Karlsruhe Institute of Technology

Kooperationsseminar

Ausgewählte Anwendungen von Textmining – Sentiment Analyse von Tweets

Adrian Oberföll, Marius Kempf

Institut für Angewandte Informatik und Formale Beschreibungsverfahren

Agenda

- Sentiment Analyse – Motivation
- Datensatz *Sentiment140*
- Vorbereitung der Tweets
- Baseline mit VADER Sentiment
- Entwickelte Modelle
  - Neuronales Netz
  - Convolutional Neural Network
- Ergebnisse
- Zusammenfassung

Sentiment Analyse - Motivation

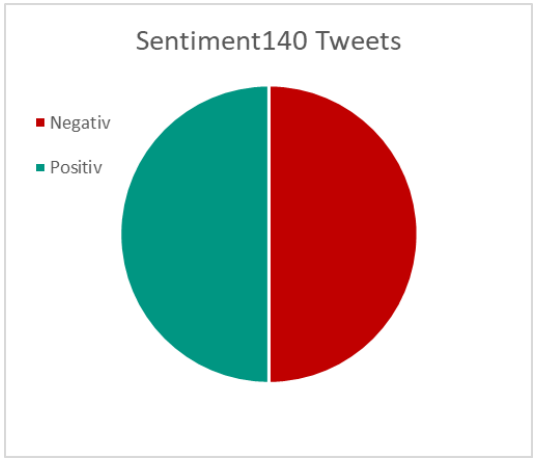
- Twitter: täglich mehr als 500 Millionen Tweets (2013)
- Facebook : 2,3 Milliarden monatlich aktive Nutzer (4/2018)

- Stimmung und Meinungen zu
  - Produkten
  - Unternehmen
  - Politiker/innen
  - Technologien

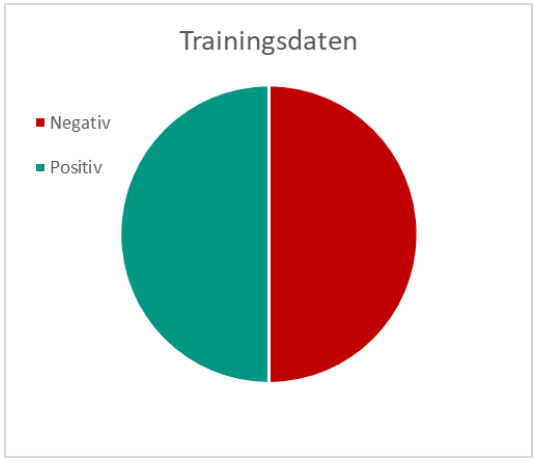
Quelle: https://de.statista.com

Datensatz – *Sentiment140*

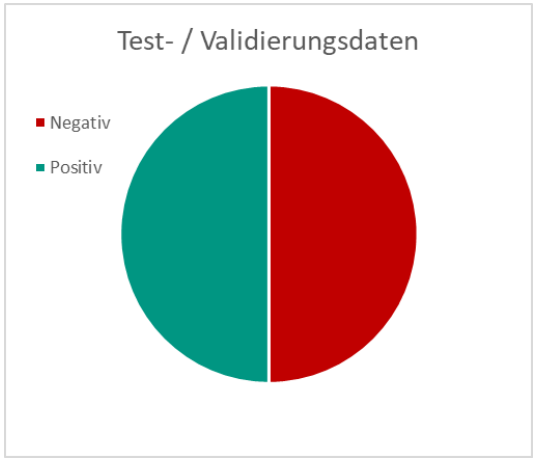
- 1.6 Millionen gelabelte Tweets
- **Binäres Klassifikationsproblem** in positiv oder negativ
- Performancebewertung: $Accuracy = \frac{Anzahl\ korrekt\ klassifizierter\ Tweets}{Anzahl\ aller\ Tweets}$



1,6 Mio. Tweets



1,564 Mio. Tweets



36 Tsd. Tweets

Vorbereitung der Tweets

Umwandlung aller Großbuchstaben in Kleinbuchstaben

Löschen der User-Tags (@user)

Löschen von „http“ & „www“ Links

Auflösen von Wortverbindungen („aren‘t“ → „are not“ etc.)

Entfernen aller Zeichen außer Buchstaben

	date	text
0	Sat Feb 02	just bought a new tesla took a pic with my apple iphone is not that awesome

Baseline Modell – VADER Sentiment

- VADER – „**V**alence **A**ware **D**ictionary for s**E**ntiment **R**easoning”
- Lexikon- und regelbasiertes Analysetool

- Beispielsatz: ***„Vader is a super cool tool!“***

Sentiment Metric	Score
Positiv	68,40%
Neutral	31,60%
Negativ	0,0 %
Compound	0,7574

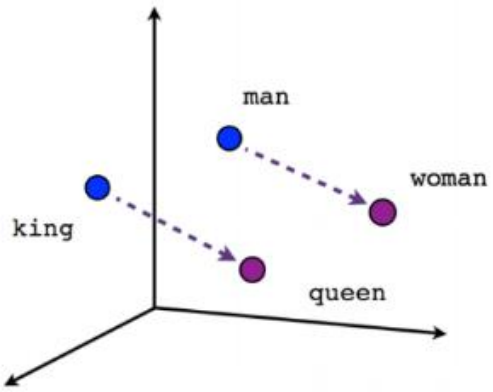
Verwendetes Tool	Accuracy Score
VADER	66,67 %
TextBlob	62,73 %
SentiWordNet	60,97 %

