HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

SemEval-2018 Task 12 - The Argument Reasoning Comprehension Task

Team HHU: Matthias Liebeck, Andreas Funke

Andreas Funke

SemEval Task 12 – Argument Reasoning Comprehension Task

- from Wikipedia:
  „**SemEval** (**Sem**antic **Eval**uation) is an ongoing series of evaluations of computational semantic analysis systems.“
- Task 12 – Organizers:
  UKP TU Darmstadt, Webis Bauhaus-Universität Weimar
- **Argument Mining:**
  Identifikation von Argumenten und Argumentationsketten in natürlichsprachlichen Texten
- **Argumentationsmodell** hier:
  reason → warrant → claim

Argumentationsmodell – Beispiel

- **claim:**
  „Cigarettes are bad for your health.“
- **reason** *supports* claim:
  „Studies show that cigarettes can cause cancer.“
- **warrant** *connects* reason and claim:
  „Cancer harms your health.“
- **alternative warrant** *negates* warrant**:**
  „Cancer doesn't harm your health.“

- **Challenge:**
  Bestimme den korrekten warrant (binäre Klassifikation).

Datensatz

- **Datenquelle:**
  - *Room For Debate* der *New York Times*
  - annotiert via Crowdsourcing
- **Felder:**
  - ID | Label (∈ {0, 1}) | warrant0 | warrant1 | reason | claim
  - zusätzlich: debate title | debate info
- **3-fach Split:**
  - train set (1210 Items)
  - dev set (316 Items)
  - test set (444 Items ohne Labels)

Trainingsinstanz – Beispiel

ID: 7951790_153_A1I4CYG5YDFTYM

Title: Do We Still Need Libraries?

Debate info: What are libraries for, and how should they evolve?

Claim: We need libraries

Reason: Libraries have lots to offer in addition to books they provide music, dvd's, magazines and more.

Warrant0: all these things are readily available to everyone online

Warrant1: none of these things are readily available to everyone online

Label: 1

Baseline Accuracy Scores

Approach	Dev	(±)	Test	(±)
externe Baselines				
Human average			0.798	0.162
Human w/ training			0.909	0.114
Attention LSTM	**0.632**	0.034	0.570	0.008
eigene Baseline				
SVM	0.588	0.028		

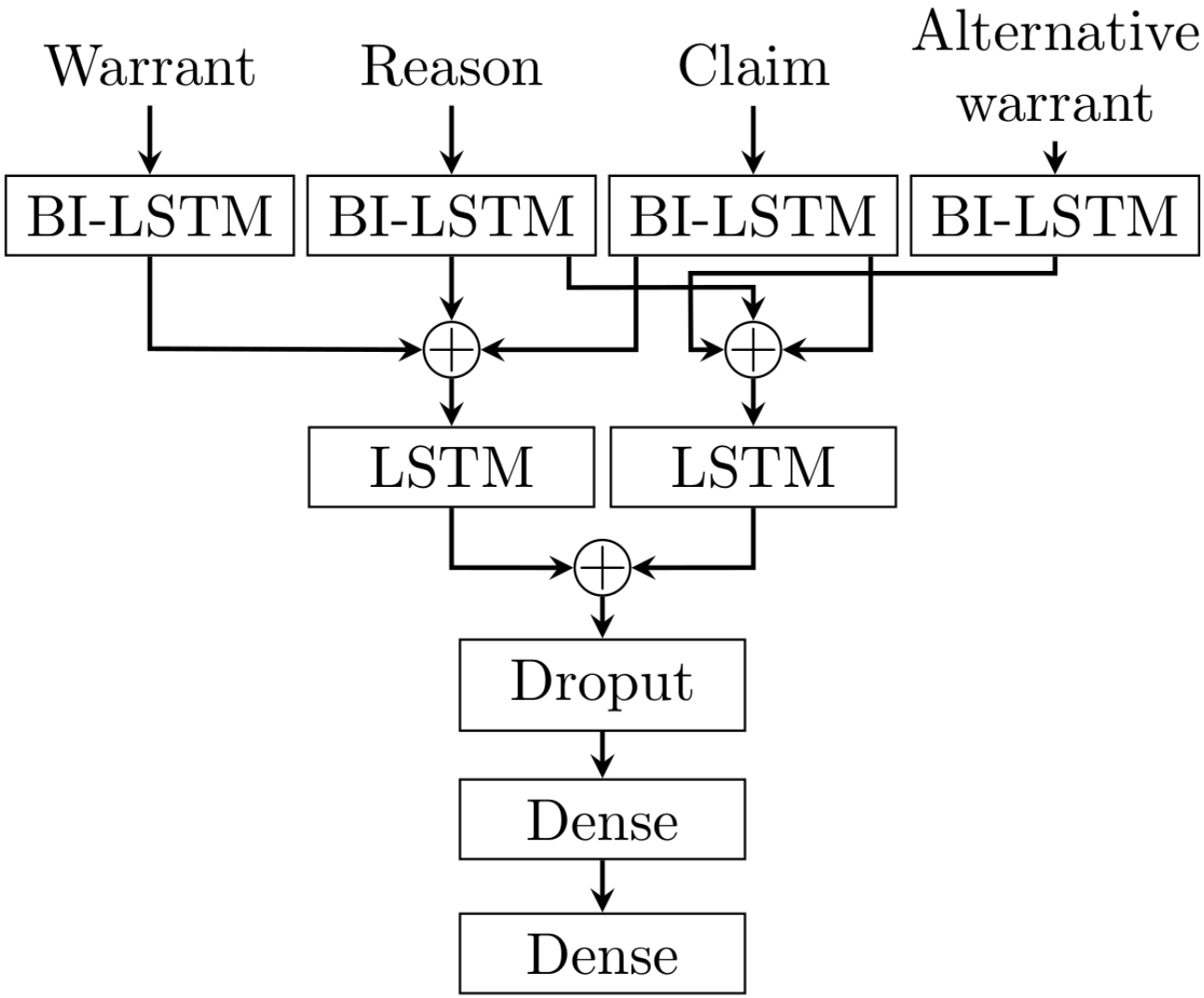
Deep Learning Ansatz – Hyperparameterwahl

