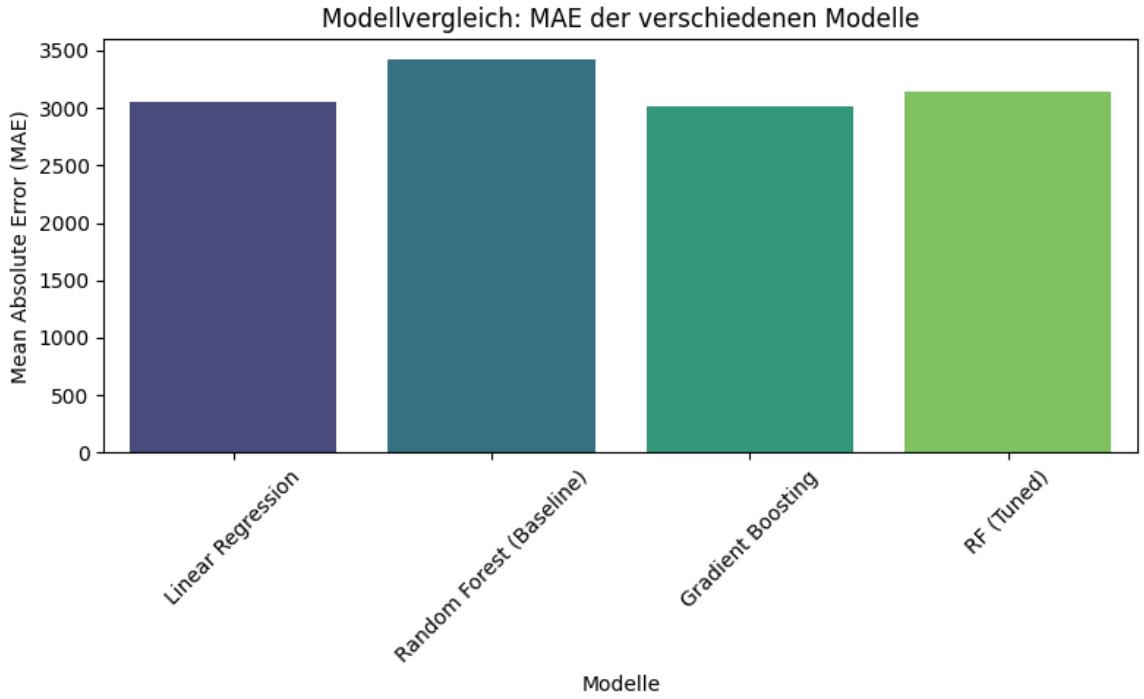
•**Effizienz:** Trotz hoher Komplexität liefert der vortrainierte Transformer schnelle Ergebnisse, sodass der zusätzliche Rechenaufwand minimal bleibt.

MODELL-EVALUATION: TRAININGSZEITEN & MAE

•**Lineare Regression:**
•Trainingszeit: ca. 0,79 s, Inferenzzeit: ca. 0.0412 s, MAE: 3057.24

•**Random Forest (Baseline):**
•Trainingszeit: ca. 315,45 s, Inferenzzeit: ca. 0.56 s, MAE: 3424.93

•**Gradient Boosting:**
•Trainingszeit: ca. 57,12 s, Inferenzzeit: ca. 0.0247 s, MAE: 3009.45