VADER Sentiment: https://github.com/cjhutto/vaderSentiment, (Hutto & Gílbert 2014)

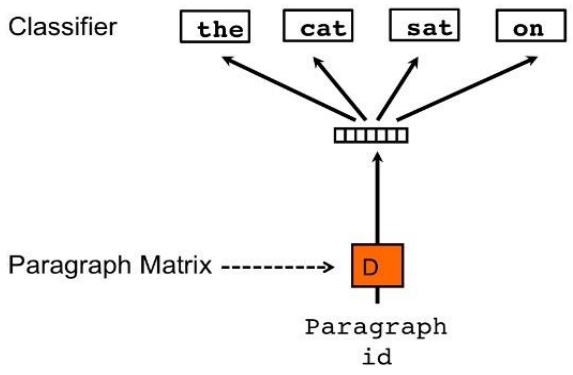
Word2Vec und Doc2Vec

- Training mit allen 1.6 Mio. Tweets möglich, da **Unsupervised Learning**
- Bessere Ergebnisse wurden jeweils mit **CBOW** (Word2Vec) bzw. **DBOW** (Doc2Vec) erzielt

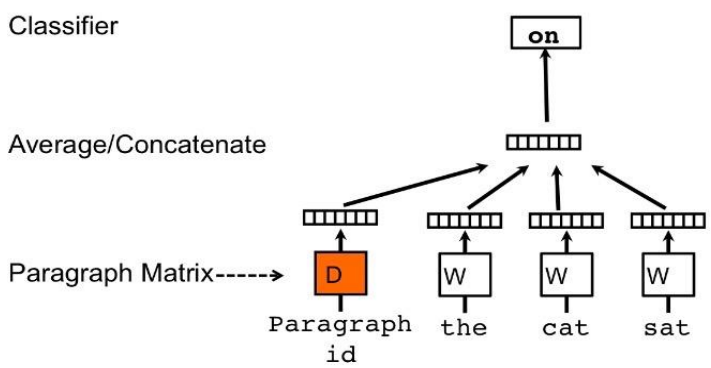
Word2Vec



Doc2Vec