- **Keras** als Frontend
- **zweistufiges Gridsearch** (approximiert)
  u.a. über folgende Hyperparameter:
  - Backend: TensorFlow or Theano
  - NN architecture (layer choices)
  - Layer sizes
  - Padding width
  - Activation function
  - Optimizer
  - Loss function
  - Dropout ratio
  - Crossvalidation split ratio
  - Batch size
  - Used data fields
  - Embedding corpus
  - Embedding dimensionality
  - Case folding …

Embedding-Auswahl

- **extern:**
  - word2vec (GoogleNews)
  - **fastText (Wikipedia)**
  - GloVe (Wikipedia)
- **eigene w2v Trainings:**
  - **Task-Datensatz**
  - **Task-Datensatz + für das Vokabular relevante Wikipedia-Artikel**
  - jeweils verschiedene Dimensionalitäten / case folding

Finale Architektur



Ensemble-Ansatz

- **finale Hyperparameter-Bestimmung:**
  - 1 Architektur
  - 4 Embeddings
  - 64 Hyperparameter-Kombinationen
  - 10 Seeds
  - = 2560 Modelle
- **Idee:**
  - statt 2559 Modelle zu verwerfen:
    kombiniere Einzelpredictions zu Ensemble
- **Ausführung:**
  - sortiere Einzelmodelle absteigend nach Accuracy-Score auf dev set
  - Majority vote über alle Modelle mit Score > 0.67 (623 Modelle)

Final Dev Set Accuracy Scores

Approach	Dev	(±)
externe Baselines		
Attention LSTM	0.632	0.034
Best trial result	0.703	-
eigene Baseline		
SVM	0.588	0.028
eigene Scores		
Single model avg.	0.677	0.022
Single model max.	0.712	-
Ensemble	**0.733**	-

Official Test Set Accuracy Scores

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Platz	Team	Test
1.	GIST	0.712
2.	blcu_nlp	0.606
3.	ECNU	0.604
4.	NLITrans	0.590
5.	Joker	0.586
6.	YNU_Deep	0.583
7.	mingyan	0.581
8.	ArcNet	0.577
...	...	...
16.	TakeLab	0.541
17.	**HHU**	**0.534**
18.	Random baseline	0.527
19.	Deepfinder	0.525
20.	ART	0.518
21.	RW2C	0.500
22.	ztangfdu	0.464

