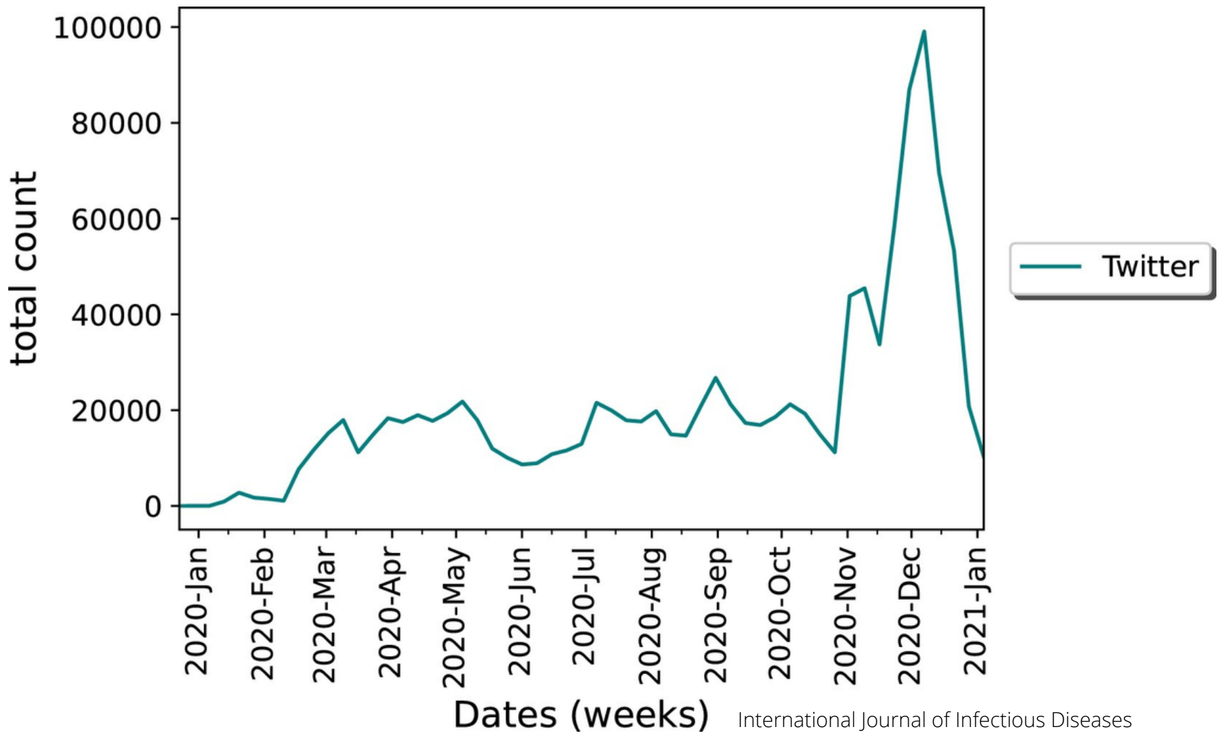
Vaccination sentiment Analysis

On Twitter

INHALT

1. Prerequisites
2. Datenerhebung/Aufbereitung
3. Sentiment Beispiele
4. Vergleich von Auswertungsmodellen
5. Datenauswertung
6. Problematik/Fazit



International Journal of Infectious Diseases
July 2021

VORAUSSETZUNGEN

- Erstellen eines Twitter Dev Accounts

- Erstellen einer Twitter app
- Speichern der Private und public keys für die API Verbindung
- Aufrufen und generieren der Authentication tokens

Hier in einem Ubuntu Linux OS

- Python3 Installation

```
apt install python3
```

- pip Installation

```
apt install python3-pip
```

- Tweepy Installation

```
pip install tweepy
```

- Kafka Installation

```
pip install kafka-python
```

- Python Twitter installation

```
pip install python-twitter
```

DATENERHEBUNG

<u>Aktuelle/Streamed Daten von Twitter API beziehen</u>

⟶ Voraussetzung Twitter Dev account

1. Hier mit setup einer VM/kann auch lokal bezogen werden
2. Installation von apache Kafka auf Linux OS (beinhaltet Zookeeper)
3. Starten des Kafka Servers

```
~$ <'kafka directory'>/bin/kafka-server-start.sh config/server.properties
```