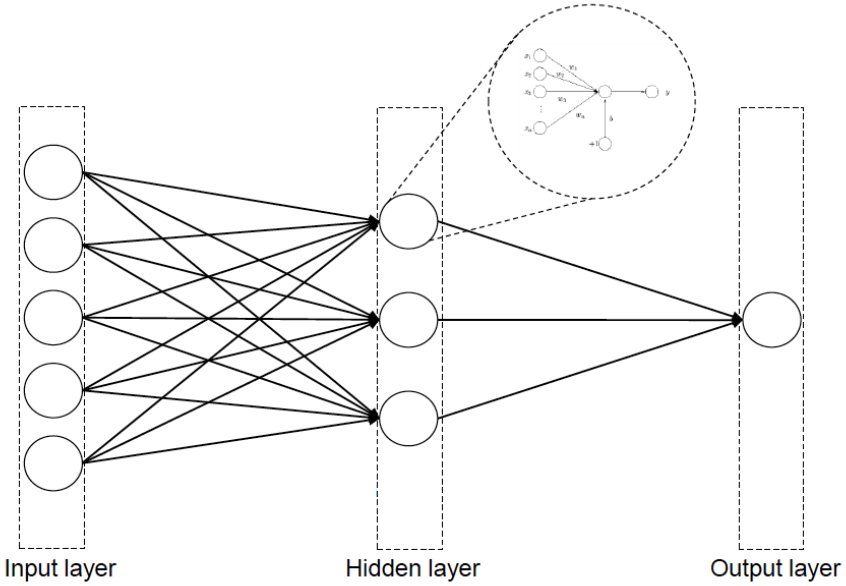
Distributed Memory - DM



Distributed Bag of Words - DBOW

Neuronale Netze mit Doc2Vec

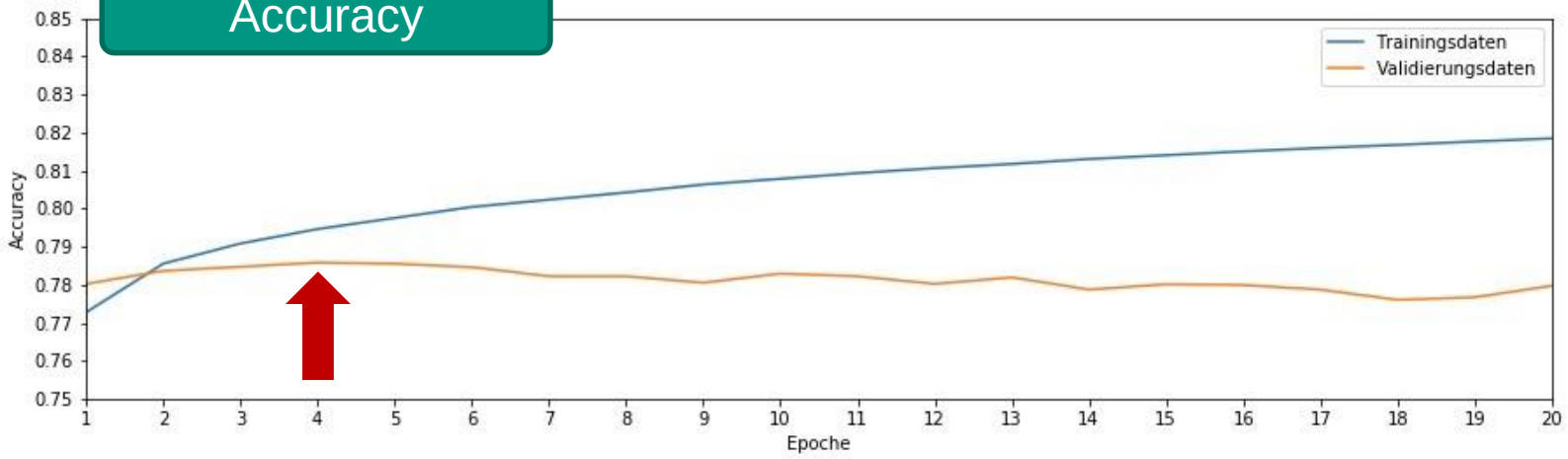
Doc2Vec Vektoren



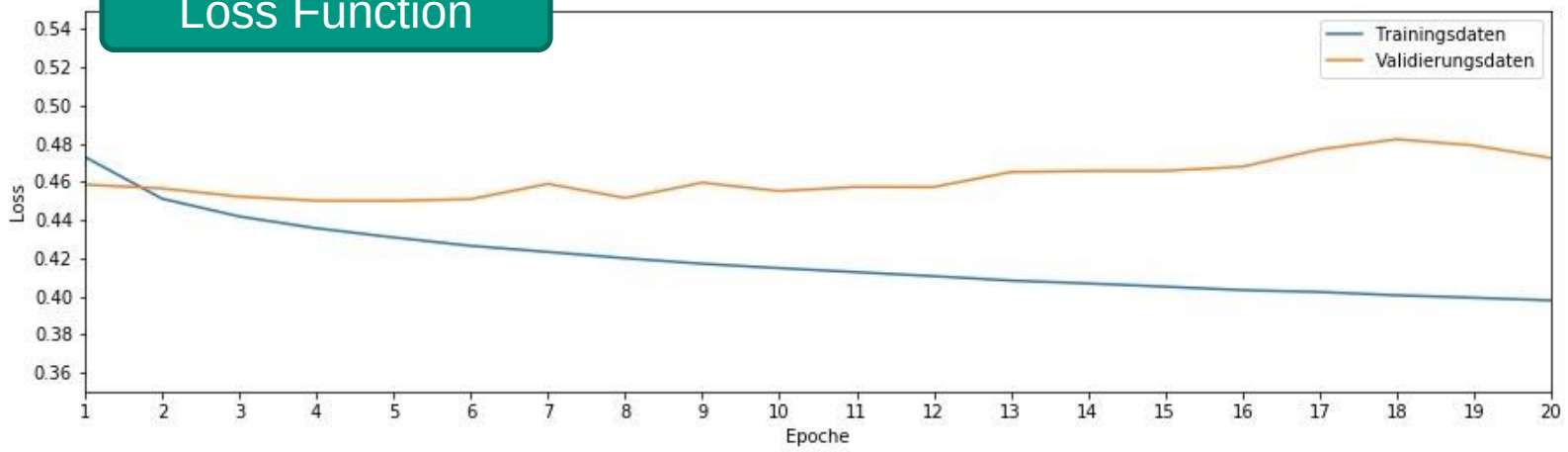
Neuronales Netz	Accuracy
NN 1 – 1 Hidden Layer (64 Neuronen)	77,87 %
NN 2 – 2 Hidden Layer (je 128 Neuronen)	78,56 %
NN 3 – 3 Hidden Layer (je 256 Neuronen)	78,58 %