Ursachenforschung

- **mögliche Gründe:**

  (a) Overfitting auf dem dev set

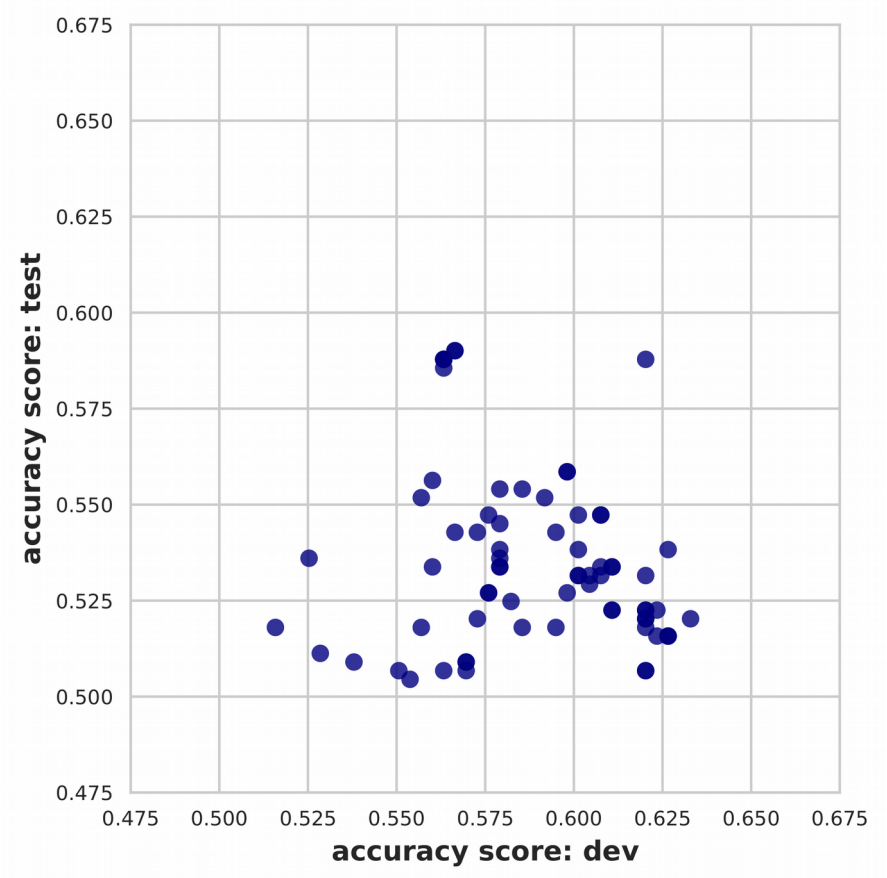
  (b) Data sets stammen aus unterschiedlichen Verteilungen

- **Hypothese:**
  - Da alle Scores auf dem test set niedriger ausfallen: **(b)**

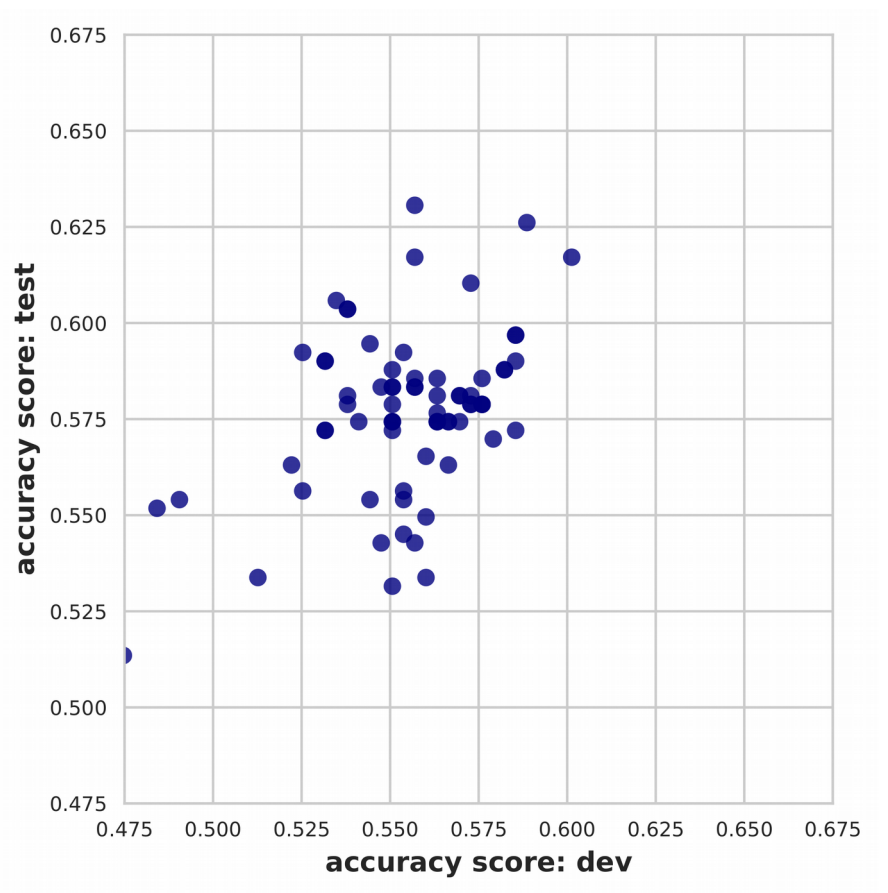
- **Validierung: Alternativer Split**
  - erzeuge neuen, randomisierten Split für train, dev und test sets
  - die Größe der sets wird vom originalen Split übernommen
  - wiederhole alle Trainings und Predictions mit alternativem Split
  - vergleiche Ergebnisse von originalem und alternativem Split