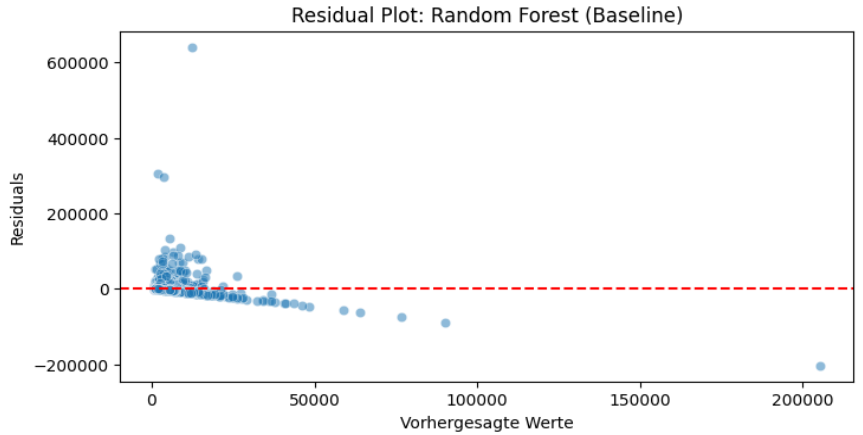


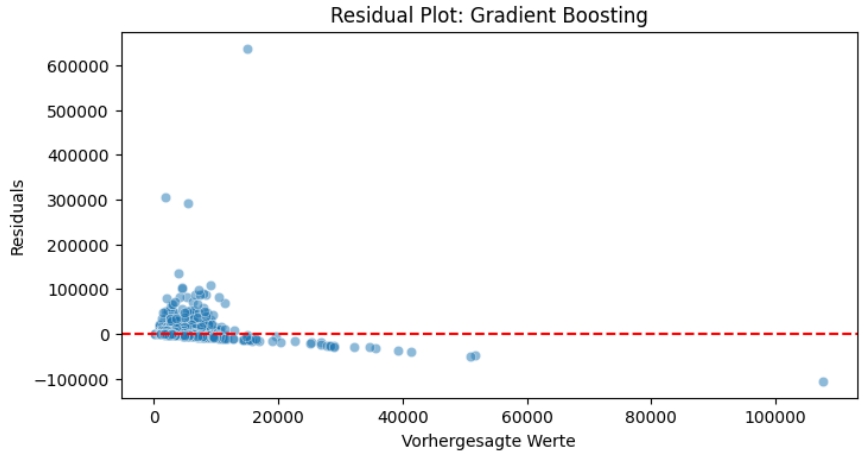
RESIDUAL PLOTS

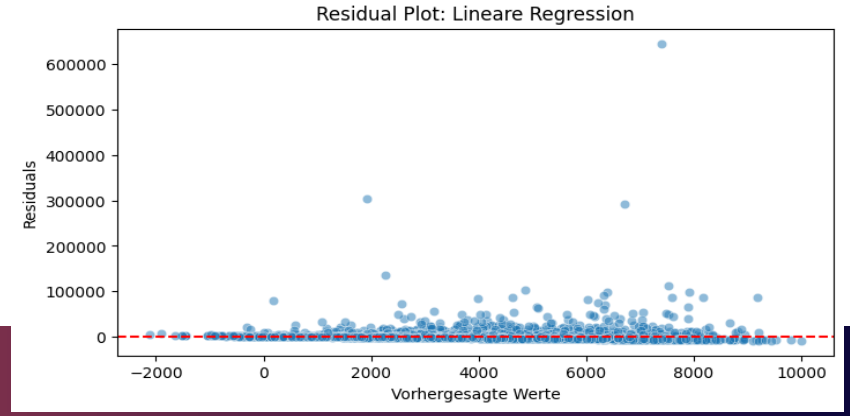
•**Lineare Regression:**
•Residuen meist nahe Null, aber starke Ausreißer bei sehr hohen Shares.
•Deutet auf recht gute Vorhersage für den Großteil der Daten hin, jedoch Schwächen bei Extremwerten.

•**Random Forest (Baseline):**
•Breiterer Streubereich, einzelne sehr große Abweichungen.
•Trotz Robustheit Schwierigkeiten bei seltenen Extremwerten.

•**Gradient Boosting:**
•Viele Residuen konzentriert um Null, dennoch Ausreißer im hohen Bereich.
•Gutes Gesamtbild, aber auch hier bleiben Extremwerte problematisch.



