

law 1.1

Juliette-Michelle Burkhardt

April 2024

1 Einleitung

Die Künstliche Intelligenz (KI), die entwickelt wurde, zielt darauf ab, die Wirksamkeit von Vertragsklauseln automatisch zu klassifizieren. Dieses Modell verwendet Textdaten als Eingabe, um zu bestimmen, ob eine Klausel wirksam (1) oder unwirksam (0) ist.

2 Datenvorbereitung

2.1 Datenbeschaffung

Die Daten für das Modell stammen aus zwei CSV-Dateien: eine enthält Beispiele für wirksame Klauseln, während die andere unwirksame Klauseln enthält. Diese werden zu einem einzigen Datensatz kombiniert.

2.2 Datenreinigung und -vorverarbeitung

Die Texte der Klauseln werden normalisiert, indem sie in Kleinbuchstaben umgewandelt und von überflüssigen Zeichen bereinigt werden. Dies dient dazu, die Konsistenz im Datensatz zu erhöhen und die Leistung des Modells zu verbessern.

3 Modellentwicklung

3.1 Feature-Extraktion

Mithilfe der Methode *CountVectorizer* werden die Textdaten in numerische Vektoren umgewandelt, die als Eingabe für das Modell dienen.

3.2 Modellausbildung

Das Modell, eine logistische Regression, wird mit den vorbereiteten Daten trainiert. Es lernt, auf Grundlage der Textfeatures zwischen wirksamen und unwirksamen Klauseln zu unterscheiden.

4 Modellbewertung

4.1 Leistungsmetriken

Die Leistung des Modells wird anhand der Genauigkeit (Accuracy) und der Konfusionsmatrix bewertet. Diese Metriken geben Aufschluss darüber, wie gut das Modell bei der Vorhersage der Wirksamkeit von Klauseln abschneidet.

4.2 Visualisierung

Die Ergebnisse werden visuell aufbereitet, insbesondere durch die Darstellung der Konfusionsmatrix, um ein klares Bild der Modelleleistung zu erhalten.

5 Results

Accuracy: 71.4%

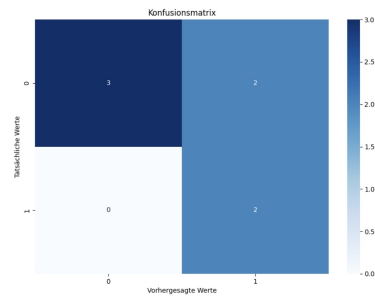


Figure 1: Hier steht die Bildunterschrift.

6 Limitations

6.1 Generalisierbarkeit

Das Modell basiert auf logistischer Regression, einem einfachen Ansatz, der möglicherweise nicht in der Lage ist, komplexe Muster in den Daten zu erfassen. Dies kann zu einer suboptimalen Leistung in realen oder komplexen Anwendungsfällen führen.

6.2 Feature-Extraktion

Die Verwendung von `CountVectorizer` beschränkt die Analyse auf die Häufigkeit von Wörtern, ohne Kontext oder Reihenfolge zu berücksichtigen, was zu einer weniger nuancierten Modelleleistung führen kann.

6.3 Balancierung des Datensatzes

Ein unausgewogener Datensatz kann dazu führen, dass das Modell die dominante Klasse bevorzugt, was die Vorhersagegenauigkeit für die unterrepräsentierte Klasse verringert.

6.4 Interpretierbarkeit

Trotz der relativen Einfachheit von logistischer Regression kann es schwierig sein, die Entscheidungsfindung des Modells im Detail zu verstehen, insbesondere bei der Bewertung der Feature-Gewichtung.

6.5 Sprachliche Vielfalt und Kontext

Die Effektivität von Klauseln kann stark von spezifischen juristischen, kulturellen und sprachlichen Kontexten abhängen, was die Übertragbarkeit des Modells einschränken kann.