

Sentimentanalyse deutscher Twitter-Korpora mittels Deep Learning

Enrico Bunde

Masterarbeit • Studiengang Wirtschaftsinformatik • Fachbereich Wirtschaft • 29.03.2019

Motivation

Ziel der Arbeit ist es eine Sentimentanalyse deutscher Twitter-Korpora mittels Deep Learning durchzuführen. Dabei wurde ein CNN und LSTM implementiert sowie vortrainierte Wortembeddings eingesetzt. Es wurden explorative Experimente für die Optimierung der Hyperparameter durchgeführt. Außerdem wurden zwei prototypische Implementierungen vorgenommen. Zum einen LIME, einem Ansatz, Vorhersagen von Machine Learning Modellen zu erklären und zum anderen eine Rastersuche zur Optimierung von Parametern.

Wortembeddings und Twitter-Korpora

Es wurden vortrainierte Wortembeddings eingesetzt. Dabei lagen zwei Ansätze von Wortembeddings vor. Zum einen Word2Vec und zum anderen fastText, was eine Erweiterung des Word2Vec Modells darstellt.

	Word2Vec	fastText
Anzahl der Worte	608.130	533.796
Trainingsdaten	Wikipedia Artikel News Artikel	50 Millionen Tweets

Tabelle 1: Übersicht zu den vortrainierten Wortembeddings

Diese Wortembeddings lassen sich mittels eines CBOW oder Skip-gram Modells umsetzen. Hier wurde Skip-gram genutzt. Dabei dient ein Wort als Eingabe und soll ein weiteres Wort in einem Kontext klassifizieren.

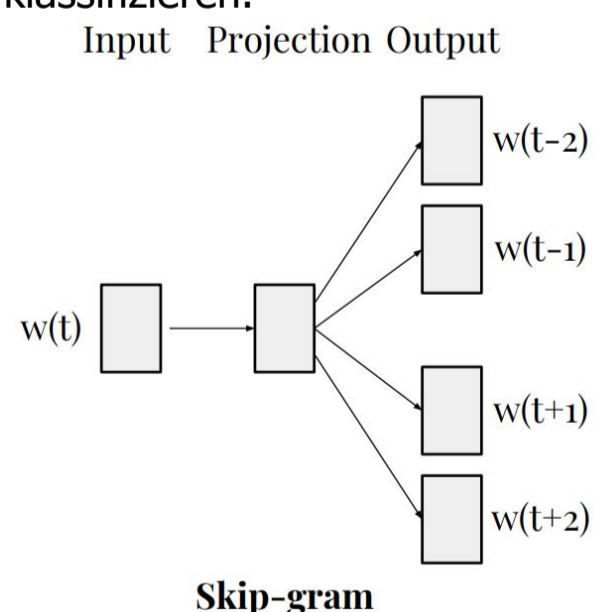


Abbildung 1: Grafische Darstellung des Skip-gram Modells

Der Twitter-Korpus wurde aus drei publizierten Korpora zusammengefügt.

Sentiment	SB-10k	DAI	GTS	gesamt
negativ	1.056	267	15.381	16.704
neutral	4.367	1.136	47.427	52.930
positiv	1609	385	21.348	23.342
gesamt	7.032	1.788	84.156	92.976

Tabelle 2: Übersicht der Tweets pro Korpus und Gesamt

Konzeption

Für die Optimierung der Parameter wurde ein Konzept erstellt, um explorative Experimente durchführen zu können. Dabei ging es um Parameter wie die Anzahl der Neuronen oder Schichten, verschiedene Optimierungsalgorithmen, Wortschatzgröße und Tweet Länge. Weiterhin wurde Regularisierung mittels Early Stopping, Dropout, L1 und L2 untersucht.

