Studienarbeit – Klassifikation von Bürgeranfragen

# 1. Einleitung

* [Platzhalter] Motivation und Relevanz
* [Platzhalter] Zielsetzung der Arbeit
* [Platzhalter] Aufbau des Dokuments

# 2. Theoretischer Hintergrund

* [Platzhalter] Grundlagen von NLP
* [Platzhalter] Textklassifikation im Überblick
* [Platzhalter] Klassische ML-Ansätze (Naive Bayes, TF-IDF etc.)

# 3. Methodisches Vorgehen (CRISP-DM)

* [Platzhalter] 3.1 Business Understanding
* [Platzhalter] 3.2 Data Understanding
* [Platzhalter] 3.3 Data Preparation
* [Platzhalter] 3.4 Modelling
* [Platzhalter] 3.5 Evaluationskonzept

# 4. Umsetzung

* [Platzhalter] 4.1 Datenquelle und Struktur
* [Platzhalter] 4.2 Textverarbeitung (Tokenisierung, Cleaning, TF-IDF)

## 4.1 Datenquelle und Struktur

## 4.2 Textverarbeitung (Tokenisierung, Cleaning, TF-IDF)

## 4.3 Erstes Modelltraining (Naive Bayes)

Zur Kategorisierung der Bürgeranfragen wurde ein Multinomial-Naive-Bayes-Modell trainiert. Die Vektorisierung erfolgte mit TF-IDF, die Trainings- und Testdaten wurden im Verhältnis 80 % zu 20 % aufgeteilt. Bei insgesamt 60 Beispielen umfasste der Testdatensatz nur 12 Anfragen.  
  
Das Modell erreichte eine Genauigkeit von 8,3 % - das liegt deutlich unter dem Erwartungswert eines zufälligen Klassifikators. Der Klassifikationsbericht (siehe Abbildung 1) zeigt, dass nur die Kategorie „Abfall“ teilweise erkannt wurde, während andere Klassen vollständig verfehlt wurden.  
  
Die geringe Anzahl an Trainingsdaten sowie die ungleichmäßige Verteilung der Klassen im Testset sind mögliche Ursachen. In einer erweiterten Version des Projekts könnte eine gezielte Balancierung der Klassen sowie die Verwendung eines komplexeren Modells (z. B. Logistic Regression oder SVM) sinnvoll sein.

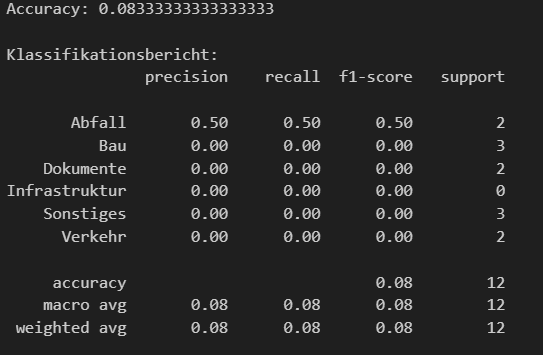
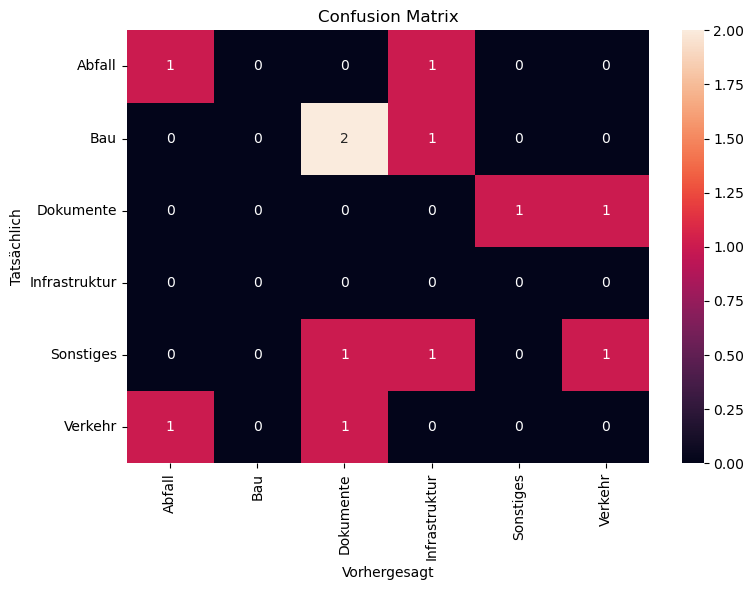


Abbildung 2: Confusion Matrix

Abbildung : Klassifikationsbericht

## 4.4 Modellverbesserung und Vergleich

### 4.4.1 Erweiterung der Textverarbeitung

zusätzliche Vorverarbeitung beschreiben (z. B. Stoppwörter, spaCy, Balancing)

### 4.4.2 Modellwechsel auf Logistic Regression

### 4.4.3 Vergleich der Ergebnisse

einen Vergleich machen (z. B. neue Confusion Matrix)

# 5. Fazit und Ausblick

* [Platzhalter] Zusammenfassung der Ergebnisse
* [Platzhalter] Bewertung & Grenzen
* [Platzhalter] Ausblick (z. B. LLM oder spaCy)

# 6. Literaturverzeichnis

[Platzhalter] Quellenangaben