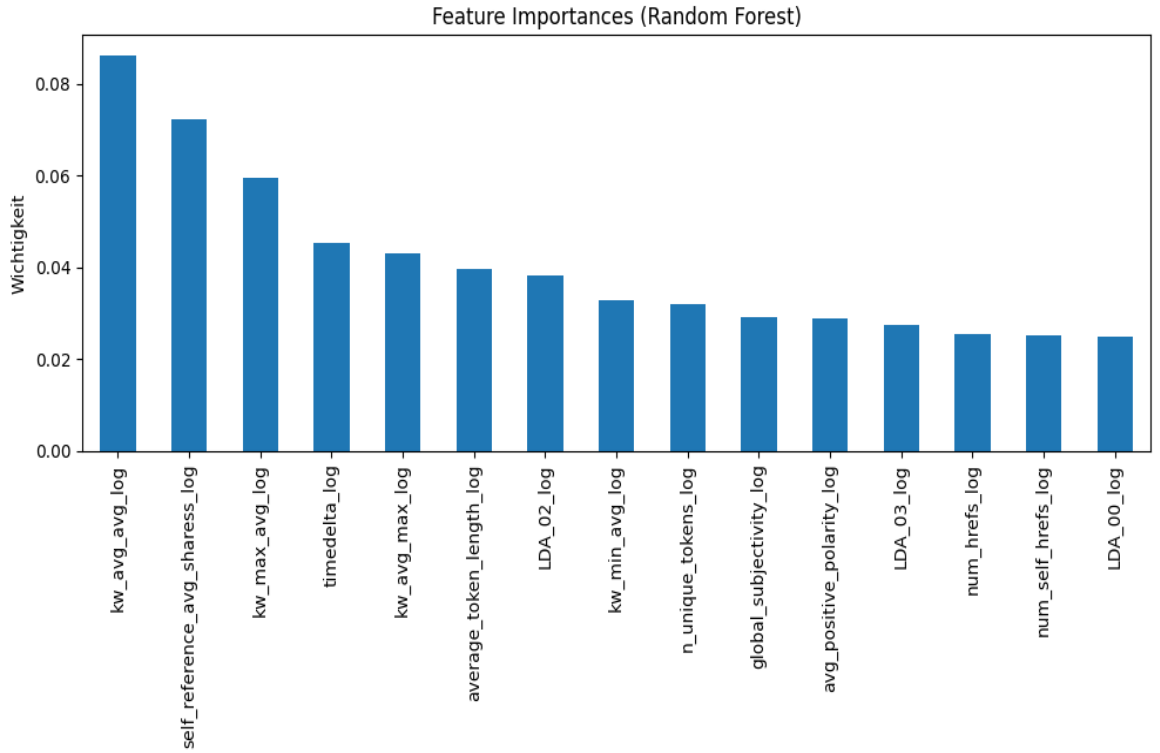




FEATURE IMPORTANCE-RANDOM FOREST

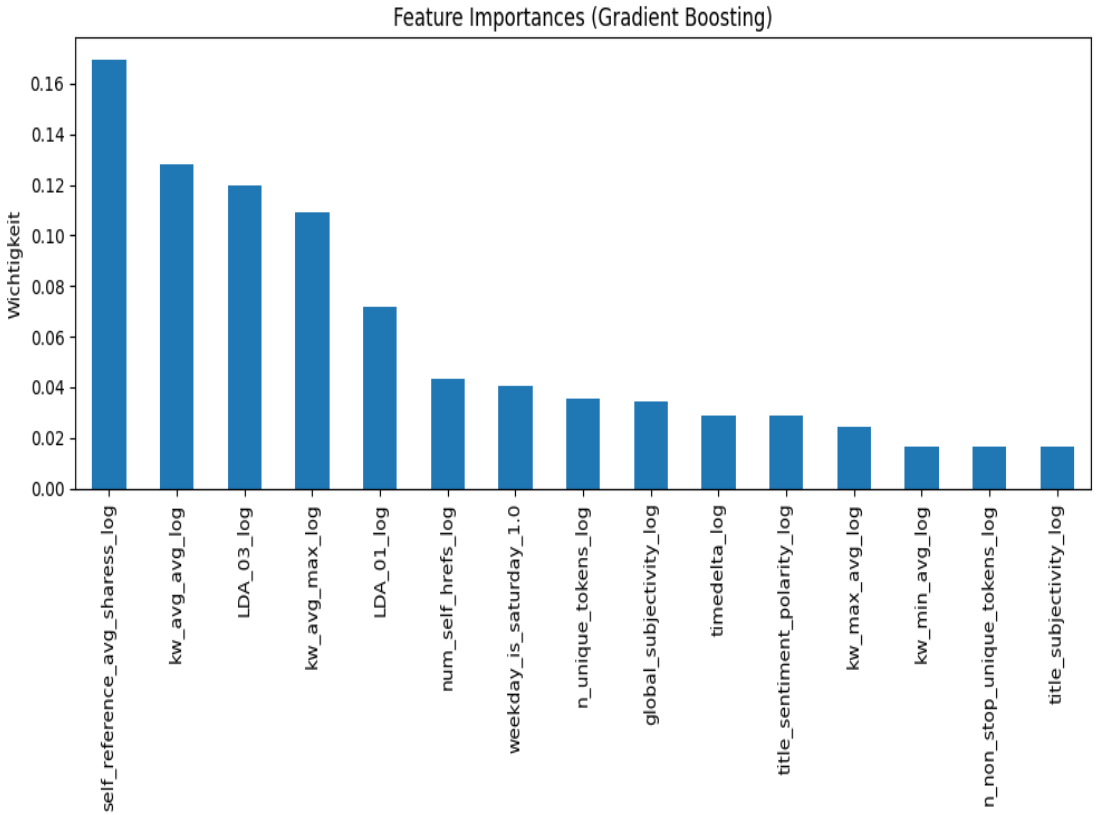
•*kw_avg_avg_log* und *self_reference_avg_sharess_log* stehen an der Spitze, was darauf hinweist, dass Durchschnittswerte bestimmter Keyword-Metriken sowie die selbstreferenzierten Shares großen Einfluss haben.
•Auch *timedelta_log* (Zeitfaktor) und *average_token_length_log* (Textlänge) spielen eine Rolle.

•**Fazit:**
•Das Modell stützt sich stark auf bestimmte Keyword-Kennzahlen und Textmerkmale.
•Dies liefert Hinweise darauf, wie Artikel strukturiert sein sollten, um mehr Shares zu generieren (z. B. Keywords oder selbstreferenzierte Shares im Artikel).



FEATURE IMPORTANCE-GRADIENT BOOSTING

•*self_reference_avg_sharess_log* ist das wichtigste Merkmal, was nahelegt, dass die selbstreferenzierten Shares (z. B. interne Verlinkungen oder frühere Shares) einen großen Einfluss auf die Vorhersage haben.
•*kw_avg_avg_log* (durchschnittliche Keyword-Metriken) und *LDA_03_log* (Themenverteilung) verdeutlichen, dass sowohl inhaltliche Aspekte (Keywords) als auch thematische Faktoren (LDA) maßgeblich sind.
•*num_self_hrefs_log* und *weekday_is_saturday_1.0* weisen darauf hin, dass sowohl interne Verlinkungen als auch der Veröffentlichungszeitpunkt (insbesondere Samstage) eine Rolle spielen.
•**Fazit:**
•Das Modell legt besonderen Wert auf Keyword-Strukturen, thematische Merkmale und Veröffentlichungsfaktoren.



