



FOM Hochschule für Oekonomie & Management

Hochschulzentrum DLS

Projektarbeit

im Studiengang Big Data & Business Analytics

über das Thema

Deep Learning zur Aktienkursprognose mit multimodalen Daten

von

Paul Hornig und Admir Dutovic

Dozent : M.Sc. Maher Hamid
Matrikelnummer : 701650 und <AdmirNr>
Abgabedatum : 19. Dezember 2024

Sentiment-Analyse mit Transformer-Encoder

Abstrakt


Diese Arbeit widmet sich der Forschungsfrage, inwieweit sich ein Transformer-Encoder-Modell mit verhältnismäßig wenig Parametern für eine Sentiment-Analyse eignet. Zusätzlich soll geprüft werden, wie groß der Einfluss des Umfangs an Trainingsdaten auf die Modellleistung ist. Aktuelle große Sprachmodelle nutzen Transformer-Architekturen, bei denen festgestellt wurde, dass deren Leistung besonders hervorsticht, wenn diese mit sehr vielen Daten trainiert wurden und Unmengen an lernbaren Parametern besitzen. Diese Anforderungen lassen sich nur von sehr finanzstarken Unternehmen erfüllen. Die Forschungsarbeit kann dazu beitragen, einen optimalen Punkt bei der Erreichung eines MVP-Modells (minimal funktionsfähige Iteration eines Modells) zu identifizieren. Als Leistungsrichtwert, der übertroffen werden muss, dient ein lexikonbasiertes Modell aus der "NLTK" Python-Bibliothek¹.

Datenaufbereitung

Als Datengrundlage dienen 3.6Mio Amazonreviews, welche auf Kaggle.com zur Verfügung gestellt werden².

Abbildung 1: Aufbereitung von Amazon Kundenbewertungen

label	review
__label__2	Grisham's Best!: Grisham knows just how to keep you reading. His amazing plots are believable and gripping. In 'A Time to Kill', he made you think of what you would d...
__label__1	Horrible buzzing!: The buzzing sound was so loud on this baby monitor that I could not hear the baby at all. I tried different positions for the receiver and monitor ...



label	review_tokens	review_features	feature_vec
1	grisham best grisham know keep reading amazing plot believable gri...	grisham best grisham know keep reading amazing plot believable gri...	0 0 0 0 0 0 0 0 0 0 0 0 0 1943 389 1943 2428 2393 3525 136 3254 ...
0	horrible buzzing buzzing sound loud baby monitor could hear baby t...	horrible sound loud baby monitor could hear baby tried different p...	0 2114 4122 2622 292 2...

Bei der Datenaufbereitung resultieren 2 Abwandlungen der Reviews und eine Zahlendarstellung (Abb. 1). Bei "review_tokens" handelt es sich um eine Schlagwortextraktion, bei

¹ Hutto, C., Gilbert, E., 2014.

² Bittlingmayer, A., 2019.

"review_features" um eine reduzierte Form von ersterem (5k mögliche Schlagwörter) und bei "feature_vec" um eine Zahlendarstellung von zweiterem. Zur Komplexitätsreduktion werden lediglich Zeilen mit mindestens 10 und maximal 40 Review-Features beibehalten. Falls ein Review weniger als 40 Features aufweist, wird der Zahlenvektor durch linksseitiges Padding mit 0 passend aufgefüllt. Die Daten werden in 2 Teilen gespeichert, mit 80% für das Training des Transformer-Encoders und 20% zum Testen.

Basis Modell

Das Basis Modell erreicht bei einer Anwendung auf Testdaten eine Genauigkeit von ca. 70% (Abb. 2), welche vom Transformer-Encoder-Modell übertroffen werden muss.

