Checkliste zu NLP-Klassifikations-Experimenten

Stand: 14. November 2024

Hinweise

- Wenn Ihr Task kein Klassifikationstask ist, dann ist dieser Fragebogen nicht für Sie.
  - Keine Klassifikationstasks sind z.B. Generierung oder Übersetzung.
- Markieren Sie alle Punkte die Sie planen umzusetzen.
- Machen Sie einen Termin in der Sprechstunde, wenn Ihnen Punkte unklar sind.
- Der Fragebogen ist keine Prüfung, sondern dient als Hilfestellung, bei der Experimentplanung an alles zu denken, auf Ideen zu kommen und ggf. die richtigen Fragen zu stellen. Diese klären wir dann idealerweise im Gespräch.
- Der Fragebogen repräsentiert eine Planung. Gewisse Abweichungen von der Planung sind normal und zu erwarten.
- Bei Fragen schreiben Sie gerne eine E-Mail an `nils.reiter@uni-koeln.de` oder melden Sie sich gerne zu einer Sprechstunde an. Wenn Sie rein technische Fragen zur Infrastruktur haben, schreiben Sie bitte an `spinfo-admin@uni-koeln.de`.

Der Task

1. Die Aufgabe heißt: Ressorterkennung anhand von Überschriften
2. Es handelt sich um ☒ Textklassifikation, ☐ Sequence Labeling, oder ☐ Sonstiges: _____
3. Die zu klassifizierenden Instanzen sind: Zeitungsüberschriften
4. Es gibt 7 Kategorien/Klassen.
5. Einer Instanz kann ☒ genau eine oder ☐ mehrere Klassen zugewiesen werden.

Die Daten

1. Annotierte Daten ☐ liegen bereits vor oder ☒ müssen noch erstellt werden.
2. In den Daten sind 100 / Klasse Instanzen (von o.g. Typ) annotiert.
3. Die Klassen sind
   - ☒ gleichverteilt (d.h. jede Klasse ist ungefähr gleich häufig)
   - ☐ unterschiedlich verteilt, und zwar: _____

Die Annotationen

Nur relevant, wenn neue Daten annotiert werden sollen.

1. Annotationsrichtlinien
   - ☐ Ich verwende die folgenden, bereits existierenden Annotationsrichtlinien: _____
     - Mit diesen wurde ein Inter-Annotator-Agreement von _____ erzielt (Metrik: _____).
   - ☐ Ich schreibe neue Annotationsrichtlinien.

2. Annotator:innen
   - ☐ Ich annotiere selbst.
   - ☐ Ich rekrutiere Annotator:innen aus meinem Freundes-/Bekanntenkreis.
   - ☐ Ich sammle Annotationen über eine Umfrage, z.B. mittels LimeSurvey.
     - ☐ Ich verfüge über einen Verteiler, über den ich die Umfrage bewerbe.
   - ☐ Ich sammle Annotationen über crowd sourcing.
3. Annotationsworkflow
   - ☐ Annotator:innen treffen eine Annotationsentscheidung auf der Basis eines Kontextes von ___ ☐ Wörtern, ☐ Sätzen, ☐ Zeilen, ☐ Absätzen, ☐ _____; oder ☐ sie verwenden den gesamten Text als Kontext.
   - ☐ Sie können dabei außerdem die folgenden Wissensquellen verwenden: ☐ Wikipedia, ☐ Lexika, ☐ Wörterbücher, ☐ Sonstige: _____
4. Anforderungen an Annotationssoftware
   - ☐ Annotator:innen müssen Spannen selbst markieren können.
   - ☐ Annotator:innen müssen neue Kategorien oder Labels ergänzen können.

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 15 %.
- ☐ Eine weitere mögliche Baseline ist: _____.
  Diese erzielt eine Accuracy von _____%.
- ☐ Eine weitere mögliche Baseline ist: _____.
  Diese erzielt eine Accuracy von _____%.

Das Experiment

Ich möchte das folgende oder die folgenden Verfahren verwenden:

- ☐ Klassisches Machine Learning
  - ☐ Entscheidungsbaum / Decision Tree (DT)
  - ☐ Naive Bayes
  - ☐ Support Vector Machines (SVM)
  - ☐ Logistic Regression
  - ☐ Neuronale Netze (NN)
    - ☐ Feed-Forward Neural Networks
    - ☐ Convolutional Neural Networks
    - ☐ Recurrent Neural Networks
- ☒ Große Sprachmodelle / Transformer (inkl. BERT & co.)
  - ☐ Welche(s) Modell(e) genau? _____
  - ☐ Fine-Tuning
  - ☒ Prompting
- ☐ Sonstige: _____

Klassisches Machine Learning
Nur relevant, wenn klassisches Machine Learning verwendet werden soll.