Neuronales Netz – NN 3

Accuracy



Loss Function



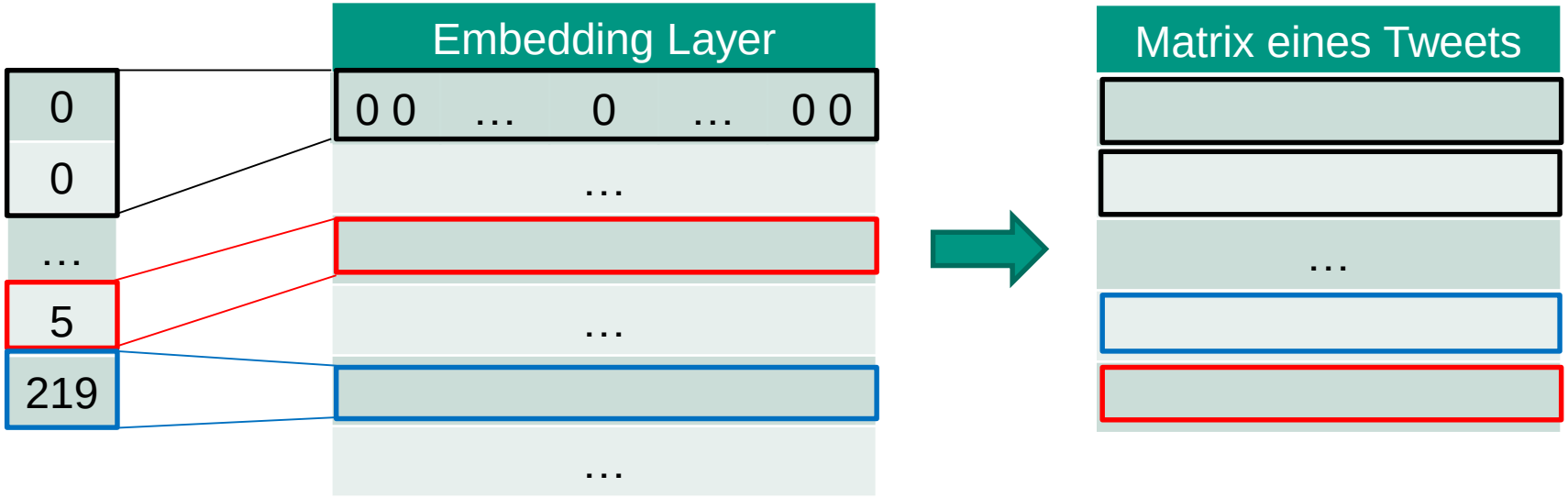
Convolutional Neural Networks (1)

this is a tweet

Jedem Wort wird ganzzahliger Wert zugeordnet

Tweets durch Nulleinträge auf einheitliche Länge bringen

[0 0 0 ... 0 0 0 30 4 5 219]



Convolutional Neural Networks (2)

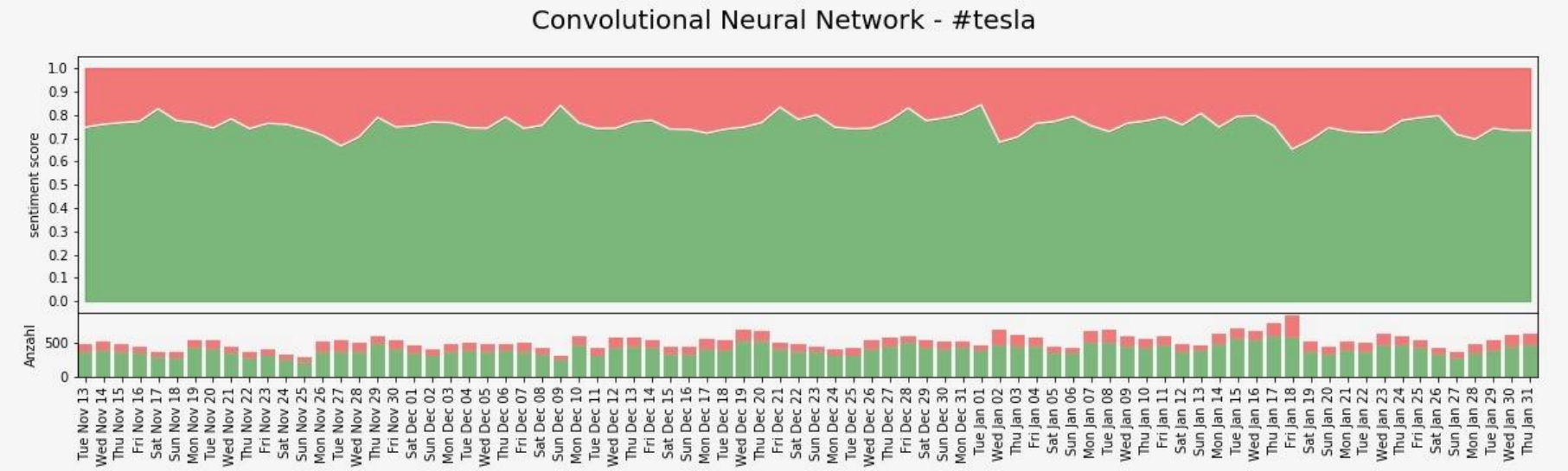
- **Ansätze**
  - CNN erlernt Word-Embeddings von Embedding Layer selbst
  - CNN mit Word2Vec
    - Vorgabe der Vektoren von Word2Vec für Embedding Layer
    - Vektoren während dem Modelltraining statisch oder dynamisch

- **Architektur** der CNNs
  - Je ein Embedding-, Convolutional- und Pooling Layer
  - Einfache Schicht mit 256 Neuronen

Convolutional Neural Network	Accuracy
CNN 1 (selbst gelernte Word-Embeddings)	82,60 %
CNN 2 (Word2Vec - static Word-Embeddings)	82,15 %
CNN 3 (Word2Vec - trainable Word-Embeddings)	83,05 %