NACHHALTIGKEITSANALYSE – RESSOURCENEFFIZIENZ

Lineare Regression:

- Kürzeste Trainingszeit und geringster SpeicherverbrauchEinfachstes Modell, aber ggf. weniger genau

Random Forest:

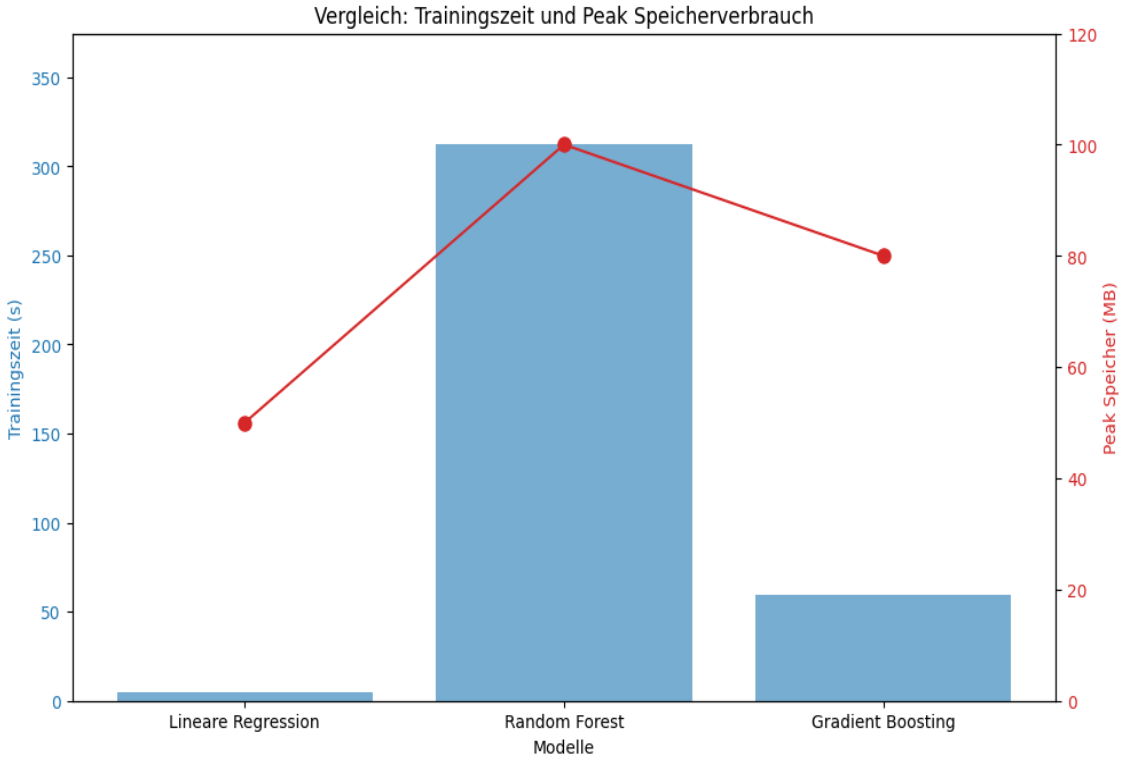
- Höchste Trainingszeit und höchster Speicherverbrauch
- Ensemble aus vielen Bäumen erfordert viel Rechen- und Speicherressourcen

Gradient Boosting:

- Moderate Trainingszeit, Speicherbedarf zwischen LR und RFLiefert in der Regel gute Vorhersagequalität bei relativ geringem Mehraufwand

Fazit:

- Gradient Boosting ist ein guter Kompromiss zwischen Modellgüte und Ressourceneinsatz.



NACHHALTIGKEITSDISKUSSION – KLASSISCHE VS. MODERNE ANSÄTZE

•**Transformer-Ansatz:**
•Vortrainierte Modelle (z. B. all-MiniLM-L6-v2) generieren Embeddings in ca. 2 Sekunden.
•Einmalige Feature-Generierung, die komplexe semantische Informationen komprimiert und ressourcenschonend ausgeführt werden kann.