1. Ich möchte die folgenden Features verwenden
   - ☐ Metadaten: _____
     _____
     _____
   - ☐ Inhaltsdaten, z.B. aus Texten:
     - ☐ Worthäufigkeiten (von allen Wörtern), auch bekannt als bag of words
     - ☐ Häufigkeiten von Wörtern aus folgenden Wortlisten: _____
     - ☐ Embeddings (z.B. Word Embeddings)
     - ☐ Sequenzielle Information (d.h. Klassifikationsergebnisse für Elemente davor oder danach)
     - ☐ N-Gram-Häufigkeiten, mit $N \leq$_____
     - ☐ Thematische Informationen aus einem Topic Model (z.B. Latent Dirichlet Allocation, LDA)
2. Meine Features haben die folgenden Datentypen:
   - ☐ Numerisch: _____ (Anzahl Features) ☐ Kategorial: _____ (Anzahl Features)

Große Sprachmodelle / Transformer
Nur relevant, wenn große Sprachmodelle verwendet werden sollen.

- ☐ Ich möchte Fine-Tuning verwenden
- ☒ Ich möchte Prompting verwenden, und verwende dabei die folgenden Strategien:
  - ☒ Hard prompting (d.h., der Prompt besteht aus Wörtern) mit den Komponenten:
    - ☒ Beschreibung der Aufgabe
    - ☒ Versprechen einer Belohnung
    - ☐ Aufforderung, Schritt für Schritt vorzugehen
    - ☒ Definition einer Rolle, die das Modell einnehmen soll
    - ☐ Beispiele – ☐ Positive oder ☐ Negative
    - ☐ Sonstiges: _____
  - ☐ Soft prompting (d.h., der Prompt besteht aus Vektoren) mit dieser/diesen Technik/en: _____
    _____
    _____
- ☒ Das Modell soll folgenden Output produzieren
  - ☒ Numerischen
  - ☐ Sprachlichen. Dieser soll wie folgt ausgewertet werden: _____
    _____
    _____
  - ☐ JSON oder YAML.
  - ☐ Sonstigen: _____
  - ☒ Mit nicht-wohlgeformtem Output möchte ich wie folgt umgehen: _____
    Nicht wohlgeformter Output wird als Fehler gezählt

Beide Varianten

1. Testdaten
   - ☒ Ich teile meinen o.g. Datensatz selbst in Trainigs- und Testdaten auf. 0 % der Instanzen werden als Trainingsdaten verwendet.
   - ☐ Ich verwende N-fold cross validation, mit $N =$ _____.
   - ☐ Trainings- und Testdaten sind bereits aufgeteilt, z.B. weil es Daten aus einem shared task sind.
2. Ich variiere und vergleiche
   - ☐ die Größe des Trainingsdatensatzes (z.B. 100, 1000, 10 000 Instanzen für den Trainingsdatensatz)
   - ☐ die Menge an oder Art von Features die verwendet werden (z.B. inhaltliche vs. sprachliche Features)
   - ☐ das Verfahren als solches oder Parameter davon (z.B. NN vs. SVM)
   - ☐ die Vorverarbeitung (z.B. Groß- und Kleinschreibung)
   - ☒ die Prompt-Komponenten (z.B. mit unterschiedlichen Belohnungen)
3. Meine Hypothese ist: _____

Die Auswertung und Evaluation

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

☐ Meine Testdaten sind stark unbalanciert, daher verwende ich die Metriken in der Mikro- und Makro-Average-Variante.

2. Für meine Fehleranalyse inspiziere ich 20 Instanzen manuell.

Die praktische Umsetzung

1. Ich verwende die Programmiersprache
   - ☒ Python
   - ☐ Java
   - ☐ R
   - ☐ _____ ×
2. Hardware-Ausstattung und Vorkenntnisse
   - ☐ Ich verfüge über einen Computer
     - ☐ der auch mal über Nacht durchlaufen kann, wenn eine Berechnung etwas länger dauert.
     - ☐ der eine GPU mit CUDA-Unterstützung hat oder ein Mac mit M1/M2-Prozessor ist.
     - ☐ der ausreichend freien Plattenspeicher hat.
   - ☐ Ich möchte Berechnungen auf einem Server der Universität laufen lassen.
     - ☐ Ich kann mich per SSH auf einem Server einloggen.
     - ☐ Ich weiß wie ich auf einer Kommandozeile ein Programm laufen lasse.