LAB 4DTE LABOR FÜR DIGITALE TRANSFORMATION & ENTREPRENEURSHIP RAVENSBURG-WEINGARTEN UNIVERSITY OF APPLIED SCIENCES

IDW INSTITUT FÜR DIGITALEN WANDEL RAVENSBURG-WEINGARTEN UNIVERSITY OF APPLIED SCIENCES

# Analyse von Kundenfeedback

## Sentiment-Analyse und Topic Detection

Prof. Dr. Wolfram Höpken,
Forschungsgruppe Data Science, Institut für Digitalen Wandel (IDW)

## Einführung

- Große Mengen an Kundenfeedback und Produktbewertungen (UGC) verfügbar in nahezu allen Branchen
- Automatische Auswertung mittels Methoden des Text Mining unumgänglich (Sentiment Analysis & Topic Detection)

## Zielsetzung

- Sentiment Analysis & Topic Detection mittels wortlisten-basierter und supervised Machine Learning Ansätze
- Dynamische Erkennung von Topics auf unterschiedlichen Abstraktionsebenen mittels unsupervised Learning (Clustering) und aspektbasierten Verfahren

## Zusammenfassung

- Text Mining als sinnvolle Möglichkeit zur automatischen Analyse von Kundenfeedback und Produktbewertungen
- Unsupervised Topic Detection erkennt Topics dynamisch (Topic Drift) und fein-granular (Topic-Hierarchie)

## Methodik

### Datenextraktion & Aufbereitung

- Extraktion von Produktbewertungen mittels Web Crawling (Reguläre Ausdrücke und XPath)
- Text-Preprocessing: Tokenization, Entfernen von Stopp-Wörtern, Reduktion auf den Wortstamm, Part-of-Speech (POS) Tagging, N-Gramme, Erzeugung eines Word-Vektors (Bag of Words)

### Supervised Learning

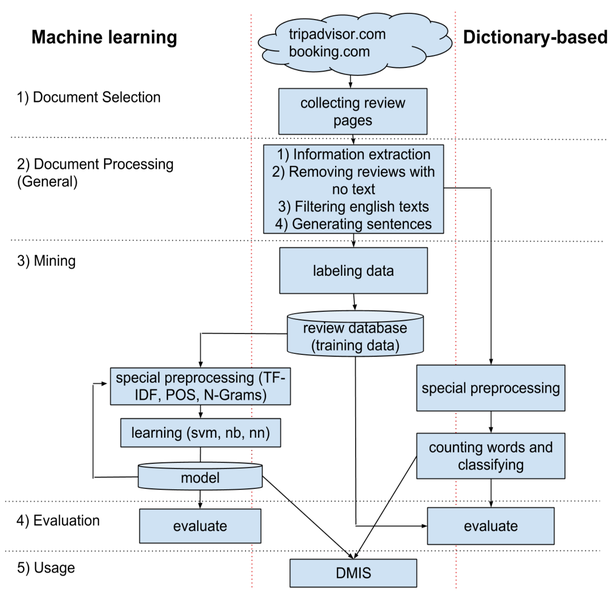
- Wortlistenbasierte Sentiment- und Topic Detection
- K-nearest-neighbors (k-NN), Naïve Bayes und Support Vector Machines (SVM)

### Unsupervised Learning

- Keyword Clustering (k-Means)
- Latent Semantic Indexing (LSI)

### Aspektorientierte Verfahren

- POS Tag Patterns
- Named Entity Recognition (NER) mit Conditional Random Fields (CRF)
- Dependency Parsing und SentiWordNet




## Ergebnis

### Ergebnisse Supervised Topic Detection

| Algorithmus | Güte | Kappa |
|---|---|---|
| Wortlisten-basiert | 71.0% | 0.642 |
| *k*-NN POS=NN.* n-Grams=no | 71.9% | 0.654 |
| SVM POS=no n-Grams=2 | 75.5% | 0.699 |
| Naïve Bayes POS=no n-Grams=2 | 51.5% | 0.427 |

