

# 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."*



*"This is a bad investment."*



*"I am going to walk today."*







# 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 <a href="https://t.co/zfdhLlsG2A">https://t.co/zfdhLlsG2A</a>
WordPress.com	Grand Theft Auto V Premium Edition – Xbox One <a href="https://t.co/CfZGK119y7">https://t.co/CfZGK119y7</a>
WordPress.com	Grand Theft Auto V Premium Edition – Xbox One <a href="https://t.co/KN0muMrN5l">https://t.co/KN0muMrN5l</a>
WordPress.com	Grand Theft Auto IV Screensaver Free For Windows [April-2022] 💎 <a href="https://t.co/K2vTJT5ord">https://t.co/K2vTJT5ord</a>
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 <a href="https://t.co/LcLeGPUojD">https://t.co/LcLeGPUojD</a>

→ 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





## DATA PRE-PROCESSING

*“@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“



## DATA PRE-PROCESSING

- 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





## 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

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE

THE PROBABILITY OF "A" BEING TRUE

THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE

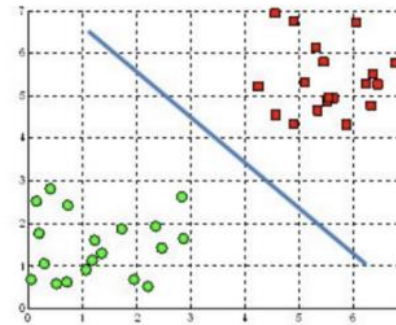
THE PROBABILITY OF "B" BEING TRUE



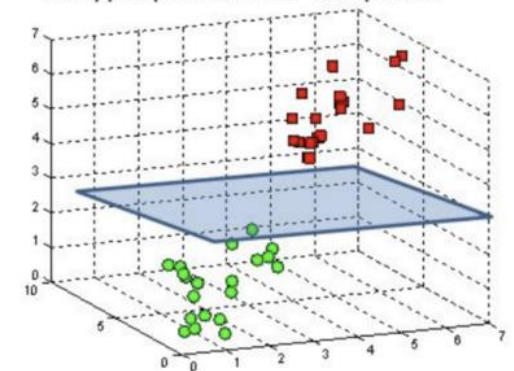
# SUPPORT VECTOR MACHINE

- 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

A hyperplane in  $\mathbb{R}^2$  is a line



A hyperplane in  $\mathbb{R}^3$  is a plane



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

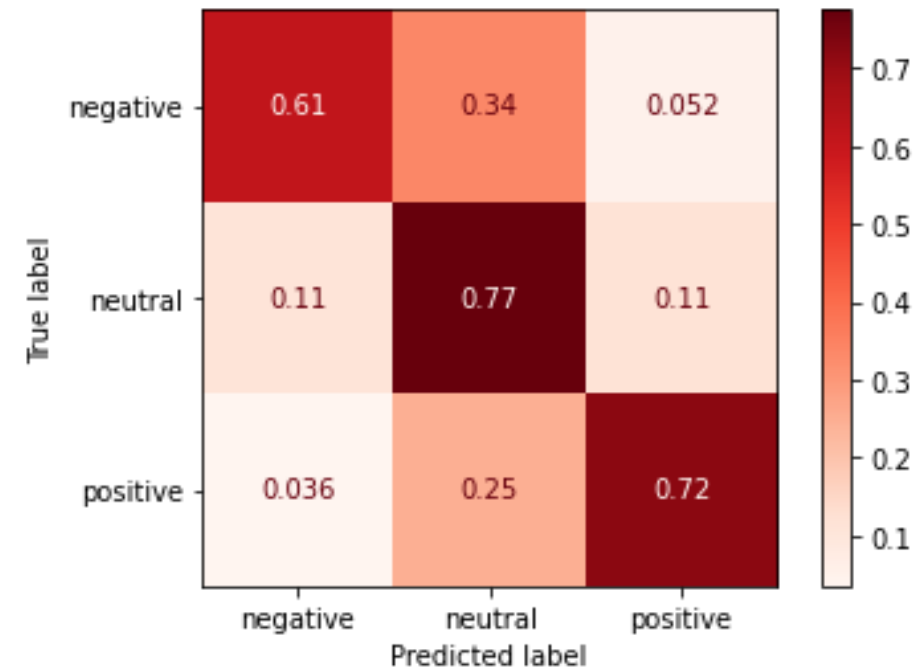


# ERGEBNISSE

- Training und Test auf Datensätzen von Kaggle

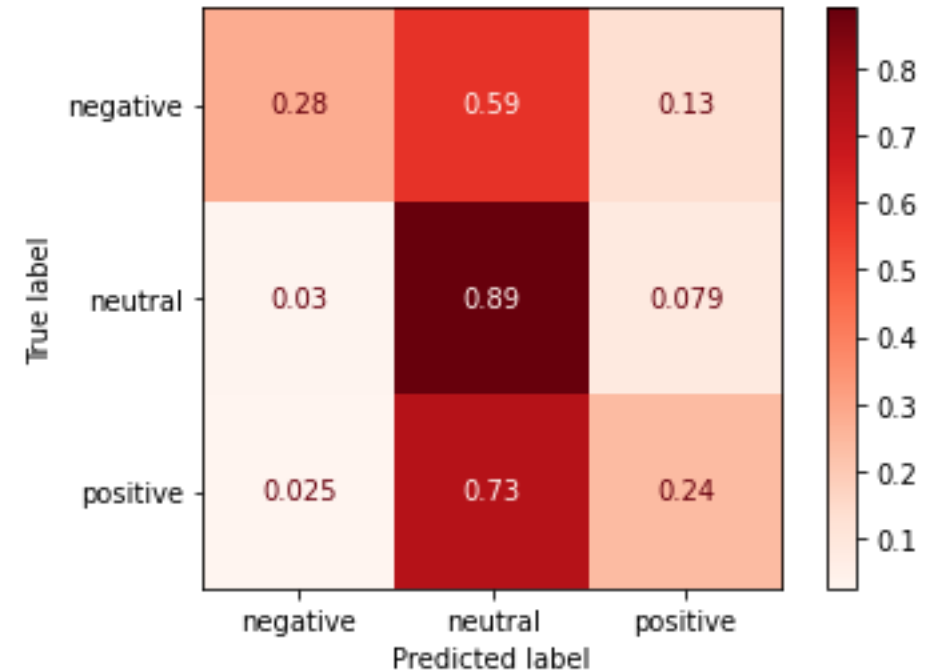
Accuracy Score	Binary	Multiclass
Naive Bayes	76,30%	65,02%
Support Vector Machine	53,66%	71,37%

- Confusion Matrix SVM Multiclass:



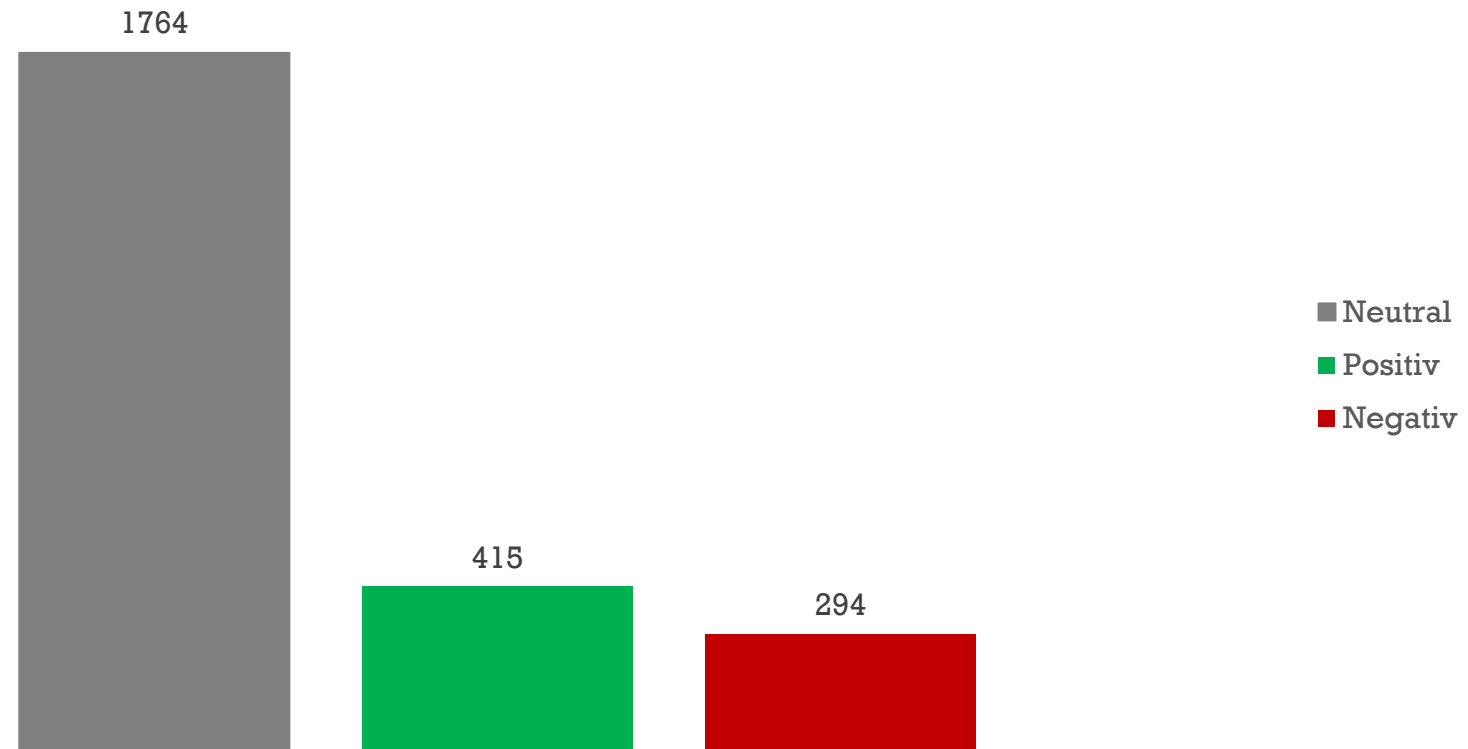
# ERGEBNISSE

- 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



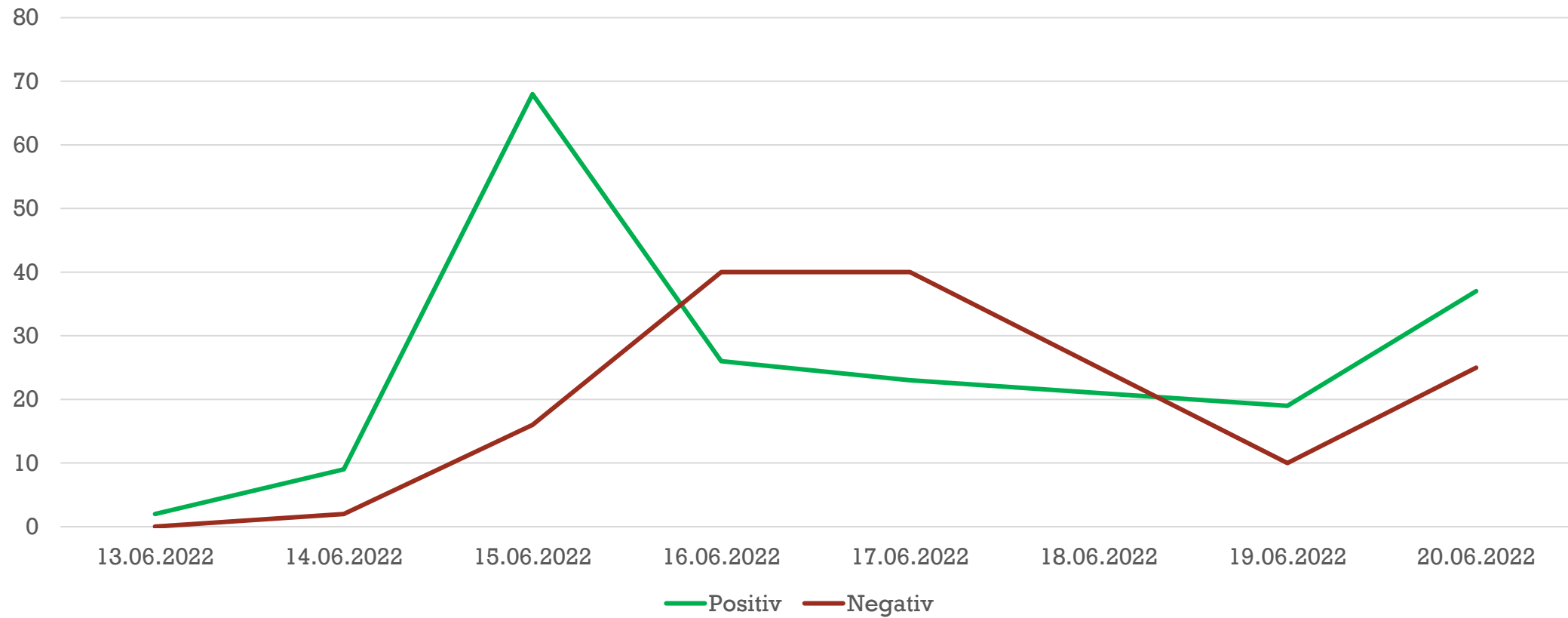
# ERGEBNISSE

- Anwendung des besten Models auf unsere Game-Tweets





# ERGEBNISSE



# 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

---



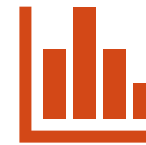
Höheren Twitter  
API Access  
erlangen



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** 😊

---