SENTIMENT ANALYSIS

Florian Frey, Frederick Neugebauer, Olena Lavrikova, Anh Vu

GLIEDERUNG

- Sentiment Analysis
- Business Use Case
- Twitter API
- Data Pre-Processing
- Algorithmen
- Ergebnisse
- Fazit & Ausblick

SENTIMENT ANALYSIS

"I am happy with this water bottle."

Positive

"This is a bad investment."

Negative

"I am going to walk today."

Neutral

BUSINESS USE CASE

- Analysieren der Twitter-Stimmung zu einem Videospiel
- Strategieplanung anhand der Analyse
- Verbesserungsmaßnahmen anstoßen
- Verbesserung der Spielerzufriedenheit

➢ Gewinnmaximierung 🤑

UMSETZUNG

- Datensätze zum Trainieren, Testen und Auswerten gewinnen
- Datenaufbereitung
- Algorithmen auswählen
- Modelle trainieren und Hyperparameter tunen
- Ergebnisse analysieren

TWITTER API

- Twitter Developer Account erstellen
- Projekt auf Twitter Developer Portal erstellen
- API Key und ein Bearer Token erhalten
- Verbindung zu den neuen Endpunkten in der Twitter-API v2
- Request zum Erhalten der Tweets erstellen

TWEETS VON DER API

- **Challenge**:
  - Full-Archive Search nur für Academic Access
  - Einige Operators stehen nicht zur Verfügung
  - Mit Elevated Access nur 100 aktuellste Tweets pro Request
- **Lösung**:
  - Statt einem Videogame, verschiedene Videospiele
  - Wiederholte Requests
  - Mergen der Daten

TWEETS VON DER API

- **Qualität**
  - Bots
  - Tweets, die sich nicht direkt auf das Spiel beziehen

source	text
WordPress.com	Grand Theft Auto V Premium Edition – Xbox One https://t.co/zfdhLIsG2A
WordPress.com	Grand Theft Auto V Premium Edition – Xbox One https://t.co/CfZGK119y7
WordPress.com	Grand Theft Auto V Premium Edition – Xbox One https://t.co/KN0muMrN5I
WordPress.com	Grand Theft Auto IV Screensaver Free For Windows [April-2022] ❖ https://t.co/K2vTJT5ord
Twitter Web App	Grand Theft Auto 6: Capital CityWith Trump as a playable character would be fun.
PlayStation®Network	Grand Theft Auto VTo Live or Die in Los Santos (Silver)Completed the final mission. #PS4share https://t.co/LcLeGPUojD

→ Entfernen einiger Sources

TWEETS VON DER API

- **Qualität**
  - „Twitter-Sprache“ oft schwer zu interpretieren
    - “need homies to fuck shit up with on GTA online lmao add me on Xbox! SUBCITYDUBZ”
  - Manche Tweets fälschlicherweise als Englisch markiert

→ Qualität bzgl. Sentiment fraglich

“@MissXu sorry! bed time came here (GMT+1) http://is.gd/fNge”

- **Lowercasing**
  - Alles in Kleinbuchstaben umwandeln
  - *“@missxu sorry! bed time came here (gmt+1) http://is.gd/fnge”*
- **Stopwords**
  - Häufig vorkommende Wörter
  - Wenig relevant für das Sentiment
  - z.B. he, is, …
  - Ausnahme: „not“ und „no“

- Entfernen von **Links, Tags** und **Sonderzeichen**
  - Mit Hilfe von Regular Expression
  - "http[s]?://\S+"
  - *“sorry bed time came here gmt”*
- **Word Stemming**
  - Wörter in ihren Wortstamm zurückführen

