# Projektdokumentation KI. Gruppe 8

## Daten

Die Daten stammen aus dem Paper „From Group to Individual Labels using Deep Features“, Kotzias et. al. Die Daten 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 nur die Website, der Kommentar in Textform und das Label, ob der Kommentar positiv oder negativ ist, vor.   
Das Label ist die Zielklasse, die von unserem Modell vorhergesagt werden soll.

## Trainingsprozess

Es liegen von jeder Website 1000 Instanzen vor, also insgesamt 3000 Instanzen.

Wir unterteilen die Daten in drei Teile, Trainingsdaten, Entwicklungsdaten und Testdaten. Für die Unterteilung verwenden wir die Prozentsätze 70%-15%-15%.

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

## Algorithmus

Wir haben den Entscheidungsbaum Algorithmus C4.5 (J48 in Weka) verwendet.

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

* Anzahl der Wörter
* Anzahl der Satzzeichen
* Anzahl positive Wörter
* Anzahl negative Wörter
* Satzbau