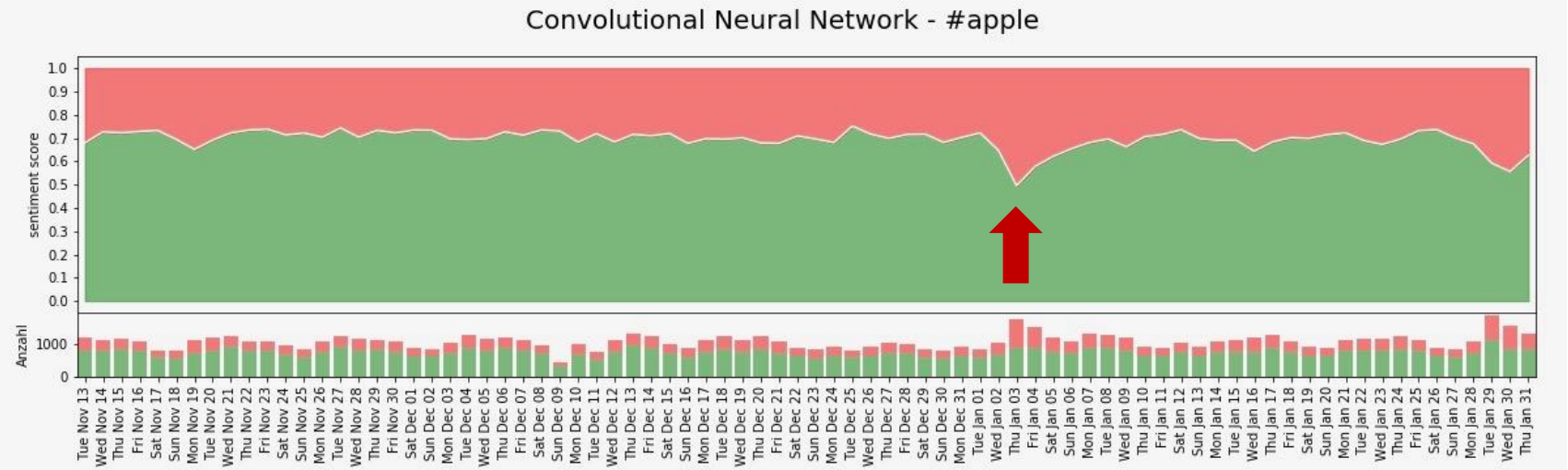
Ergebnisse - #tesla

- Kurve zeigt den relativen Anteil von positiven Tweets pro Tag
- Balkendiagramm verdeutlicht die tägliche Anzahl von positiven und negativen Tweets mit dem Hashtag #tesla
- Durchschnittlicher Score: **75,67 %**



(Zeitraum: 13.11.2018 – 31.01.2019)

Ergebnisse - #apple (1)



Apple senkt Umsatzerwartung

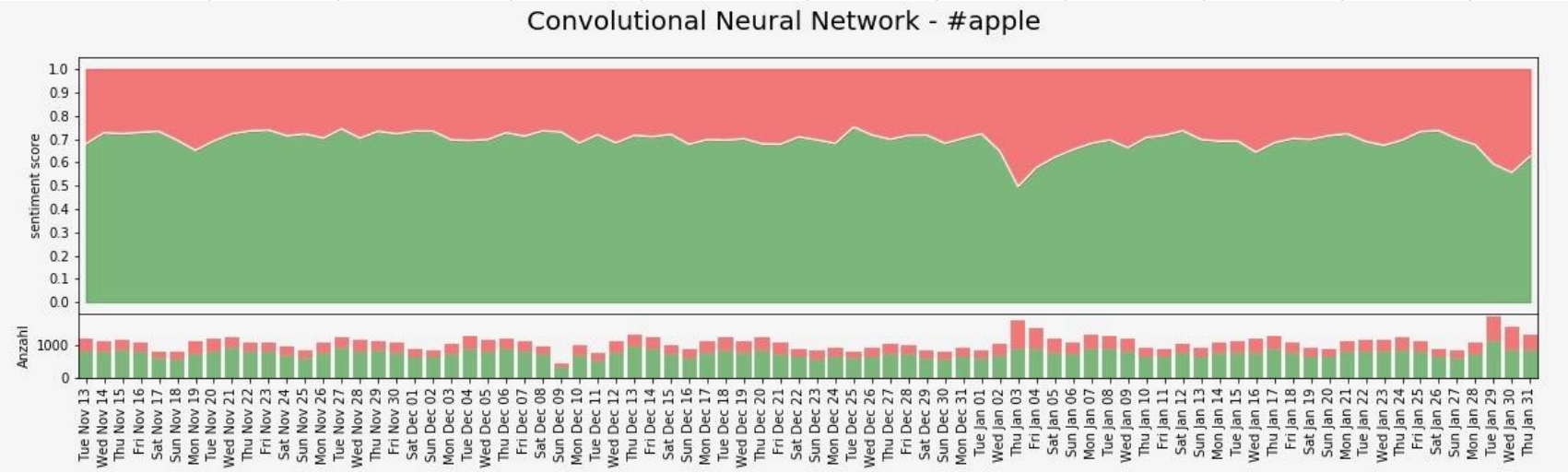
Teure iPhones verkaufen sich schlecht

Stand: 03.01.2019 04:27 Uhr

Das Weihnachtsgeschäft lief nicht gut für Apple. Der Konzern muss deshalb seine Umsatzerwartung senken. Der Handelsstreit mit China sei schuld, sagt Apple-Chef Cook. Analysten sehen das etwas anders.

https://www.tagesschau.de/wirtschaft/apple-287.html (03.01.2019)

Ergebnisse - #apple (2)



Zusammenfassung

Modell	Accuracy
VADER	66,67 %
Neural Network	78,58 %
Convolutional Neural Network	83,05 %

- Sentiment Analyse konnte erfolgreich auf Tweets zu den Unternehmen Apple und Tesla durchgeführt werden
- Zusammenhänge zwischen Twitter Sentiment und unternehmensbezogenen Ereignissen sind identifizierbar

Vielen Dank für die Aufmerksamkeit!