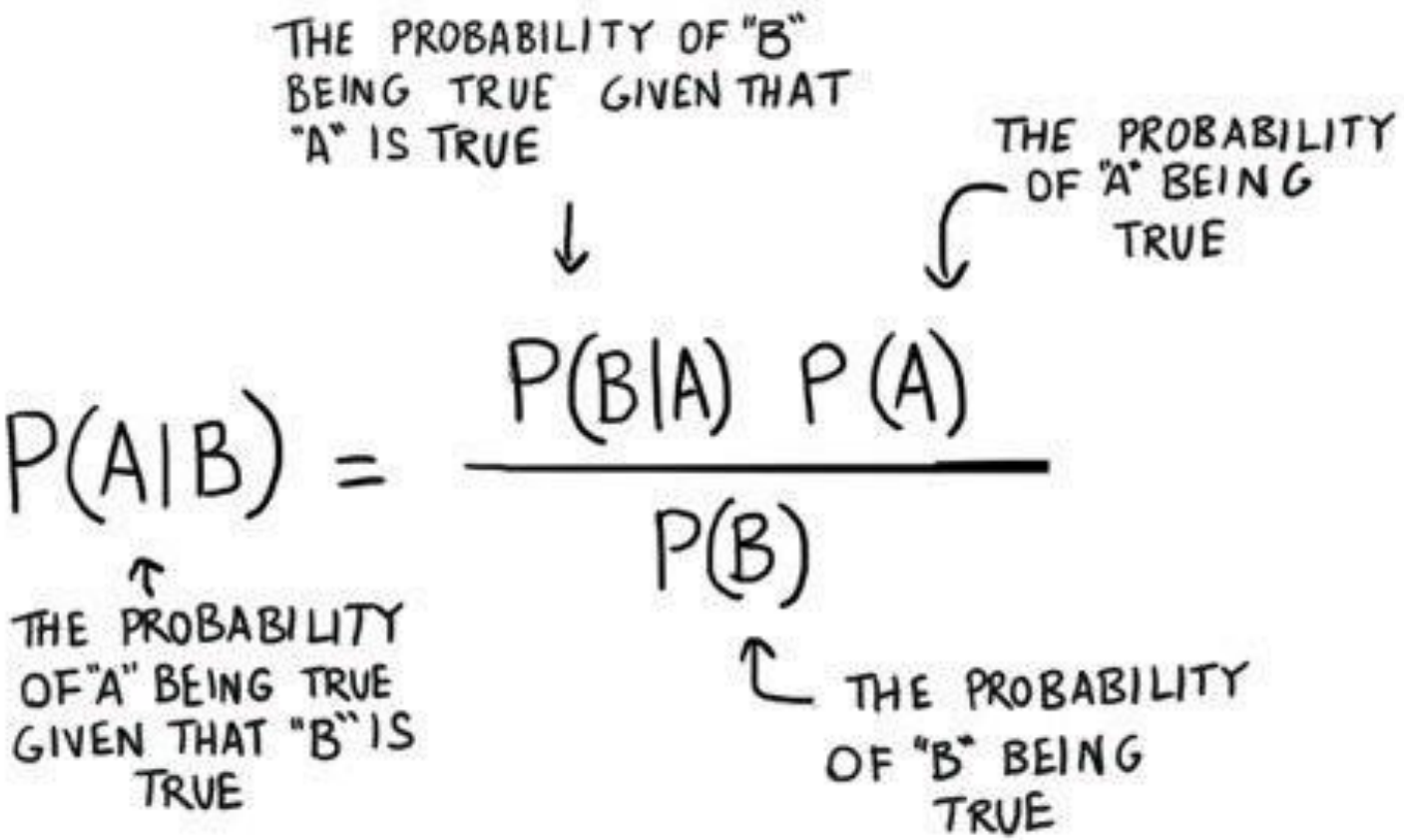
  →Vereinheitlichung von Konjugationen
- **Tokenizing**
  - Satz in Liste von Wörtern umwandeln
  - *[‘sorry’, ‘bed’, ‘time’, ‘came’, ‘here’, ‘gmt’]*

ALGORITHMEN

- TF-IDF
- Naive Bayes
- Support Vector Machine
- Implementierung in Python mit sklearn

TF-IDF

- TF steht für Term-Frequency, also der Häufigkeit der Wörter in einem Dokument
- IDF steht für Inverse-Document-Frequency, das bedeutet Wörter die in viele Dokumenten vorkommen wird ein geringeres Gewicht zugeordnet
- Wandelt Wörter anhand ihrer Häufigkeit und anhand der Häufigkeit der Dokumente in denen das Wort vorkommt in einen Vektor um