Prototypische Implementierung Rastersuche und Lime

Mittels der Rastersuche wird es möglich ein Deep Learning Modell mit verschiedenen Parametereinstellungen auszuführen. Dabei werden z. B. die Optimierungsalgorithmen SGD, RMSprop und Adagrad nacheinander eingesetzt und verglichen. Bei Lime handelt es sich um einen Ansatz Vorhersagen zu erklären. Diese Erklärung kann als HTML-Seite gespeichert werden. Dabei zeigt die Abbildung 2 eine solche Seite. Hier wurden keine Wortembeddings implementiert. Es wurde Klasse 1 (neutral) vorhergesagt, wobei Klasse 0 (negativ) korrekt wäre. Weiterhin ist zu erkennen, welche Worte, was für Einfluss auf diese Vorhersage gehabt haben.

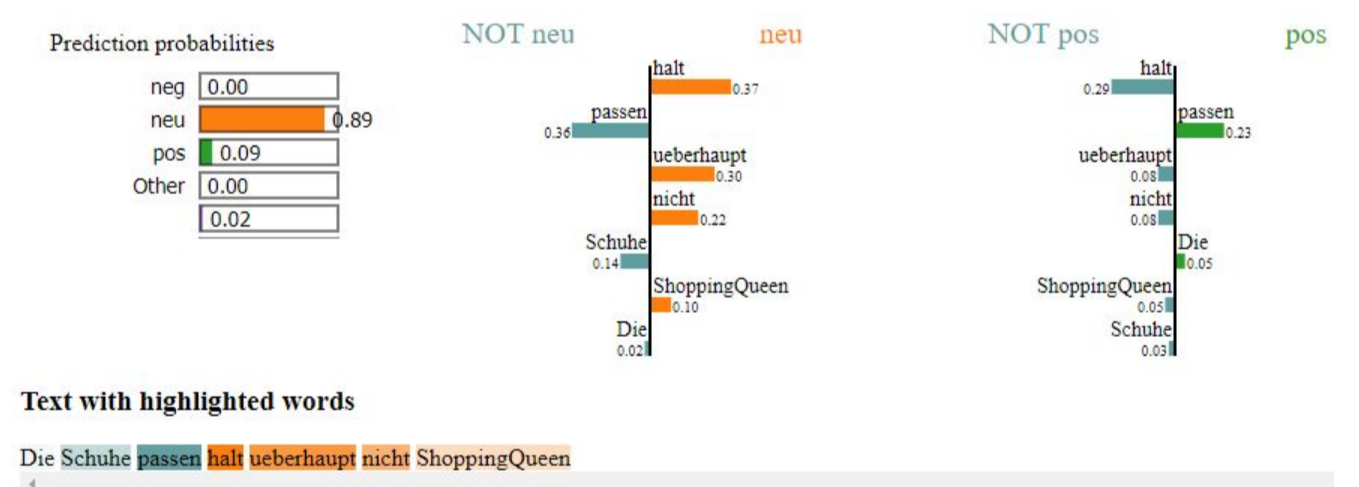


Abbildung 2: HTML-Seite der LIME Erklärung an einem Beispiel Tweet

Ergebnisse, Fazit und Ausblick

Das beste Ergebnis für den Testdatensatz lautet 54,61 % für den F1-Wert. Ein publiziertes Ergebnis für den GTS-Korpus lautet 53,60 % und die Ergebnisse zu SB-10k sind im Bereich zwischen 47,30 % und 65,09 %. Somit sind die Ergebnisse durchaus vergleichbar. Wobei dort wesentlich größere Systeme und Wortembeddings zum Einsatz kamen. Die Wortembeddings zu SB-10k wurden beispielsweise an 300.000 Millionen Tweets trainiert.

Mittels größeren Wortembeddings oder einer praktischen Implementierung von LIME oder der Rastersuche könnte die Leistung noch gesteigert werden. Auch die Ergänzung bzw. Vergrößerung des Datensatzes wäre möglich. Die drei Klassen sind sehr unausgeglichen, mit zusätzlichen positiven und negativen Tweets könnte der Korpus vergrößert und die Klassenverteilung angeglichen werden.