4. Starten einer zweiten shell um ein topic in Kafka zu erstellen

```
~$ <'kafka directory'>/bin/kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic <'name'>
```

5. Anpassen der 'server.properties'-file (+Bootstrap connection)
6. Schreiben des Python scripts

SCRIPT ZUR DATENERHEBUNG

```python
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
from kafka import KafkaProducer
import json

access_token = "904485960-R6V78****************GgjlX"
access_token_secret = "GRr***********cxyM1VgxB"
api_key = "DQi0t***********utb"
api_secret = "pGZosWxW******************K5aO3I081R5"

class StdOutListener(StreamListener):
    def on_data(self, data):
        json_ = json.loads(data)
        producer.send("covid", data.encode('utf-8'))
        return True
    def on_error(self, status):
        print (status)
```

```python
producer = KafkaProducer(bootstrap_servers='192.*****:9092')
l = StdOutListener()
auth = OAuthHandler(api_key, api_secret)
auth.set_access_token(access_token, access_token_secret)
stream = Stream(auth, l)
stream.filter(track=["vaccine", "covid", "booster shot", "Vaccine",
"Covid-19", "third shot", "booster-shot", "Vaccination", "third-shot",
"moderna", "pfizer", "sputnik", "BioNTech", "Johnson&Johnson",
"Johnson & Johnson", "Pfizer-BioNTech", "vaxxed", "unvaccinated"])
```

Code is slightly censored due to privacy concerns

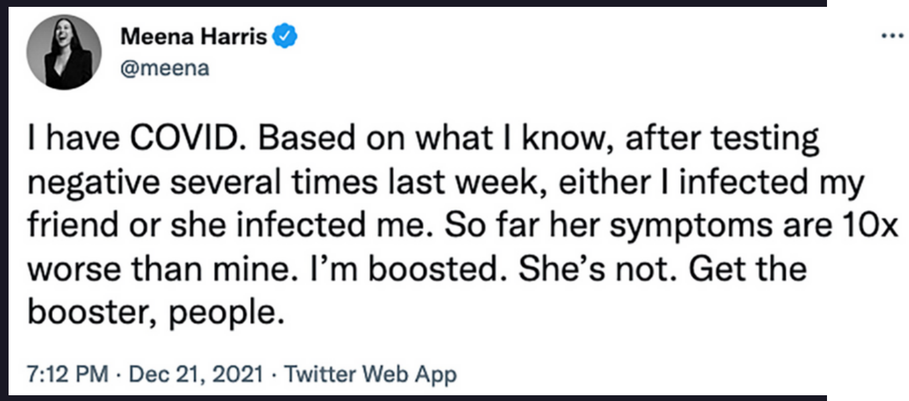
VORLIEGENDE DATEN


```
{
  "created_at": "Wed Dec 01 18:33:22 +0000 2021",
  "id": 1466113207043596292,
  "id_str": "1466113207043596292",
  "text": "RT @DonaldJTrumpJr: Dr. Fauci claims that \"anything and everything\" is on the table to stop the spread of a new COVID variant. American bus\u2026",
  "source": "\u003ca href=\"http:\/\/twitter.com\/#!\/download\/ipad\" rel=\"nofollow\"\u003eTwitter for iPad\u003c\/a\u003e",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 510152045,
    "id_str": "510152045",
    "name": "Texas Deplorable \ud83c\uddfa\ud83c\uddf8\ud83c\uddfa\ud83c\uddf8\ud83c\uddfa\ud83c\uddf8\ud83c\uddfa\ud83c\uddf8",
    "screen_name": "rlfcars",
    "location": "Texas",
    "url": null,
    "description": "Constitutional Conservative & Christian. Proud Texan & Cancer survivor. #DallasCowboys #MAG #2ndAmendment \ud83c\uddfa\ud83c\uddf8\ud83c\uddfa\ud83c\uddf8\ud83c\uddfa\ud83c\uddf8",
    "translator_type": "none",
    "protected": false,
    "verified": false,
    "followers_count": 3722,
    "friends_count": 4145,
    "listed_count": 3,
    "favourites_count": 18209,
    "statuses_count": 11051,
    "created_at": "Thu Mar 01 12:19:15 +0000 2012",
    "utc_offset": null,
    "time_zone": null,
    "geo_enabled": false,
    "lang": null,
```