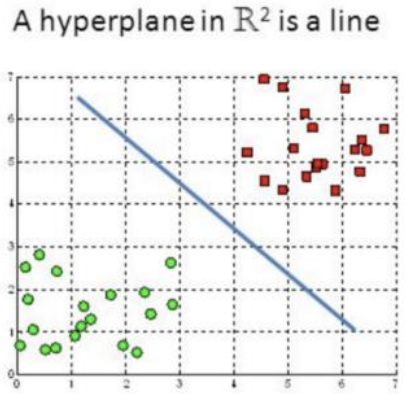
https://blog.clairvoyantsoft.com/mlmuse-naivety-in-naive-bayes-classifiers-9c7f6ba952bf

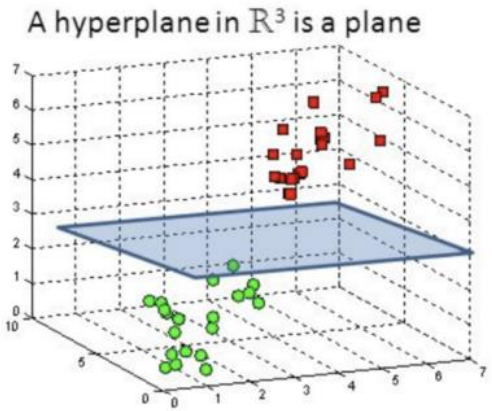
NAIVE BAYES

- Einfacher probabilistischer Klassifikator der auf Bayes' Theorem basiert
- Betrachtet Features unabhängig voneinander
- Gibt für jede Klasse eine Wahrscheinlichkeit, dass die Beobachtung zu dieser Klasse gehört

SUPPORT VECTOR MACHINE

- Trennt Daten mithilfe einer Hyperplane (Gerade bzw. Ebene) in Klassen
- Die Margin ist die Entfernung zu den nächstgelegenen Datenpunkten der verschiedenen Klassen
- Optimal ist eine möglichst große Margin zwischen Hyperplane und den Datenpunkten der Klassen



