Steckbrief für NLP-Experimente

Stand: 15. Dezember 2022

Der Task

1. Die Aufgabe heißt: _____
2. Es handelt sich um
   - ☐ Textklassifikation, ☐ Sequence Labeling, oder ☐ Sonstiges: _____
3. Die zu klassifizierenden Instanzen sind: _____
4. Es gibt _____ Kategorien/Klassen.

Die Daten

1. Annotierte Daten
   - ☐ liegen bereits vor, oder ☐ müssen noch annotiert werden.
2. In den Daten sind _____ Instanzen (von o.g. Typ) annotiert.
3. Die Klassen sind
   - ☐ gleichverteilt (d.h. jede Klasse ist ungefähr gleich häufig)
   - ☐ unterschiedlich verteilt, und zwar: _____

Die Baseline

- ☐ Weil die Klassen ungleich verteilt sind, bietet sich eine *majority baseline* an. Diese erzielt eine *accuracy* von _____ %.
- ☐ Weil die Klassen gleich verteilt sind, bietet sich eine *random baseline* an. Diese erzielt eine *accuracy* von _____ %.
- ☐ Eine weitere mögliche Baseline ist: _____. Diese erzielt eine *accuracy* von _____%.
- ☐ Eine weitere mögliche Baseline ist: _____. Diese erzielt eine *accuracy* von _____%.

Das Experiment

1. Ich möchte das folgende oder die folgenden Verfahren verwenden:
   - ☐ Entscheidungsbaum / decision tree
   - ☐ Naive Bayes
   - ☐ Support Vector Machines
   - ☐ Logistic Regression
   - ☐ Neural Networks
   - ☐ Sonstige: _____
2. Ich möchte die folgenden Features verwenden
   - ☐ Metadaten: _____
   - ☐ Textdaten:
     - ☐ Worthäufigkeiten (von allen Wörtern)
     - ☐ Häufigkeiten von Wörtern aus folgenden Wortlisten: _____
     - ☐ Word Embeddings
     - ☐ Sequenzielle Information (d.h. Klassifikationsergebnisse für Elemente davor oder danach)

- ☐ N-Gram-Häufigkeiten, mit $N \leq$_____.

3. Meine Features haben die folgenden Datentypen:
   - ☐ Numerisch: _____ (Anzahl an Features)
   - ☐ Kategorial: _____ (Anzahl an Features)
4. Testdaten
   - ☐ Ich teile mein o.g. Datensatz in Trainigs- und Testdaten auf. _____% des Datensatzes sind Trainingsdaten
   - ☐ Ich verwende N-fold cross validation, mit $N =$_____.
   - ☐ Trainings- und Testdaten sind bereits aufgeteilt.

Die Auswertung

1. Ich verwende die Evaluationsmetrik(en)
   - ☐ Accuracy
   - ☐ Precision
   - ☐ Recall
   - ☐ F-Measure
   - ☐ Area under curve (AUC)
   - ☐ _____

- ☐ Für meine Fehleranalyse inspiziere ich _____ Instanzen manuell.