SVM Baseline – original vs. alternative Split

- original Split

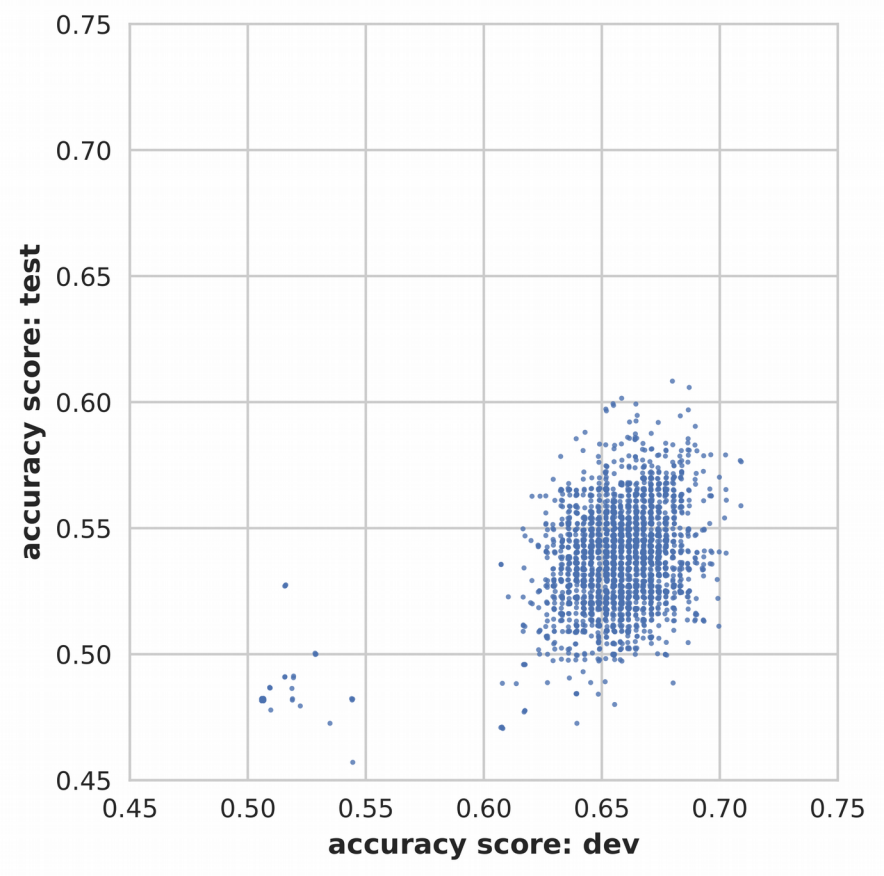


- alternative Split

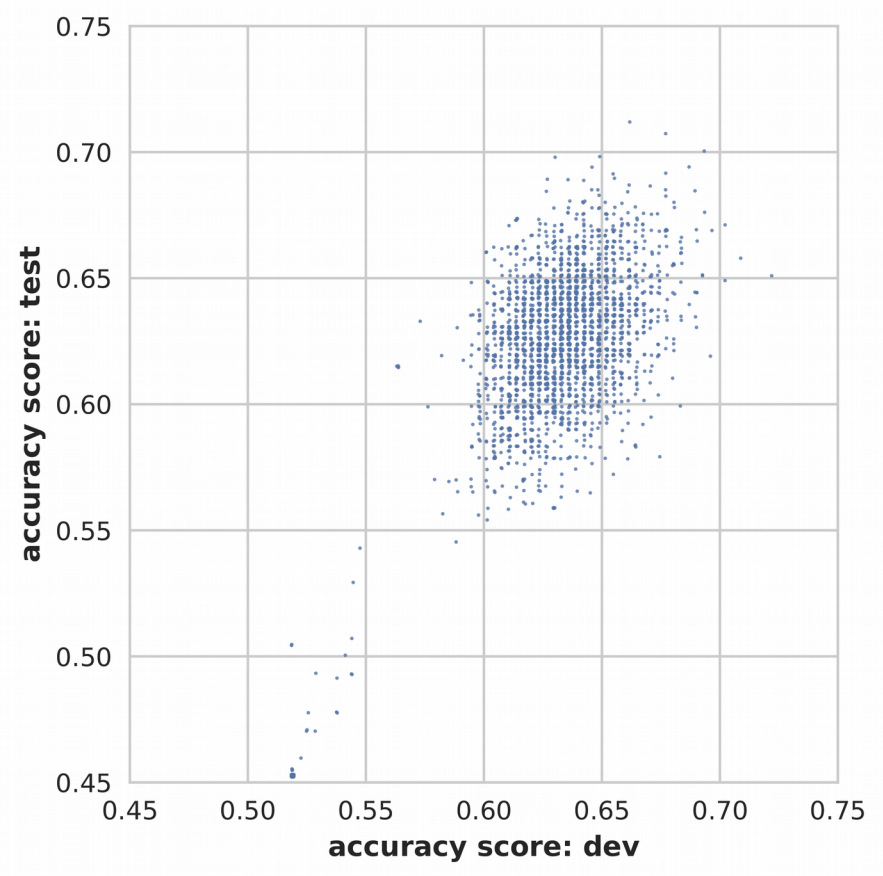


Single NN model – original vs. alternative Split

- original Split

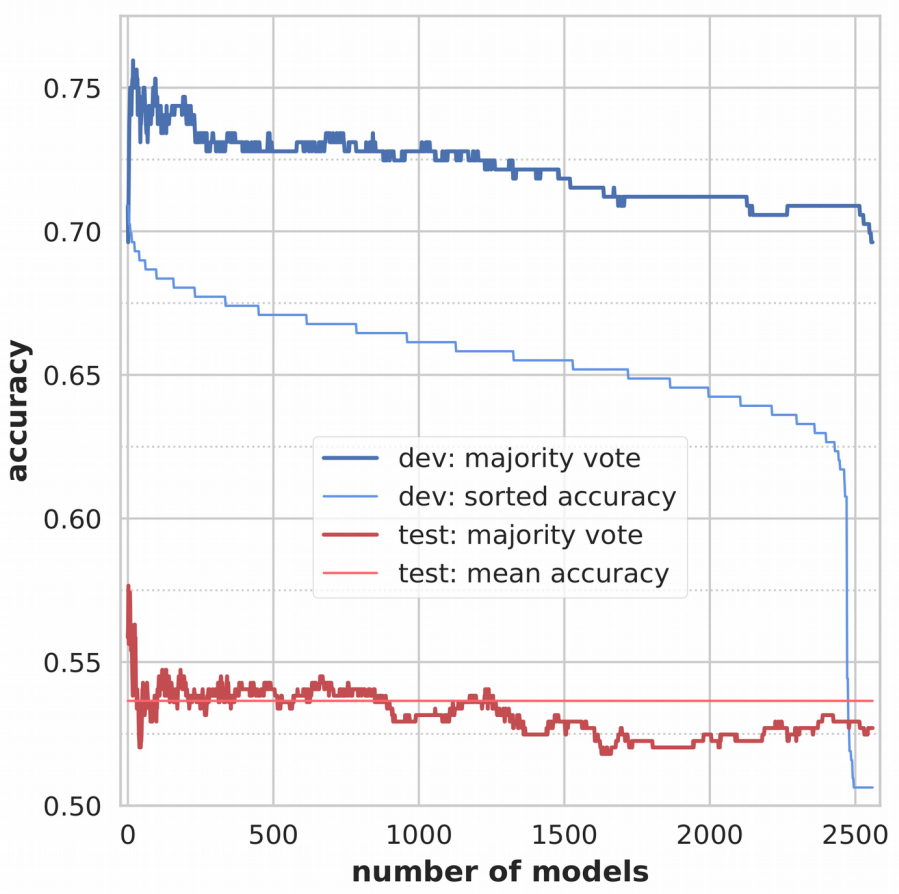


- alternative Split

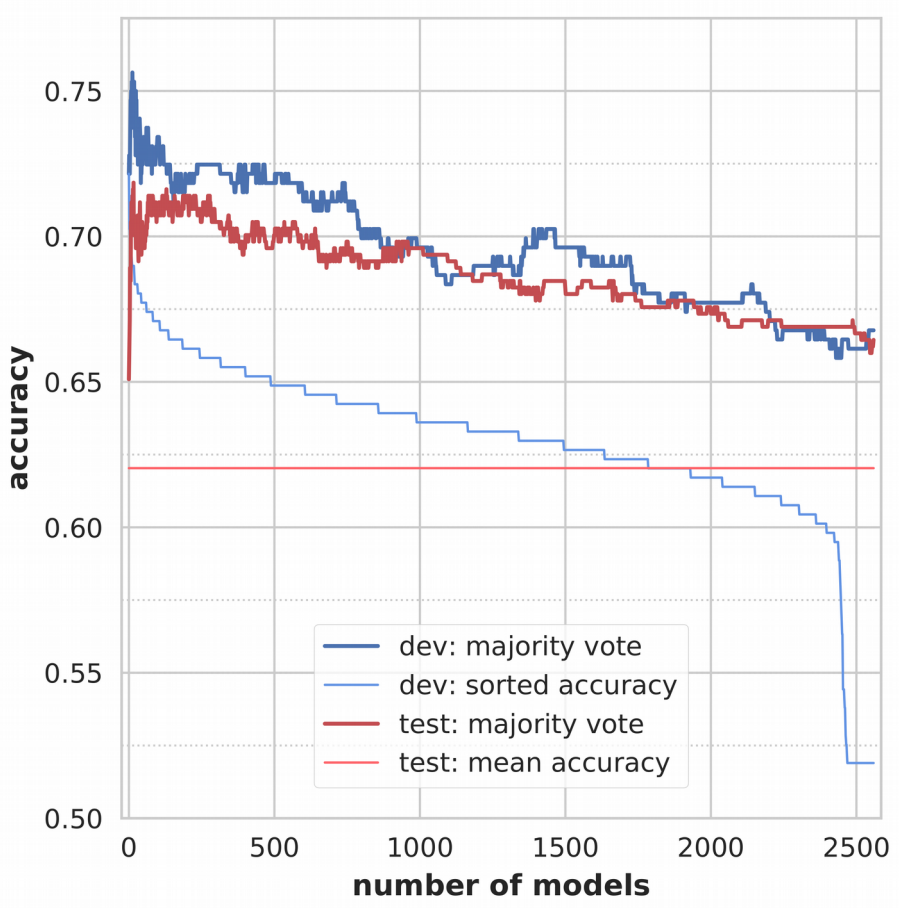


Ensemble Performance – original vs. alternative Split

- original Split



- alternative Split



Fazit und Learnings

- **Ensemble-Ansatz:**
  - Grundsätzlich liefert das Boosting durch ein DL-Ensemble sehr gute Scores.
  - nur geringes Overfitting auf dem dev set festzustellen
  - Voraussetzung hierfür:
    train, dev und test set stammen aus der gleichen Verteilung
  - Die Peak-Performance auf dem dev set liefert auch die beste Performance auf dem Test set => wenige Modelle genügen

- **Argument Reasoning Comprehension Task:**
  - Ein sorgfältigeres Feature-Engineering vor dem Training hätte bei Daten aus unterschiedlichen Verteilungen (wie hier gegeben) bessere und verlässlichere Ergebnisse erzielt.

Vielen Dank

für Ihre Aufmerksamkeit!

Quellen

- SemEval-2018. International Workshop on Semantic Evaluation. http://alt.qcri.org/semeval2018/
- SemEval-2018 Task 12 - The Argument Reasoning Comprehension Task. https://competitions.codalab.org/competitions/17327
- Habernal, Wachsmuth, Gurevych, Stein: The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In Proceedings: NAACL 2018. https://arxiv.org/abs/1708.01425
- Habernal et.al: competition source: https://github.com/habernal/semeval2018-task12
- Habernal et al: task source: https://github.com/UKPLab/argument-reasoning-comprehension-task/
- Liebeck, Funke, Condrad: HHU at SemEval-2018 Task 12: Analyzing an Ensemble-based Deep Learning Approach for the Argument Mining Task of Choosing the Correct Warrant. Accepted for publication in Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)
- Andreas Funke: HHU submission source: https://github.com/andifunke/semeval18task12