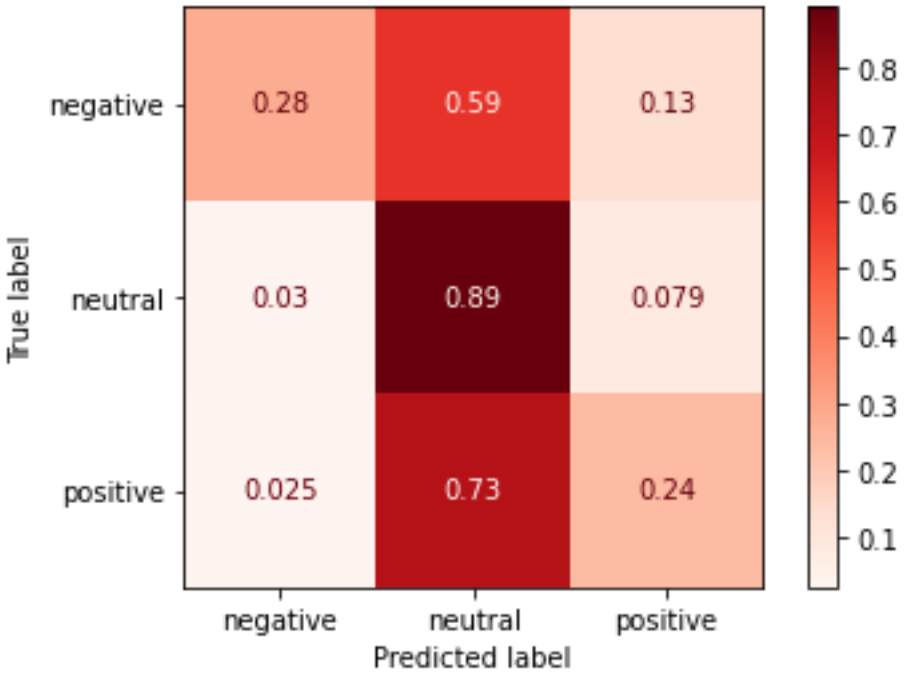


https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c

ERGEBNISSE

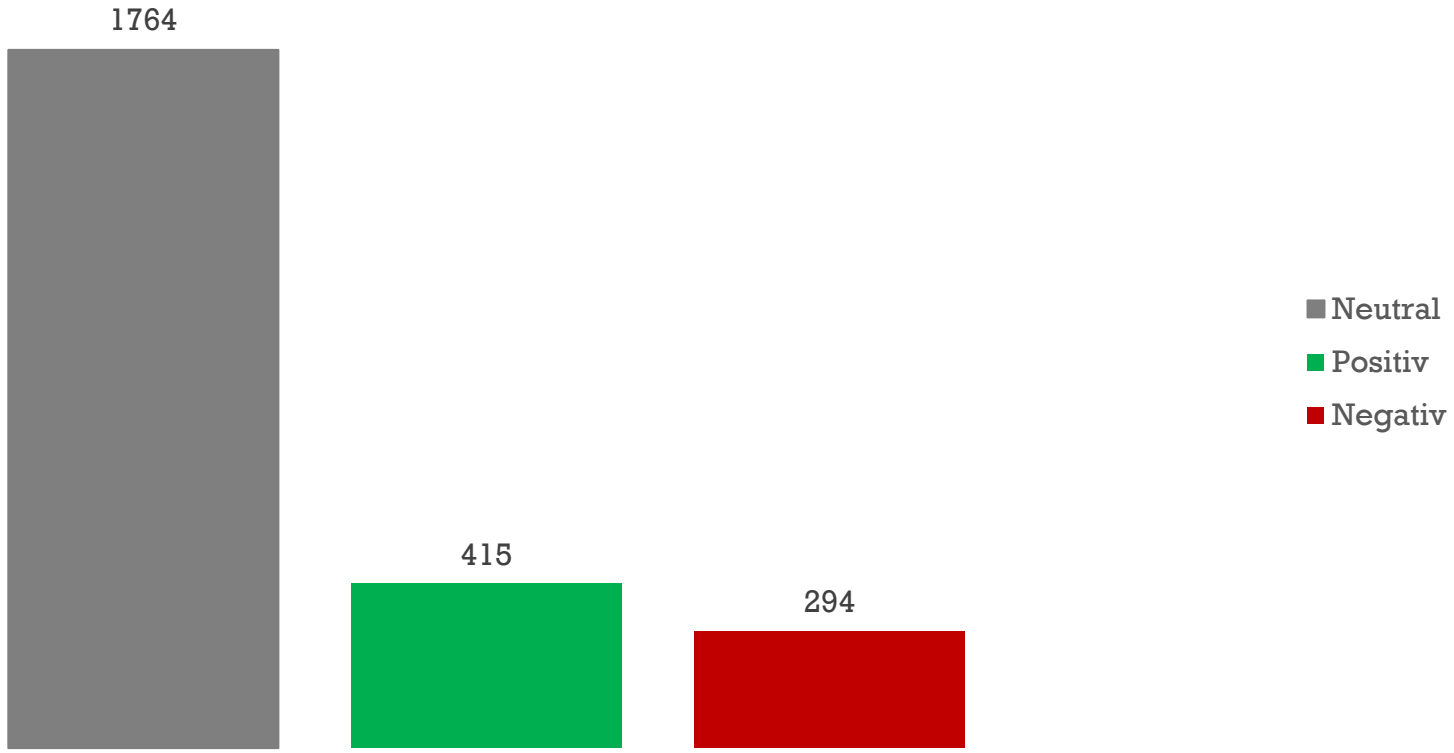
- Testen auf 600 gelabelten Game-Tweets
- Trainingsdatensatz: 27k Tweets
  - neutral 11118
  - positive 8582
  - negative 7781
- Tweets möglicherweise unterschiedlich zu den Game-Tweets