Quellen

Statista: https://de.statista.com/themen/138/facebook/ &
https://de.statista.com/themen/99/twitter/

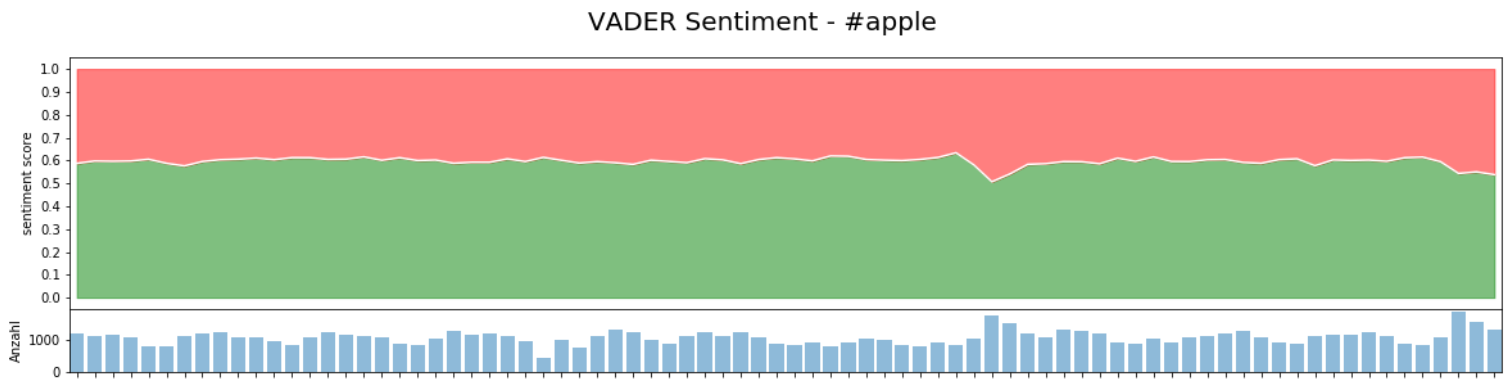
Nachrichten: https://www.tagesschau.de/wirtschaft/apple-287.html &
https://www.onvista.de/aktien/Apple-Aktie-US0378331005 &
https://www.zeit.de/news/2019-01/18/jobabbau-und-weniger-gewinn-bei-tesla-190118-99-621792

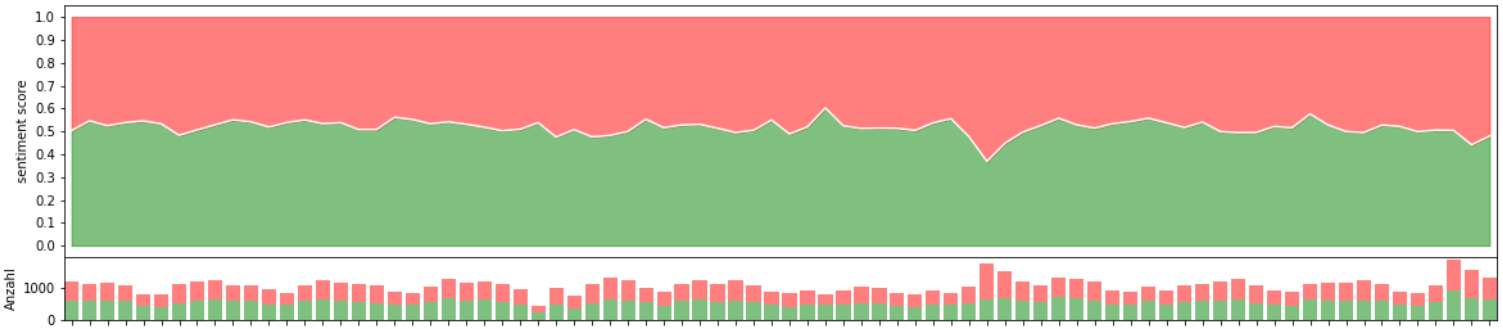
(Hutto & Gilbert 2014) - VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