Abbildung 2: Basis Modell Performance

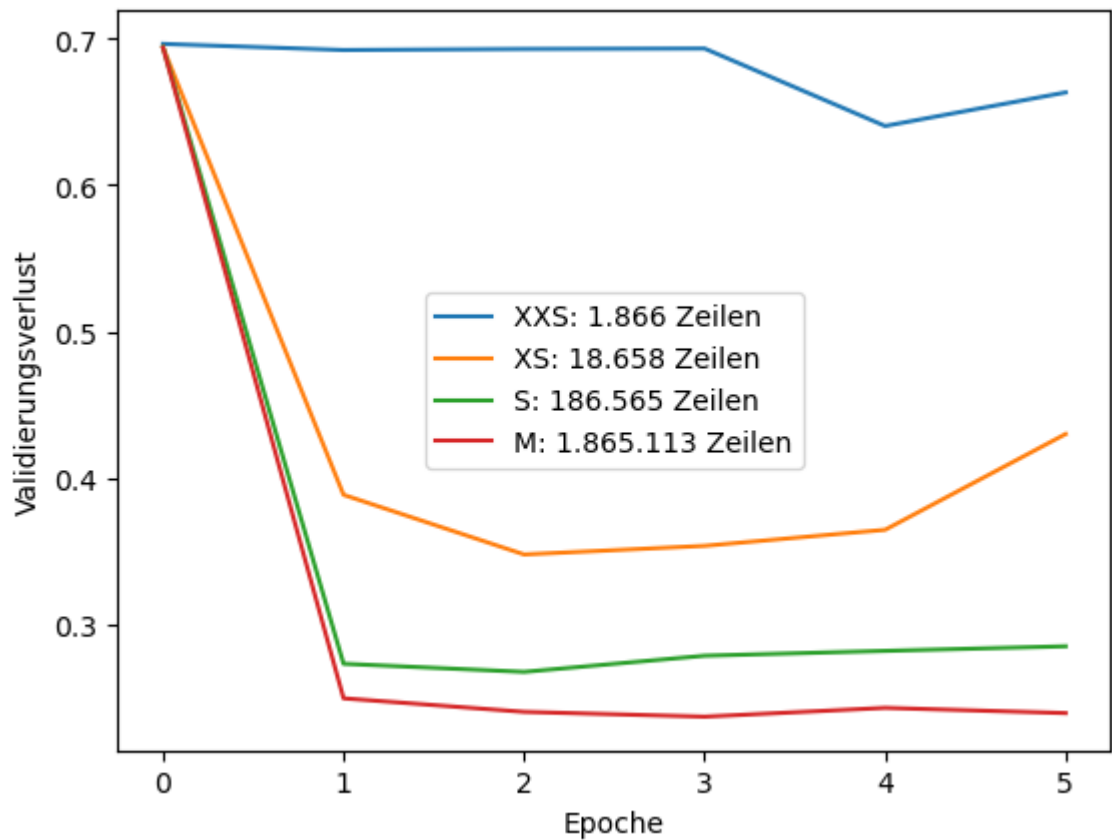
Eingabe	Genauigkeit (%)
review_tokens	69,8
review_features	69,2

Transformer-Encoder

Genaue Architekturdetails des umgesetzten Transformer-Encoders befinden sich im Anhang als Bild "model_architecture.png" und orientieren sich an einer Vorlage von Andrej Karpathy³, wobei mit dem Python-Paket "PyTorch" gearbeitet wird⁴. Als Eingabe für den Transformer-Encoder dienen die Feature-Vektoren, welche eine einheitliche Kontextlänge von 40 aufweisen. Zur Untersuchung des Einflusses der Datenmenge auf Modellleistungen werden 4 Datengrößen angewandt (Abb. 3).

³ Karpathy, A., 2024.

⁴ PyTorch Foundation, 2024.

Abbildung 3: Verlauf des Validierungsverlusts

Es ergeben sich die in Abbildung 4 dargestellten Genauigkeiten.

Abbildung 4: Transformer-Encoder-Modell Performance

Trainingsdatengröße	Genauigkeit (%)
XXS	63,5
XS	85,4
S	88,8
M	90,3

Daraus geht hervor, dass bereits wenige Trainingsdaten (XS) ausreichen, um ein lexikon-basiertes Verfahren bei einer Sentimentanalyse zu übertreffen. Außerdem wird deutlich, dass die notwendige Steigerung der Trainingsdatenmenge für eine spürbare Verbesserung der Modellleistung exponentiell mit der bereits vorhandenen Performance wächst.

Dabei ist mit einem Sättigungspunkt zu rechnen, wobei man spätestens dann dazu übergehen sollte, die Modellkomplexität durch z.B. zusätzliche Schichten zu erhöhen.

Fazit

Es sind nur wenige Trainingsdaten notwendig, um mithilfe der Transformer-Architektur ein brauchbares MVP-Modell umzusetzen. Liegt diese Voraussetzung vor, sollte man sich auch für diesen Weg entscheiden, da man dann mehr Potenzial zur Verbesserung des Produktes besitzt. Vor allem große Sprachmodelle wie GPT-4 von OpenAI sind der Beweis für die hervorragende Skalierbarkeit von Transformer-Modellen.

Quellenverzeichnis

Bittlingmayer, Adam (2019): Amazon Reviews for Sentiment Analysis, <<https://www.kaggle.com/datasets/bittlingmayer/amazonreviews>> (2019) [Zugriff: 2024-07-02]

Hutto, C.J., Gilbert, E.E. (2014): VADER Sentiment Analysis, <<https://github.com/cjhutto/vaderSentiment>> (2014) [Zugriff: 2024-07-03]

Karpathy, Andrej (2024): nanoGPT, <<https://github.com/karpathy/nanoGPT>> (2024-06-03) [Zugriff: 2024-07-03]

PyTorch Foundation (2024): PyTorch, <<https://pytorch.org/>> (2024) [Zugriff: 2024-07-04]