➢ Bias auf neutrale predictions

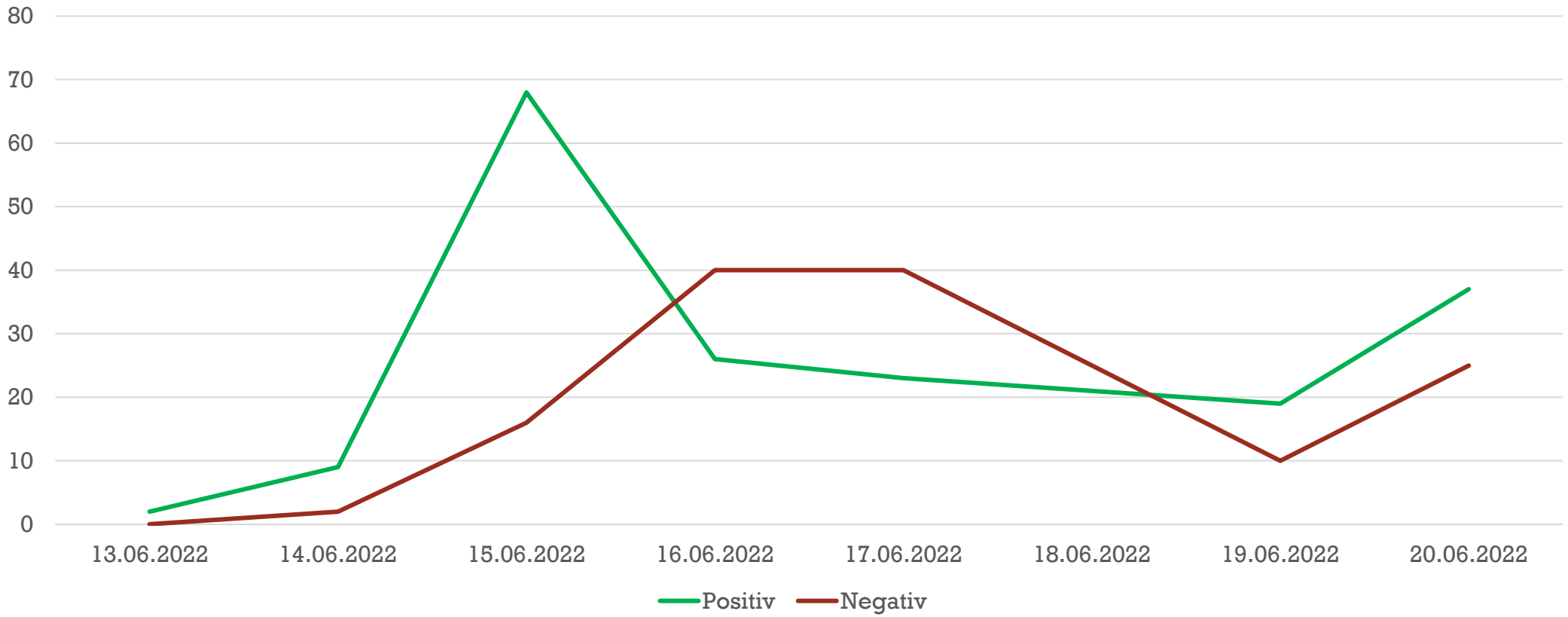


ERGEBNISSE

- Anwendung des besten Models auf unsere Game-Tweets



ERGEBNISSE


80
70
60
50
40
30
20
10
0
13.06.2022
14.06.2022
15.06.2022
16.06.2022
17.06.2022
18.06.2022
19.06.2022
20.06.2022
Positiv
Negativ

FAZIT

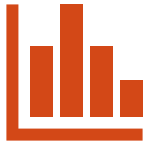
- Daten schwierig manuell zu labeln
- Twitter API benutzerunfreundlich
- Herausforderung bei der Abfrage von Tweets
- Viele Tweets wurden als neutral bewertet
- Videospiele geeignetes Thema?

OUTLOOK

Höheren Twitter API Access erlangen

Automatisierung der API Abfragen

Tweet-Filterung und Data-Preprocessing ausbauen



Grafische Aufbereitung

QUELLEN

- sammit (2020): Naive Bayes' Classifiers | Supervised learning algorithms | Clairvoyant Blog. In: Clairvoyant Blog, 18.02.2020. Online verfügbar unter https://blog.clairvoyantsoft.com/mlmuse-naivety-in-naive-bayes-classifiers-9c7f6ba952bf (abgerufen am 29.06.2022)
- Pier Paolo Ippolito (2019): SVM: Feature Selection and Kernels. In: towardsdatascience. Online verfügbar unter https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c (abgerufen am 29.06.2022)

THE END ☺