Implementierung und Evaluation von Machine Learning-Ansätzen zur Sentiment-Analyse im Deutschen

Antrittsvortrag Bachelorarbeit

Niklas Donhauser
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Universität Regensburg

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Vorstellung

Name	Niklas Donhauser
Semester	6. Fachsemester
Fächerkombination	Medieninformatik (1.Hf.)
	Medienwissenschaft (2.Hf.)
Betreuer	Jakob Fehle
Erstgutachter	Prof. Dr. Christian Wolff
Zweitgutachter	Prof. Dr. Udo Kruschwitz
Status	Vertiefte Einarbeitung

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Thema & Hintergrund

- Vergleich verschiedener Machine Learning (ML)-Ansätze im Deutschen
  - Traditionelle Methoden
  - Neurale Netzwerke
  - Transformer-basierte Methoden

→ Fokus: auf Transformer-basierten Methoden

- Deutsch als „underresourced language“
- Bisherig keine flächendeckende Untersuchung verschiedener ML-Ansätze auf unterschiedlichen Korpora

→ Wie performen die verschiedenen ML-Ansätze auf den verschiedenen Korpora?
Welche Rückschlüsse kann man daraus ziehen?

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Verwandte Arbeiten

- Literaturrecherche über Google Scholar
- Untersuchung von lexikonbasierten Lösungen im Deutschen [2]
- Überblick über Deep Learning-Methoden im Englischen [1]
- Auswertung von ML-Ansätzen auf Produktbewertungen [5]
- Transformer-basierte Methode BERT [3]
- Deutsche Versionen von BERT /ELECTRA [4]

[1] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.
[2] Fehle, J., Schmidt, T., & Wolff, C. (2021). Lexicon-based sentiment analysis in german: Systematic evaluation of resources and preprocessing techniques.
[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[4] Chan, B., Schweter, S., & Möller, T. German's Next Language Model. arXiv 2020. arXiv preprint arXiv:2010.10906.
[5] Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In Cognitive informatics and soft computing (pp. 639-647). Springer, Singapore.

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Korpora

- 20 Korpora aus 5 verschiedenen Kategorien [1]
  - Literarische Texte
  - Texte aus gemischten Domänen
  - Nachrichtenartikel
  - Produktbewertungen
  - Social Media
- Insgesamt über 1,1 Millionen Texteinheiten mit Angabe des Sentiment

[1] Fehle, J., Schmidt, T., & Wolff, C. (2021). Lexicon-based sentiment analysis in german: Systematic evaluation of resources and preprocessing techniques.

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Preprocessing

- Vorverarbeitung der Korpora für bessere Ergebnisse
  - Stopp Wort Entfernung [2]
  - Stemming [3]
  - Emoji Umwandeln [1]
  - Entfernen von Hashtags / URL / Nutzernamen

[1] Parveen, H., & Pandey, S. (2016, July). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. In 2016 2nd international conference on applied and theoretical computing and communication technology (ICATCCT) (pp. 416-419). IEEE.
[2] Basarslan, M. S., & Kayaalp, F. (2020). Sentiment Analysis with Machine Learning Methods on Social Media.
[3] Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing (pp. 639-647). Springer, Singapore.

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Traditionelle Machine Learning-Ansätze

- Support Vector Machines [1]
- Naïve Bayes [1]

[1] Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In *Cognitive informatics and soft computing* (pp. 639-647). Springer, Singapore.

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Neurale Netzwerke und Transformer-basierte Ansätze

- Convolutional Neural Networks [1]
- Recurrent Neural Networks [2]
- (G)BERT
- (G)ELECTRA
- GPT-2