- JSON format
- lots of Tweet parameters
- file already pre formatted

BEISPIEL TWEEETS (POSITIV)

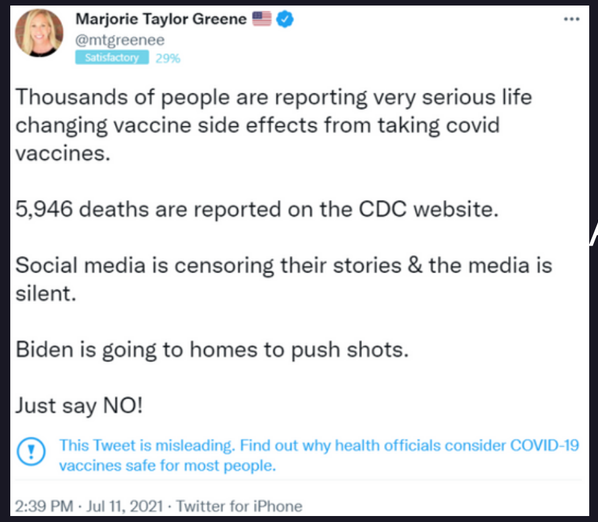
Analysed sentiment: positive



Analysed sentiment: positive

Analysed sentiment: positive

BEISPIEL TWEEETS (NEGATIV)