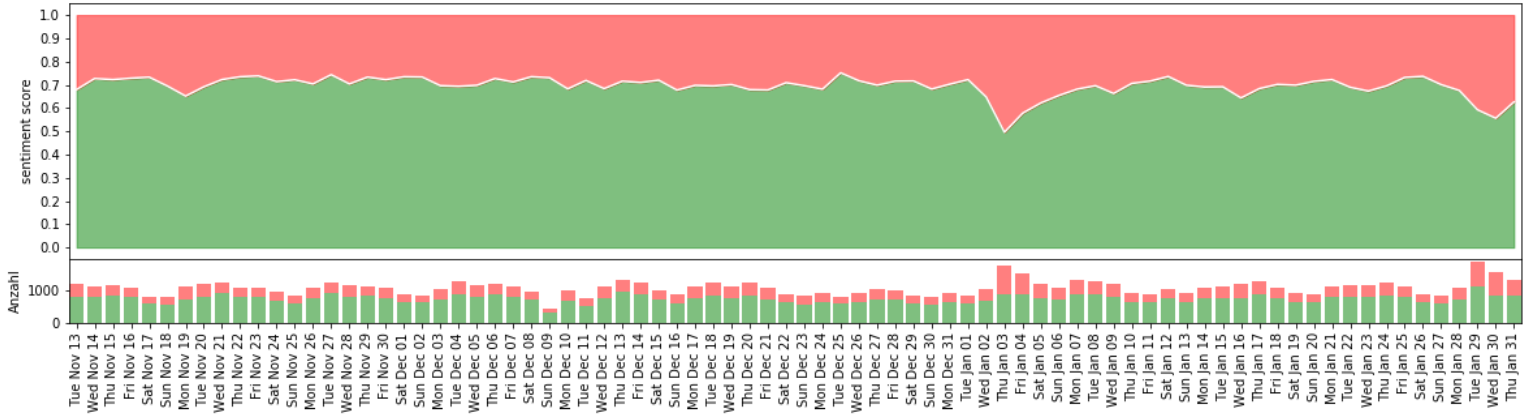
(Le und Mikolov 2014) - Distributed Representations of Sentences and Documents

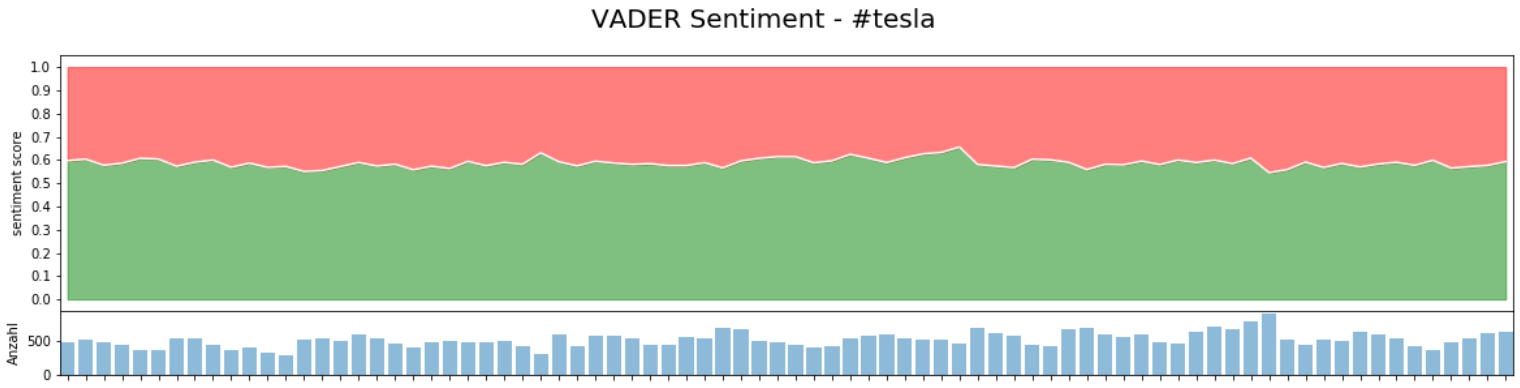
VADER Sentiment - https://github.com/cjhutto/vaderSentiment