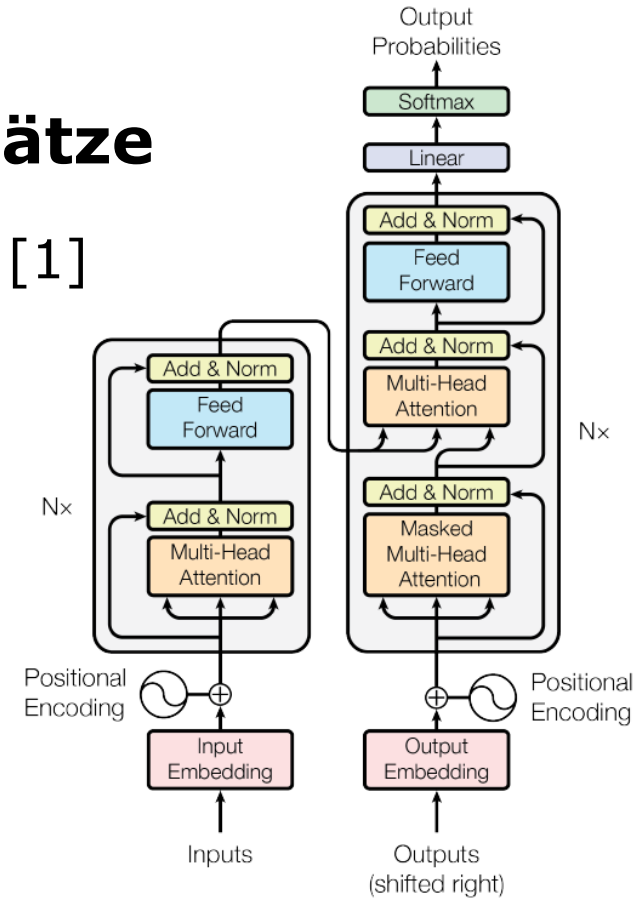


Figure 1: The Transformer - model architecture.

[1] Dos Santos, C., & Gatti, M. (2014, August). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers* (pp. 69-78).
[2] Arras, L., Montavon, G., Müller, K. R., & Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206*.
Figure: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

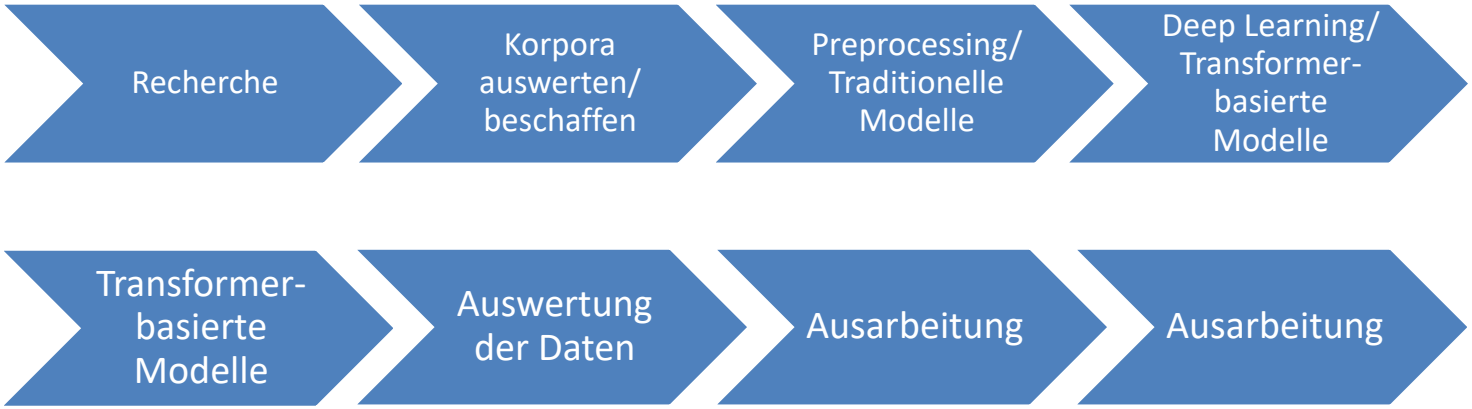
Vorgehensweise

- Eigene Modelle trainieren
- Vortrainierte Modelle (wie BERT) fein-tunen
- Evaluation der Modelle
- Aufteilen der Daten mit k-fold Cross Validation
- Metriken:
    - Accuracy
    - Precision
    - Recall
    - F1 Wert

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Zeitplan



Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Zusammenfassung

- Auswertung verschiedener ML-Ansätze und Korpora im Deutschen
- Implementierung verschiedener ML-Ansätze und Vorbereitung der Korpora
- Nächste Schritte:
  - Vorverarbeitung abschließen
  - Literaturliste erweitern

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Fragen:

- Sollte nur ein traditioneller Ansatz gewählt werden?
- Alternativen zu GPT-2?

Universität Regensburg

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Vielen Dank für Eure Aufmerksamkeit!