Analysed sentiment: negativ

Analysed sentiment: negativ

Analysed sentiment: negativ

HUMAN VS. ALGORITHM

Model evaluation: negativ

Human evaluation: rather positiv

Model evaluation: positiv/neutral

Human evaluation: negativ

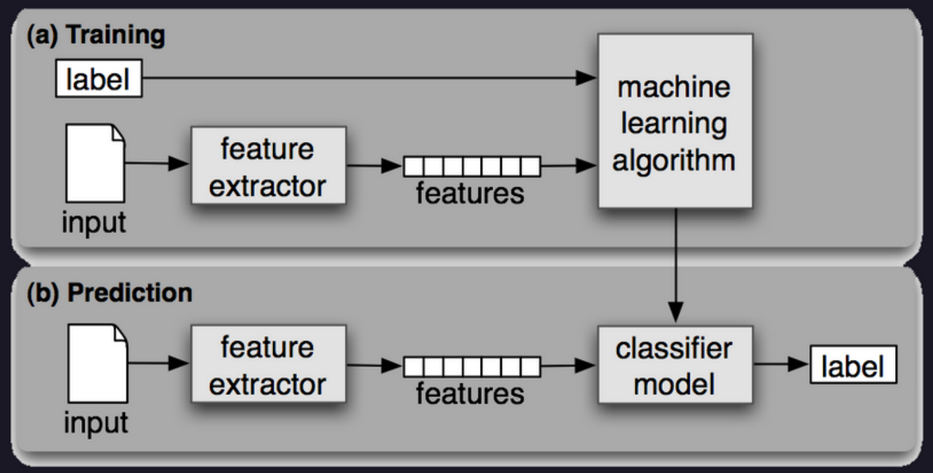
Model evaluation: positiv

Human evaluation: negativ/neutral

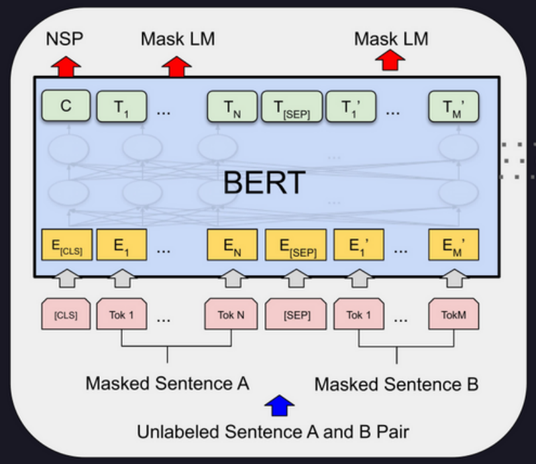
MODELLVERGLEICH

<u>Übereinstimmungen:</u>

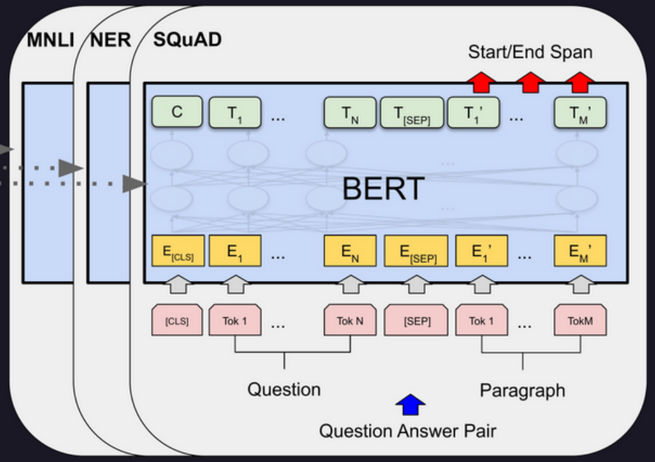
1. Nicht gelabelte Daten
2. Vortrainierte Modelle
3. Sprache der Tweets beeinflussen Analyse nicht
4. Identischer Datensatz
5. Englischer Tweet Corpus



NLTK



Pre-Training



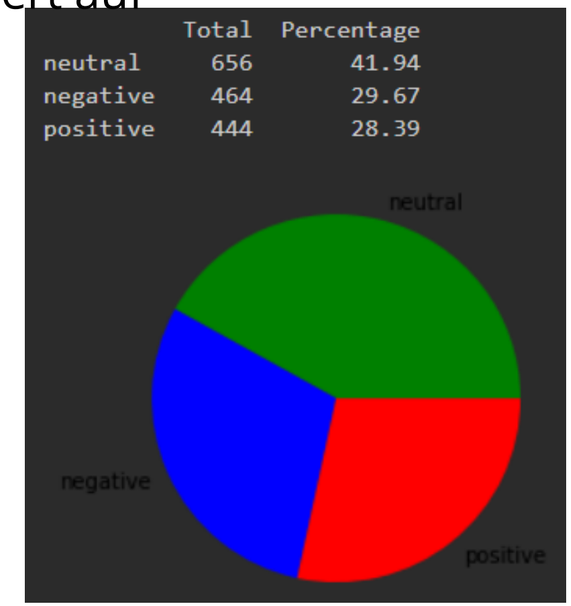
Fine Tuning

NLTK

Bekannte Python Bibliothek

1. Tiefes Tool kann im gesamten NLP Bereich eingesetzt werden
2. Tokenization
3. Speech tagging
4. Naive Bayes Classification
5. Vortrainiert auf immensen Datenmengen verschiedener Gebiete
6. unter anderem spezialisiert auf Sentiment Analysen
7. Vorliegendes Sentiment (positiv/neutral/negativ)

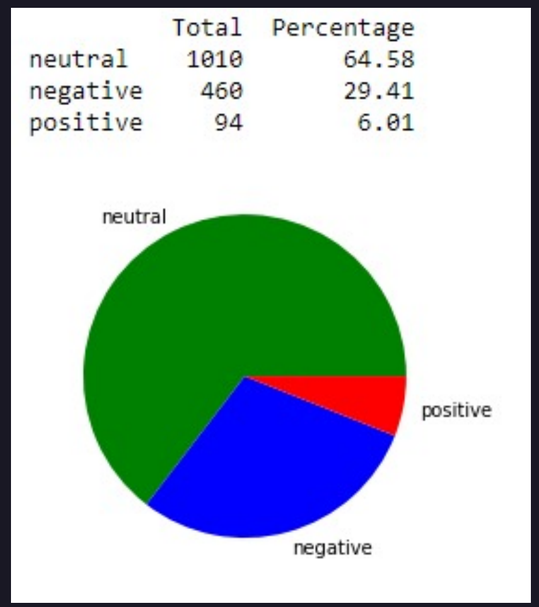
	Total	Percentage
neutral	656	41.94
negative	464	29.67
positive	444	28.39



