

Künstliche Intelligenz

Abschlussprojekt Teil 1

Sommersemester 2022

Gruppe 8

Julian Teschner, …

Marco Haas, 1000290

Medjen Izairi, …

Simon Vincent Abel, …

## Daten

Die Daten stammen aus dem Paper „From Group to Individual Labels using Deep Features“, Kotzias et. al. und wurden 2015 erhoben.

Für die Daten wurden Nutzerkommentare von 3 verschiedenen Websites, imdb.com, amazon.com und yelp.com, zufällig aus den Review Datensätzen ausgewählt. Bei den Daten wurde darauf geachtet, dass die Bewertung eindeutig in positiv und negativ unterteilbar ist. Es wurden insgesamt 500 positive und 500 negative Bewertungen von jeder Website ausgewählt.

Bei den Daten liegen lediglich Informationen zur Website sowie der Kommentar in Textform und das Label vor, welches anzeigt, ob der Kommentar positiv oder negativ ist.   
Das Label ist die Zielklasse, die von unserem Modell vorhergesagt werden soll.

## Trainingsprozess

Von jeder Website liegen 1000 Kommentare vor, wodurch sich insgesamt 3000 Instanzen ergeben.

Wir unterteilen die Daten wie in der Vorlesung dargestellt in die drei Teile Trainingsdaten, Entwicklungsdaten und Testdaten. Für die Unterteilung verwenden wir die Prozentsätze 70%-15%-15%. Somit ergeben sich folgende absolute Zahlen:

* Trainingsdaten: 2100 Instanzen
* Entwicklungsdaten: 450 Instanzen
* Testdaten: 450 Instanzen

Zur Aufteilung der Datensätze wurde in die unterschiedlichen Teile wurde Weka verwendet.

## Algorithmus

Wie in der Vorlesung wurde der Entscheidungsbaum Algorithmus C4.5 (J48 in Weka) verwendet.

Hierbei wird aus den Trainingsdaten ein Entscheidungsbaum erzeugt, mit welchem weitere Instanzen (zunächst die Entwicklungsdaten, anschließend die Testdaten und abschließend bei guter Evaluation produktive Instanzen) kategorisiert werden können. Hierzu wird eine Zielklasse und Attribute verwendet, welche dabei helfen, dass Model zu trainieren.

*C4.5 is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. These accuracy of each rule is then evaluated to determine the order in which they should be applied. Pruning is done by removing a rule’s precondition if the accuracy of the rule improves without it.*

<https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart>

## Merkmalsauswahl

Um unser Modell zu trainieren haben wir unterschiedliche Features zusammengetragen, welche eine Klassifizierung der Kommentare ermöglichen sollen. Hierbei handelt es sich einerseits um quantitative Merkmale, wie die Kommentarlänge aber auch um quantitative Merkmale, wie ein Abgleich mit positiven bzw. negativen Wortlisten.

* Anzahl der Wörter im Kommentar
* Anzahl der Satzzeichen
  + Anzahl der Satzzeichen insgesamt
  + Anzahl der Punkte
  + Anzahl der Kommas
  + Anzahl der Ausrufezeichen
  + Anzahl der Fragezeichen
* Anzahl positiver Wörter (zum Abgleich wurde eine Liste positiver Wörter verwendet -> https://gist.github.com/mkulakowski2/4289437)
* Anzahl negative Wörter (zum Abgleich wurde eine Liste negativer Wörter verwendet -> https://gist.github.com/mkulakowski2/4289441)
* Anzahl von Verneinungen im Kommentar (no / non)
* Kommentaranfänge (werden Kommentare mit I begonnen)
* Großschreibung (werden ganze Wörter mit Großbuchstaben geschrieben)
* Satzbau (z.B. fehlende Verben)
* Tokenizing
* Anzahl der Sätze/Nebensätze im Kommentar

Die oben genannten Features wurden mittels Python implementiert.

Bereits vor der ersten Verwendung des Weka-Tools zum Trainieren des Models hatten wir eine grobe Vorstellung, welche der genannten Merkmale erfolgsversprechender als andere sein könnten. Besonders vielversprechend war hierbei der Abgleich mit den positiven bzw. negativen Wortlisten.

Aber auch die Verwendung von Verneinungen wie no oder non konnten damit begründet werden, dass Menschen in einem negativen Kommentar eher diese Wörter verwenden als in einem positiven Kommentar.

Die Verwendung vieler Satzzeichen oder das Schreiben von Wörtern in Großbuchstaben wurde ebenfalls eher mit negativen Kommentaren assoziiert, da hierbei dem Kommentar besonderer Ausdruck verliehen werden sollte.

## Abschlussevaluation