

Projektpräsentation – Optimierung der Mundart-Chat-Analyse

24.10.2025 | Raphael Weiss

1. Ausgangslage – Projektstand bei Zwischenpräsentation

Idee

Automatisierte **Sentiment-Analyse** für Schweizerdeutsche Chatnachrichten.

Verwendete Modelle

- **BoW + Logistic Regression**
 - mit Dialekt-Preprocessing
- **TF-IDF + Logistic Regression**
 - Uni- und Bigramme, Preprocessing
- **SBERT + Logistic Regression**
 - *paraphrase-multilingual-MiniLM-L12-v2*

Herausforderungen

- Chatnachrichten sind **extrem kurz** (oft 2–5 Wörter)
- **Kein Kontext** → Sentiment sehr schwer zu bestimmen
- **Hohe Accuracy**, aber künstlich verzerrt
→ Modelle erkennen triviale Muster,
schlechte Generalisierbarkeit (Overfitting)
- Fehlende Vielfalt im Training (wenig Variationen, viele Hapax-Wörter)

2. Erweiterung & Optimierung des Projekts

2.1 Datenerweiterung und Analyse

- **Generierung zusätzlicher Mundartnachrichten mit mehr Kontext**
→ komplexere Sätze → bessere Modellrobustheit
- **Kein Einsatz von Emojis**
Begründung: Emojis tragen starke Sentimentsignale → Modelle lernen „Emoji = Sentiment“ statt semantische Inhalte → schlechte Generalisierung.
- **Mehr Daten- und Modell-Analysen:** z.B. Confusion Matrix und Zipf

2.2 Einführung einer mehrstufigen Klassifikation

Sentiment mit Unterklassen (Intents)	Vorteile
<ul style="list-style-type: none">• Positiv → <i>danke, lob, freud</i>• Negativ → <i>beschwerde, frust, hilfe_bitten</i>• Neutral → <i>smalltalk, orga, frage_info, sonstiges</i>	<ul style="list-style-type: none">• Erhöhte semantische Differenzierung• bessere Auswertbarkeit und feinere Chat-antwortoptionen.

2.3 Neue Modelle / Methoden

a) Predict-Next-Word (N-Gramm-Language-Model)

- Unterstützung beim Generieren von Text
- Nutzung eines 1-, 2-, 3-Gramm-Modells mit Backoff-Strategie
- „welches Wort ist wahrscheinlich als nächstes?“

b) Predict Answer (SBERT-Embedding Retrieval)

- semantisch ähnliche Nutzeranfragen finden
- passende Antwort aus vordefiniertem Antwortkorpus auswählen

Debug-Modus

- erklärt welche Nachbarn ähnlich waren
- und welche Ähnlichkeitswerte vergeben wurden → wichtig für Interpretierbarkeit

Warum Retrieval und kein generatives Modell?

Zu kleine Datenmenge: für generative Modelle viel grössere Daten nötig

Funktionsfähigkeit: schnell und stabil bei klar definierte Intents und Standardantworten

Erklärbarkeit: Mit SBERT-Nachbarn sieht man transparent, warum eine Antwort gewählt wurde.

2.4 Integration in Streamlit-App (Frontend)

<https://mundartchat.streamlit.app>

3. Pipeline – Gesamtprozess

1 Datengrundlage <ul style="list-style-type: none">• 600 Sätze (manuell + generiert)• Chatpairs für Antwortretrieval	2 Preprocessing <ul style="list-style-type: none">• Dialektstandardisierung• Lowercasing• Tokenisierung auf Wortebene• Entfernung von Stopzeichen / Noise-Token
3 Feature-Erstellung & Modelltraining <p>BoW (Bag-of-Words): CountVectorizer und Klassifikation via multinomiale Logistic Regression</p> <p>TF-IDF: TfidfVectorizer und Klassifikation via multinomiale Logistic Regression</p> <p>SBERT: SentenceTransformer und Logistic Regression auf Embedding-Raum</p>	
3 Neu: N-Gram Language Model <ul style="list-style-type: none">• Training von 1-, 2-, 3-Grammen• Backoff für Next-Word-Vorhersagen• Nutzung im UI zur Textgenerierung	3 Neu: Antwort-Retrieval (SBERT-basiert) <ul style="list-style-type: none">• Suche nach semantisch ähnlicher Chatnachricht und Auswahl der passenden Antwort• Debug-Ansicht für „Nearest Neighbors“

4

Evaluations & Visualisierungen in der App

- Accuracy, Precision, Recall, F1
- Confusion Matrices (Heatmaps) für SBERT
- Label-Verteilung
- Textlängenstatistik
- Token-Statistik (Hapax, Typen-Token)
- Zipf-Analyse
- N-Gramm-Statistik

4. Fazit – Optimierte Lösung und Verbesserungspotential

Erreichte Optimierungen

- bessere Generalisierbarkeit
- robustere Modelle für kurze Dialekttexte
- semantische Tiefe dank Intent-Klassen
- Transparenz durch Debugging & Statistiken
- Demo-App für Sentimentanalyse, Next-Word, Antwortgenerierung

Verbesserungspotential

- grössere & vielfältigere Datengrundlage
(mehr Dialekte, echtes Chatmaterial,)
- mehr Kontextverarbeitung (Dialogmodelle)
- Erweiterung der Intents und Multi-Label-Klassifikation
- Kombination aus Retrieval und leichter Generierung (RAG)