### Ergebnisse Sentiment Detection

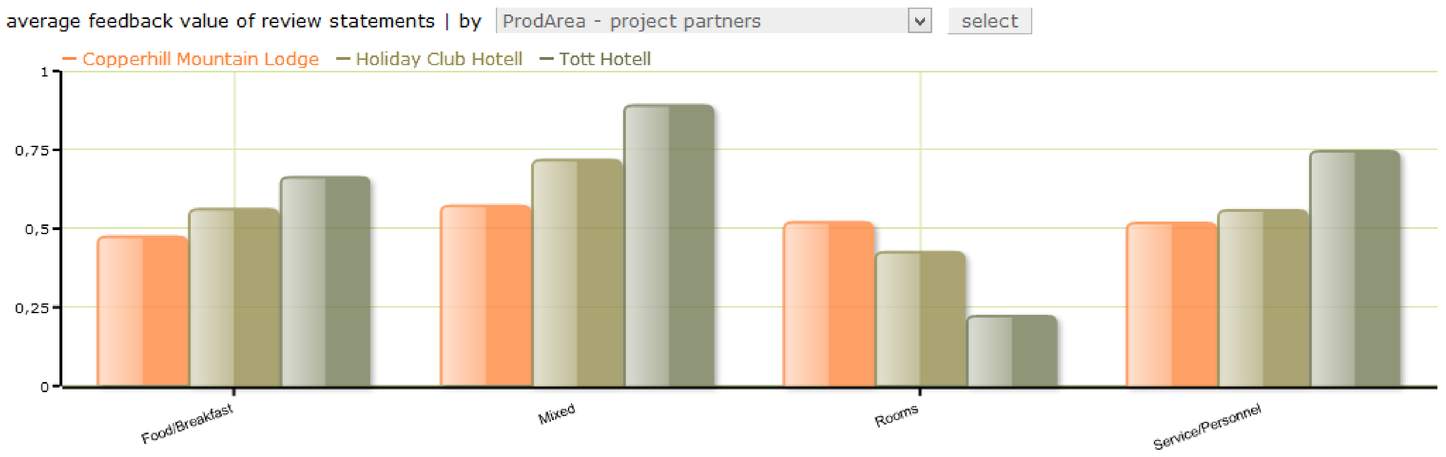
| Algorithmus | Güte | Kappa |
|---|---|---|
| Wortlisten-basiert | 68.6% | 0.449 |
| *k*-NN POS=no n-Grams=2 | 67.4% | 0.342 |
| SVM POS=no n-Grams=2 | 70.9% | 0.451 |
| Naïve Bayes POS=no n-Grams=2 | 59.1% | 0.305 |

### Ergebnisse Unsupervised Topic Detection mit k-means Clustering

| Unsupervised Topics (k=10) | | | | | Unsupervised Topics (k=25) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Topic vordefiniert* | *Cluster Keywords* | *Größe (%)* | *WCV* | *Güte* | *Topic* | *Cluster Keywords* | *Größe (%)* | *WCV* | *Güte* |
| Common Facilities | area, pool, spa | 8.8 | 1.05 | 73.4% | spa | area, pool, sauna, spa, treatment | 6.1 | 1.07 | 84.4% |
| FoodAnd Beverages | breakfast | 24.9 | 0.99 | 49.3% | breakfast | breakfast, place | 8.3 | 0.90 | 61.8% |
| | | | | | restaurant | restaurant | 3.7 | 1.13 | |
| | | | | | dinner | dinner, food | 3.5 | 1.10 | |
| | | | | | beverages | coffea, lobby, tea | 2.2 | 1.07 | |
| | | | | | food | breakfast, buffet, dinner, event, lunch | 3.2 | 0.87 | |
| Location | city, halmstad, location, station, town, train, walk | 9.2 | 1.04 | 71.5% | beach | beach, hotel, location | 3.1 | 1.18 | 68.5% |
| | | | | | centrality | center, city, halmstad, minutes, station, train, walk | 3.7 | 1.04 | |
| | | | | | town | town | 2.1 | 1.14 | |
| | | | | | parking | hotel, lot, parking, space | 3.0 | 1.10 | |
| Room | bathroom, floor, room, sea view, shower | 7.7 | 1.01 | 73.7% | bathroom | bathroom, floor, shower | 4.7 | 0.97 | 73.7% |
| | | | | | view | room, sea view | 3.0 | 1.17 | |
| | bed, room | 14.1 | 1.07 | | bed | bed, room | 3.5 | 0.75 | |
| | | | | | room costs | bit, room | 8.7 | 1.13 | |
| Staff | staff | 5.3 | 1.05 | 49.8% | reception | reception, staff | 4.6 | 1.25 | 76.2% |
| | food, service | 4.0 | 1.16 | | service | food, service, staff | 3.3 | 1.20 | |

*Keywords = Wörter mit TF-IDF-Wert > 0.05; WCV = Within Cluster Variation; Güte = Güte der Zuordnung zu vordefiniertem Topic*

### Benchmarking auf Basis des durchschnittlichen Sentiments pro Topic und Anbieter




## Literatur

1. Höpken, W., Fuchs, M., Menner, Th. and Lexhagen, M. 2017b. "Sensing the Online Social Sphere - the Sentiment Analytical Approach", Xiang, Z. and Fesenmaier, D.R. (Ed.s.), Analytics in Smart Tourism Design – Concepts and Methods, Springer, Cham: 129-146.
2. Menner, T., Höpken, W., Fuchs, M. and Lexhagen, M. 2016. "Topic detection – Identifying relevant topics in tourism reviews", in Inversini, A. and Schegg, R. (Ed.s.), Information and Communication Technologies in Tourism 2016, Springer, New York: 411-423.
3. Schmunk, S., Höpken, W., Fuchs, M. and Lexhagen, M. 2014. "Sentiment Analysis – Implementation and Evaluation of Methods for Sentiment Analysis with Rapid-Miner®", in Xiang, Ph. and Tussyadiah, I. (Ed.s.), Information and Communication Technologies in Tourism 2014, Springer, New York: 253-265)

## Kontakt

Prof. Dr. Wolfram Höpken
Leiter IDW / Forschungsgruppe Data Science
+49 751 501 9764
wolfram.hoepken@rwu.de