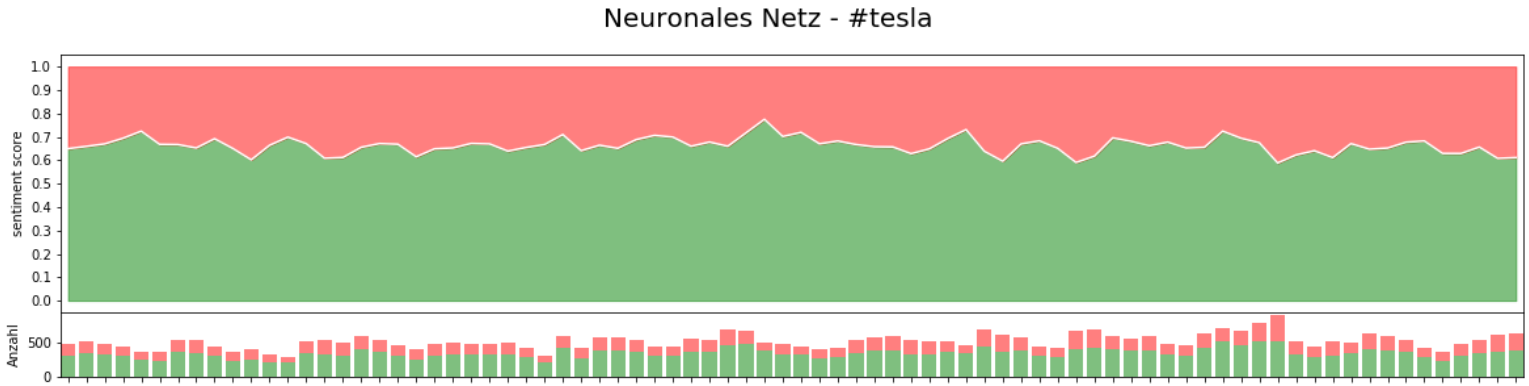
BACKUP FOLIEN

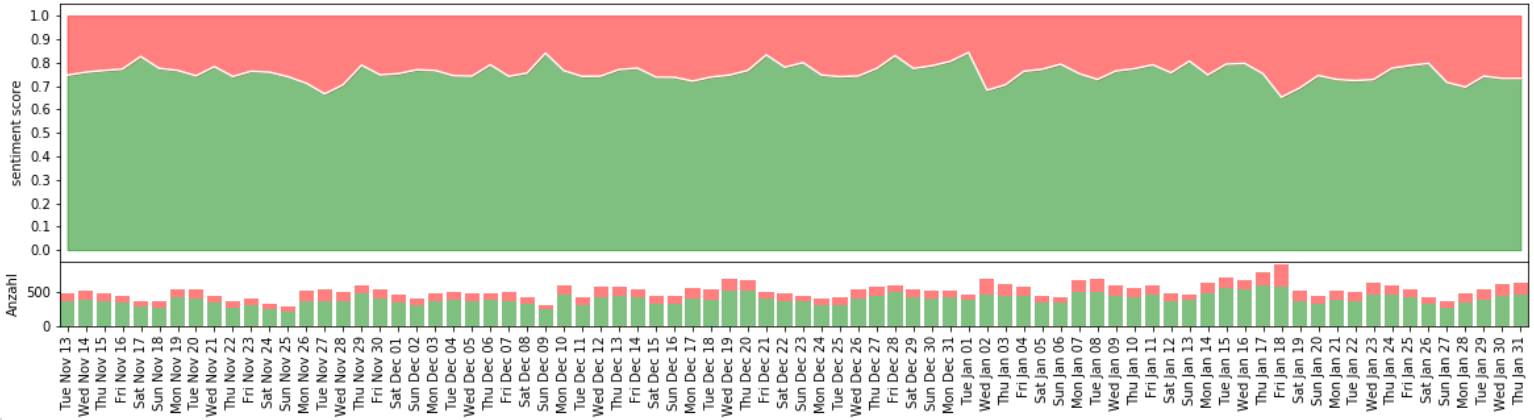

VADER Sentiment - #apple
sentiment score
Anzahl


Neuronales Netz - #apple
sentiment score
Anzahl


Convolutional Neural Network - #apple
sentiment score
Anzahl


VADER Sentiment - #tesla
sentiment score
Anzahl


Neuronales Netz - #tesla
sentiment score
Anzahl


Convolutional Neural Network - #tesla
sentiment score
Anzahl

NN – Callback Function

```
#Definieren der checkpoint und early_stop Funktion
checkpoint = ModelCheckpoint('models/nn/nn_model.h5', monitor='val_acc', verbose=1, save_best_only=True, mode='max')
early_stop = EarlyStopping(monitor='val_acc', patience=5, mode='max')
callbacks_list = [checkpoint, early_stop]
#Initialisierung des finalen Neuronalen Netzes
nn = Sequential()
nn.add(Dense(256, activation='relu', input_dim = 200))
nn.add(Dense(256, activation='relu'))
nn.add(Dense(256, activation='relu'))
nn.add(Dense(1, activation='sigmoid'))
nn.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
#Starte Training von nn mit ausreichend vielen Epochen
nn.fit(train_vecs, y_train, validation_data = (validation_vecs, y_validation), epochs = 100, batch_size = 32, verbose =
2, callbacks = callbacks_list)
```

CNN – Callback Function

```
checkpoint = ModelCheckpoint('models/cnn/cnn_model-e{epoch:02d}.h5', monitor='val_acc', verbose=1, save_best_only=True,
mode='max')
early_stop = EarlyStopping(monitor='val_acc', patience=5, mode='max')
callbacks_list = [checkpoint, early_stop]
cnn_03 = Sequential()
e = Embedding(100000, 200, weights=[embedding_matrix], input_length=58, trainable=True)
cnn_03.add(e)
cnn_03.add(Conv1D(filters=100, kernel_size=2, padding='valid', activation='relu', strides=1))
cnn_03.add(GlobalMaxPooling1D())
cnn_03.add(Dense(256, activation='relu'))
cnn_03.add(Dense(1, activation='sigmoid'))
cnn_03.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
cnn_03.fit(x_train_seq, y_train, validation_data=(x_val_seq, y_validation), epochs=100, batch_size=32, verbose=2, callb
acks = callbacks_list)
```

Aktie unter Druck

Jobabbau und weniger Gewinn bei Tesla

18. Januar 2019, 16:27 Uhr / Quelle: dpa

Palo Alto (dpa) - Der E-Autobauer Tesla hat einen großen Stellenabbau angekündigt und Aktionäre auf weniger Gewinn eingestellt. «Der Weg vor uns ist sehr schwierig», teilte Tesla-Chef Elon Musk am Freitag im