BERT

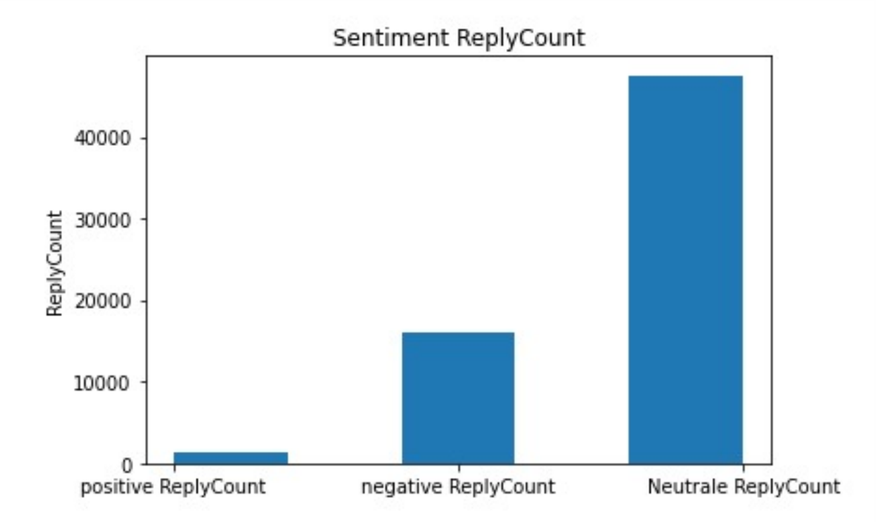
1. Google AI open-sourced Bidirektionales Language Processing
2. Hervorragend geeignet für fine tuning and deep neural networks
3. Pretrained contextualised word embeddings
4. Eher bei semi-supervised learning anzuwenden
5. Bis zu 340 Millionen anpassbare Parameter
6. Nicht Vor trainiert auf aktuellste covid News und Entwicklungen
7. Vorliegendes Sentiment (1/0/-1)

