Sentimentanalyse –
Anwendung und Kategorisierung auf existierende Texte

Dieses Projekt behandelt die Sentimentanalyse, welche auch unter dem Begriff *Opinion Mining* bekannt ist. Mithilfe der Sentimentanalyse werden subjektive Informationen aus Texten extrahiert und Wörter, Texte und Dokumente klassifiziert, üblicherweise wird hier zwischen positiver, neutraler oder negativer Stimmung unterschieden. Der Fokus liegt dabei auf den Anwendungen und Kategorisierungen von Texten.
Anwendungsgebiete der Sentimentanalyse sind beispielsweise Produkt- oder Dienstleistungsbewertungen und auch verschiedenste Textarten, wie etwa Zeitungsartikel, Foreneinträge, Kommentare oder Tweets.

In der heutigen Zeit kann man dank der vielen Online-Plattformen schnell und unkompliziert seine Meinungen über diverse Produkte, Dienstleistungen, Veranstaltungen oder sonstige Themen teilen. Da schon so gut wie jedes Produkt oder ähnliches mindestens einmal bewertet wurde, existiert eine Vielzahl von Einträgen. Durch die hohe Menge an Daten gestaltet sich die Untersuchung ohne technische Unterstützung als sehr aufwändig und kompliziert. Durch die Sentimentanalyse soll dieser Prozess erleichtert werden.
Einerseits können beispielsweise Standpunkte und Feedbacks der Kunden schnell und effizient erfasst und daraufhin vom Unternehmen umgesetzt werden. Andererseits können auch die Meinungen, welche sich in einem Artikel oder Kommentar reflektieren, herausgefunden werden.

Im Bereich der Sentimentanalyse wurden über die Jahre hinweg diverse Ansätze entwickelt, welche in lexikon- und lernbasierte Ansätze unterteilt werden. Das Ziel dieses Projekts ist es, diese Ansätze miteinander zu vergleichen und daraufhin darzustellen, welcher Ansatz für welche Kategorie von Texten geeignet ist.

Folgende Ansätze wurden für dieses Projekt in Betracht gezogen:

- Lexikonbasierte Ansätze
  - ☐ TextBlob
- Lernbasierte Ansätze
  - ☐ Decision Tree
  - ☐ Support Vector Machines
  - ☐ Neuronales Netz
  - ☐ Naive Bayes
  - ☐ Maximum Entropy

Diese Ansätze wurden vom Projektteam implementiert und anhand von Trainings- und Testdaten ausgewertet. Mithilfe einiger Kennzahlen wurden die Ergebnisse der Ansätze verglichen und somit erfolgte eine erste Selektierung, mit welchen Ansätzen weitergearbeitet wurde. Dabei fiel die Entscheidung auf **Decision Tree, Support Vector Machines und Neuronales Netz.**

Nach weiteren Verfeinerungen und Verbesserungen konnten diese drei Ansätze weiterentwickelt werden. Durch die Veränderungen im Code konnten wesentlich bessere Kennzahlen erzielt werden und somit wurden die Ansätze in weiterer Folge mit den realen Artikeln und Kommentaren getestet.
Dabei war ersichtlich, dass der **Support Vector Machines** Klassifikator sich nicht für große Datenmengen eignet, da die Datei der Kommentare einen Programmabbruch herbeiführte. Auch im Bezug auf die Zeitungsartikel schnitt er, im Vergleich zu den anderen, am schlechtesten ab.

Somit wurde der Fokus auf die zwei übrig gebliebenen Ansätze **Decision Tree** und **Neuronales Netz** gesetzt, welche in Folge noch einmal überarbeitet und mit den realen Daten getestet wurden.
Wie in *Abbildung 1* ersichtlich, wurden die Kennzahlen der Ansätze gegenübergestellt, um ein eindeutiges Ergebnis festhalten zu können:

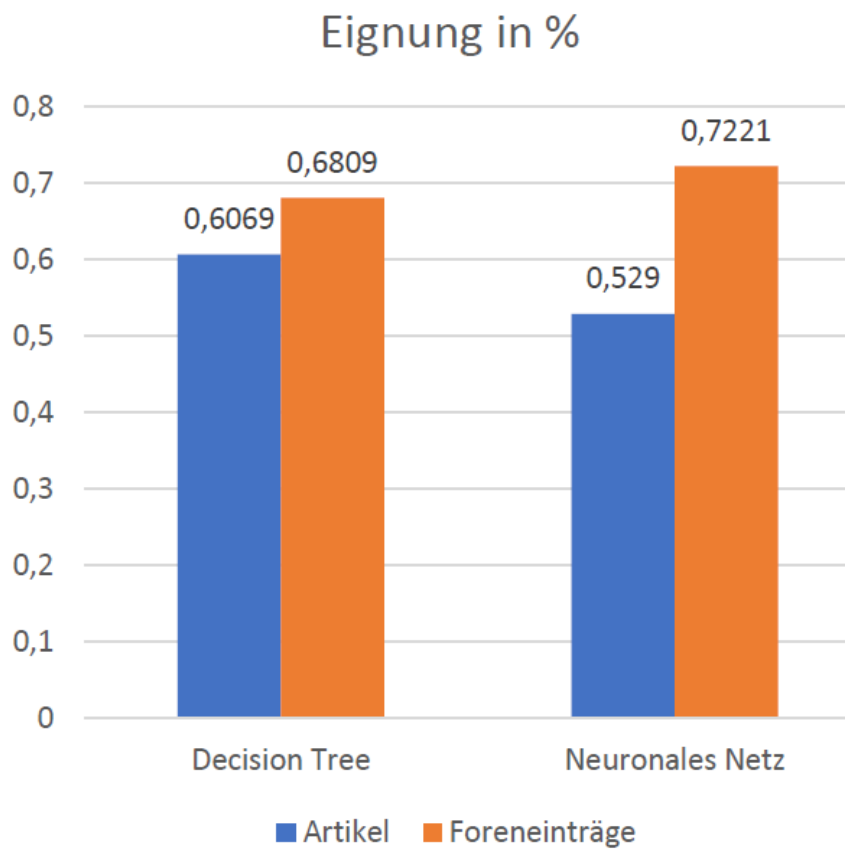

Abbildung 1: Eignung in %

Zusammenfassend lässt sich sagen, dass sich der **Decision Tree** Klassifikator aufgrund des höheren Ergebnisses besser für die Analyse von Zeitungsartikeln eignet. Dafür schneidet der Klassifikator des **Neuronalen Netzes** bei Kommentaren bzw. Foreneinträgen besser ab.