



Projektbericht

Sentimentanalyse von Produktrezensionen

Studiengang
"Angewandte Künstliche Intelligenz"

Natural Language Processing
DLBAIPNLP01_D

Dawid Jedlinski

Matrikelnummer: IU14113900

dawid.jedlinski@iu-study.org

Tutor: Maja Popovic

Abgabedatum: dd.mm.yyyy

Inhaltsverzeichnis

I. Abbildungsverzeichnis	III
II. Tabellenverzeichnis	IV
III. Abbreviations	V
1. Einleitung	1
1.1. Problemstellung	1
1.2. Zielsetzung	1
1.3. Vorgehensweise	1
2. Datenbeschreibung	2
2.1. Quelle	2
2.2. Ausgewählte Produktkategorien	2
2.3. Datenumfang und Aufteilung	2
2.4. Verteilung der Bewertungen	2
3. Datenvorverarbeitung	3
3.1. Trennung von Eingabe- und Ausgabedaten	3
3.2. Textaufbereitung	3
3.3. Numerische Repräsentation	3
4. Modell und Methodik	4
4.1. Lineare Regression	4
4.2. Naive Bayes	4
4.3. Trainingsprozess	4
5. Evaluation und Ergebnisse	5
5.1. Evaluationsmetriken	5
5.2. Gesamtergebnisse	5
5.3. Ergebnisse nach Produktkategorien (optional)	5
5.4. Ergebnisse von zwei Algorithmen (optional)	5
6. Schluss	6

I. Abbildungsverzeichnis

II. Tabellenverzeichnis

III. Abbreviations

EDA	Explorative Datenanalyse
GMM	Gaussian Mixture Model
HR	Human Resources
LLE	Locally Linear Embedding
MDS	Multidimensional Scaling
MH	Mental Health (Mentale Gesundheit)
ML	Machine Learning
NLP	Natural Language Processing
OSMI	Open Sourcing Mental Illness
PCA	Principal Component Analysis

1. Einleitung

- irgendwas... _____

1.1. Problemstellung

- Bedeutung von Online-Produktrezensionen - Herausforderung der manuellen Auswertung - Relevanz automatischer Sentimentanalyse

1.2. Zielsetzung

- Entwicklung eines Textklassifikationssystems - Vorhersage numerischer Bewertungen (1–5 Sterne) - Fokus auf Mehrklassen-Klassifikation

1.3. Vorgehensweise

- Überblick über Daten, Methodik und Evaluation - Kurze Beschreibung der einzelnen Kapitel

2. Datenbeschreibung

- irgendwas ... ——————

2.1. Quelle

- Amazon Reviews'23 Datensatz (McAuley Lab) - Allgemeine Beschreibung des Datensatzes - Wahl des 0-core-Splits (oder 5-core?) und Begründung

2.2. Ausgewählte Produktkategorien

- Beschreibung der verwendeten Kategorien (mind. drei) - Begründung der Auswahl

2.3. Datenumfang und Aufteilung

- Anzahl der Rezensionen pro Kategorie - Aufteilung in Trainings- und Testsatz

2.4. Verteilung der Bewertungen

- Häufigkeit der Ratings (1–5 Sterne) - Vergleich von Trainings- und Testsatz

3. Datenvorverarbeitung

- irgendwas über Preprocessing ——————

3.1. Trennung von Eingabe- und Ausgabedaten

- Definition der Eingabedaten (Titel + Rezensionstext) - Definition der Zielvariable (numerische Bewertung)

3.2. Textaufbereitung

- Zusammenführung von Titel und Text - Tokenisierung - Entfernen von Stopwörtern - Umgang mit Sonderzeichen und Groß-/Kleinschreibung

3.3. Numerische Repräsentation

- Verwendung von TF-IDF - Wahl der n-Gramm-Länge (z. B. Unigramme) - Begründung der Feature-Darstellung

4. Modell und Methodik

- Problemformulierung als Klassifikationsaufgabe - Supervised Learning - Mehrklassen-Klassifikation (5 Klassen)
-

4.1. Lineare Regression

- [1] discriminative - [1] kategorien selbst bestimmen - Beschreibung des gewählten Klassifikators - Begründung der Eignung für Textklassifikation

4.2. Naive Bayes

- [1] generative - [1] rechnet mit wahrscheinlichkeiten - Beschreibung des gewählten Klassifikators - Begründung der Eignung für Textklassifikation

VERGLEICH - Both methods are simple; Naive Bayes is the simplest one. - Both methods are interpretable: you can look at the features which influenced the predictions most (in Naive Bayes - usually words, in logistic regression - whatever you defined). - Naive Bayes is very fast to train - it requires only one pass through the training data to evaluate the counts. For logistic regression, this is not the case: you have to go over the data many times until the gradient ascent converges. - Naive Bayes is too "naive" it assumed that features (words) are conditionally independent given class. Logistic regression does not make this assumption - we can hope it is better. - Both methods use manually defined feature representation (in Naive Bayes, BOW is the standard choice, but you still choose this yourself). While manually defined features are good for interpretability, they may be no so good for performance - you are likely to miss something which can be useful for the task.

4.3. Trainingsprozess

- Training auf dem Trainingsdatensatz - Keine separate Validierungsmenge - Begründung des Vorgehens

5. Evaluation und Ergebnisse

- irgendwas... ——————

5.1. Evaluationsmetriken

- Accuracy - Confusion Matrix - Begründung der Metrikauswahl

5.2. Gesamtergebnisse

- Accuracy auf dem Testsatz - Darstellung und Interpretation der Confusion Matrix

5.3. Ergebnisse nach Produktkategorien (optional)

- Vergleich der Modellleistung zwischen Kategorien - Kurze Analyse der Unterschiede

5.4. Ergebnisse von zwei Algorithmen (optional)

- Vergleich der Modellleistung zwischen Algorithmen (z.B. Naive Bayes und NN) - Kurze Analyse der Unterschiede

6. Schluss