	Total	Percentage
neutral	1010	64.58
negative	460	29.41
positive	94	6.01

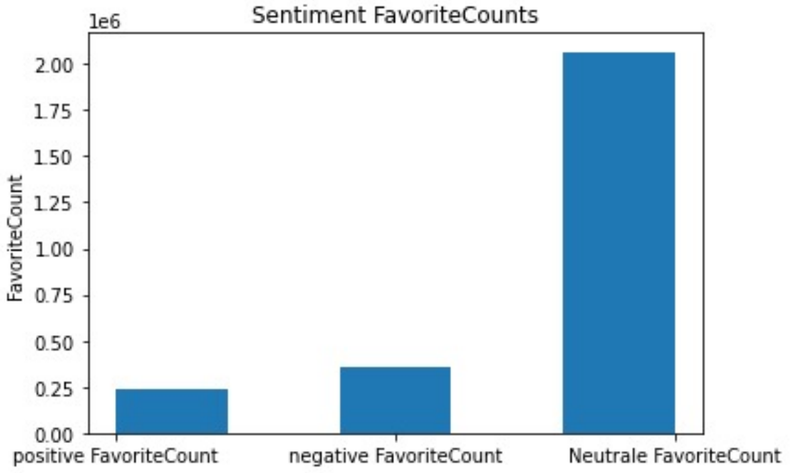


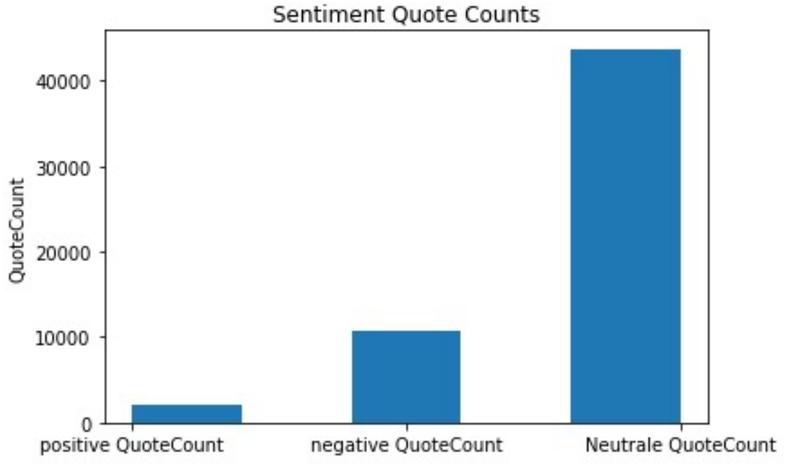
DATENAUSWERTUNG

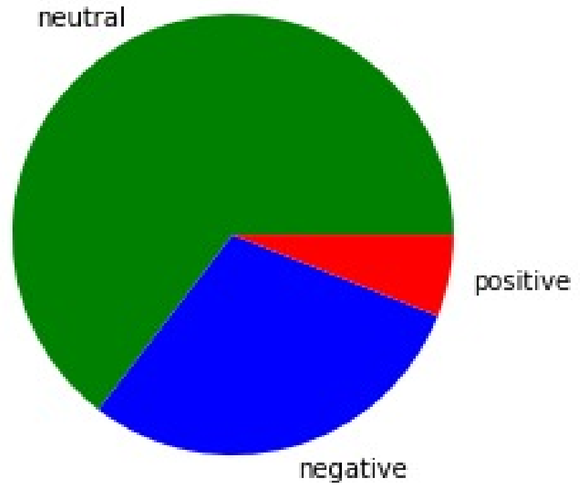
BERT Modell



	Total	Percentage
neutral	1010	64.58
negative	460	29.41
positive	94	6.01

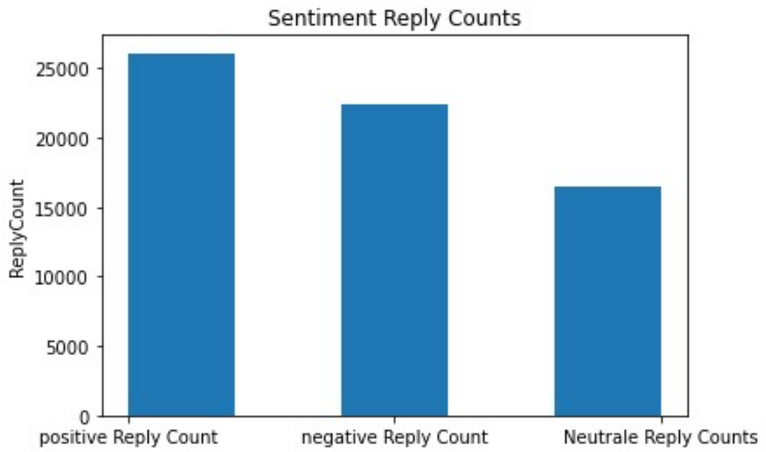


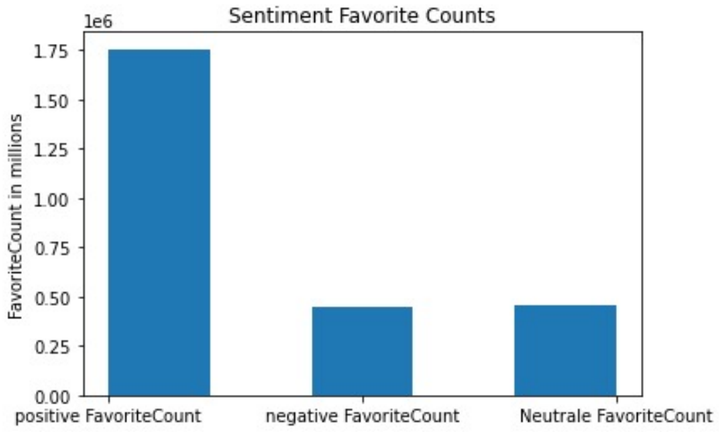


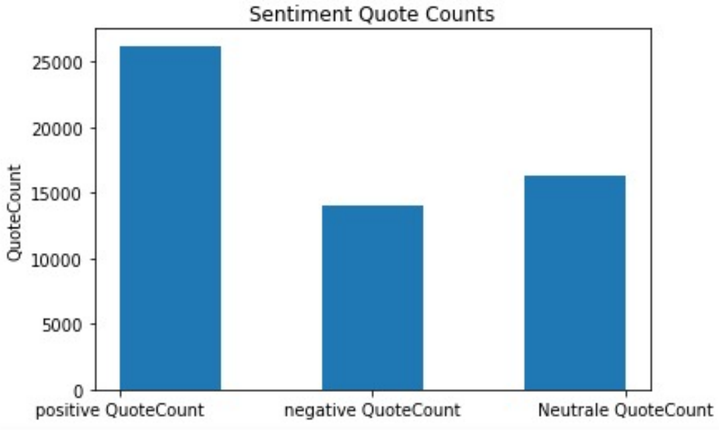


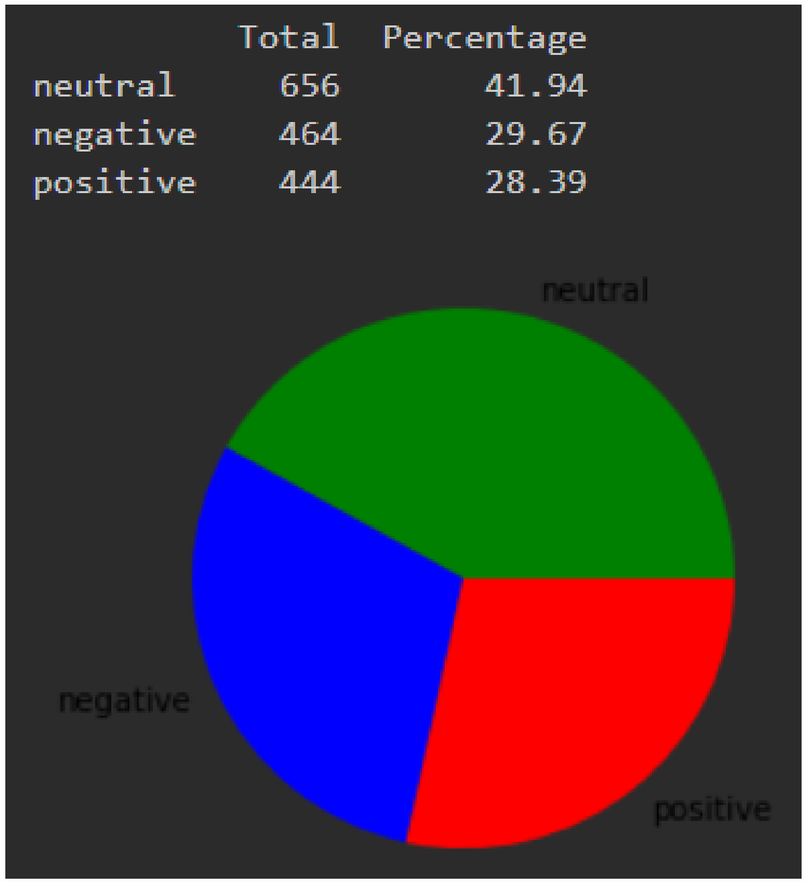
DATENAUSWERTUNG

NLTK Modell









PROBLEME EINES TWITTER SENTIMENTS

- Die Daten sind mehrheitlich von einer Quelle
- Auf Twitter werden hauptsächlich Informationen von Befürwortern veröffentlicht (biased)
- Werbung ist oft schwer zu filtern
- Memes oder Sarkasmus lassen sich von einem Algorithmus nur schwer analysieren
- Ohne das evaluieren durch einen Menschen können schlecht Handlungsempfehlungen geben werden

Vielen lieben Dank für Eure Aufmerksamkeit!