•**Baseline Modelle:**
•**Lineare Regression:** Sehr kurze Trainingszeit (~0,8 s) und minimaler Speicherverbrauch, aber einfache Modellstruktur.
•**Random Forest:** Sehr lange Trainingszeit (~315 s) und hoher Speicherverbrauch aufgrund des Ensembles von Bäumen.
•**Gradient Boosting:** Moderater Trainingsaufwand (~57 s) mit gutem Kompromiss zwischen Genauigkeit und Ressourcenverbrauch.

Nachhaltigkeitsvergleich:

•**Transformer + Klassisches Modell:** Durch die einmalige, schnelle Berechnung der Transformer-Embeddings können diese als zusätzliche Features in ressourcenschonende Modelle (z. B. Gradient Boosting) integriert werden.

•Der Transformer-Ansatz liefert einen signifikanten Mehrwert im semantischen Verständnis, ohne den Gesamtressourcenverbrauch übermäßig zu erhöhen.

LIMITATIONEN & AUSBLICK

Limitationen:

- **Lange Trainingszeiten & hoher Ressourcenverbrauch:**
  - Insbesondere bei komplexen Modellen wie Random Forest und beim Hyperparametertuning.

- **Datenqualität:**
  - Unausgewogene Verteilungen und Ausreißer können die Modellleistung beeinträchtigen.

Ansätze zur Optimierung:

- Einsatz effizienterer Tuning-Strategien (z. B. Bayesian Optimization statt GridSearchCV).
- Erweiterung des Feature-Engineerings durch zusätzliche, domänenspezifische Merkmale.
- Skalierung auf größere Datensätze und Nutzung spezialisierter Hardware (GPU, parallele Verarbeitung).

LIMITATIONEN & AUSBLICK

Ressourcenschonende Infrastruktur – Ausblick

•**Erneuerbare Energien:**
•Nutzung von Solar- und Windenergie in Rechenzentren, um den CO_2-Fußabdruck zu minimieren.
•Integration von Energiespeichersystemen zur Stabilisierung des Netzbetriebs.

•**Moderne Rechenzentrumsarchitektur:**
•Einsatz von energieeffizienten Kühlsystemen (z. B. Flüssigkeitskühlung, Free Cooling) zur Reduzierung des Stromverbrauchs.
•Virtualisierung und Containerisierung, um Hardware optimal auszulasten und den Energieverbrauch zu senken.

•**Skalierbare Cloud-Infrastrukturen:**
•Nutzung von Cloud-Plattformen, die auf erneuerbare Energien setzen und eine dynamische Skalierung erlauben.
•Automatisierte Lastverteilung und Optimierung der Workloads, um Überkapazitäten zu vermeiden.

•**Zukünftige Entwicklungen:**
•Forschung an selbstoptimierenden, grünen Rechenzentren.
•Weiterentwicklung von energieeffizienten Algorithmen und Hardwarelösungen.

FAZIT

- **Wichtigste Ergebnisse:**

- **Modellvergleich:**
  - Lineare Regression: Sehr ressourcenschonend, aber einfache Modellstruktur.
  - Random Forest: Hoher Ressourcenverbrauch, aber robuste Vorhersagen.
  - Gradient Boosting: Guter Kompromiss zwischen Genauigkeit und Effizienz.
- **Feature Importance:**
  - Relevante Features (z. B. Keyword-Metriken, interne Verweise) haben einen signifikanten Einfluss auf die Vorhersagen.
- **Nachhaltigkeitsaspekte:**
  - Der Einsatz vortrainierter Transformer-Modelle ermöglicht eine schnelle und effiziente Generierung von semantischen Textfeatures, was den Ressourcenverbrauch deutlich reduziert.

- **Schlussfolgerung:**
- Der innovative, ressourcenschonende Ansatz kombiniert klassische ML-Modelle mit modernen Transformer-Methoden, um sowohl hohe Vorhersagegenauigkeit als auch Nachhaltigkeit zu erreichen.

VIELEN DANK

Monika Bernecker

Monika.Alber@gmx.net