LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

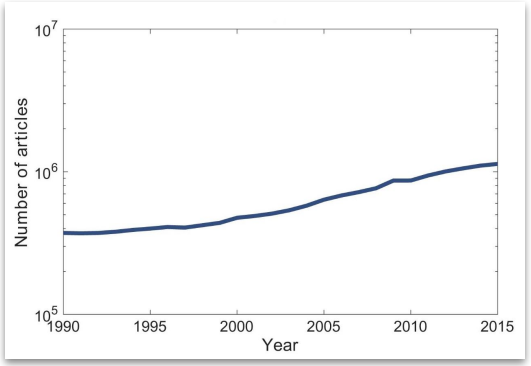
Argument Mining in Scientific Reviews

Team **five**

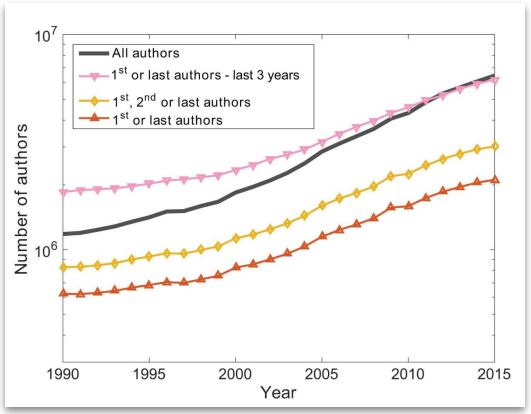
Project Supervisor: Michael Fromm
Lukas Dennert, Ruoxia Qi, Siddharth Bhargava,
Sophia Selle, Yang Mao, Yao Zhang

Motivation

- 63.4 million hours of peer-reviews in 2015 [1]
- Example:
  - LMU had 51.606 students in 2018 [2]
  - ~1228½ hours of reviewing
  - Or review ~273 papers per year
- +3.5% articles every year [1]
- **Sustainable?**
- Solution: partial Automation ⇒ Argument Mining
  - Argument Recognition (argumentative vs. non-arg.)
  - Stance Detection (pos-arg vs. neg-arg.)

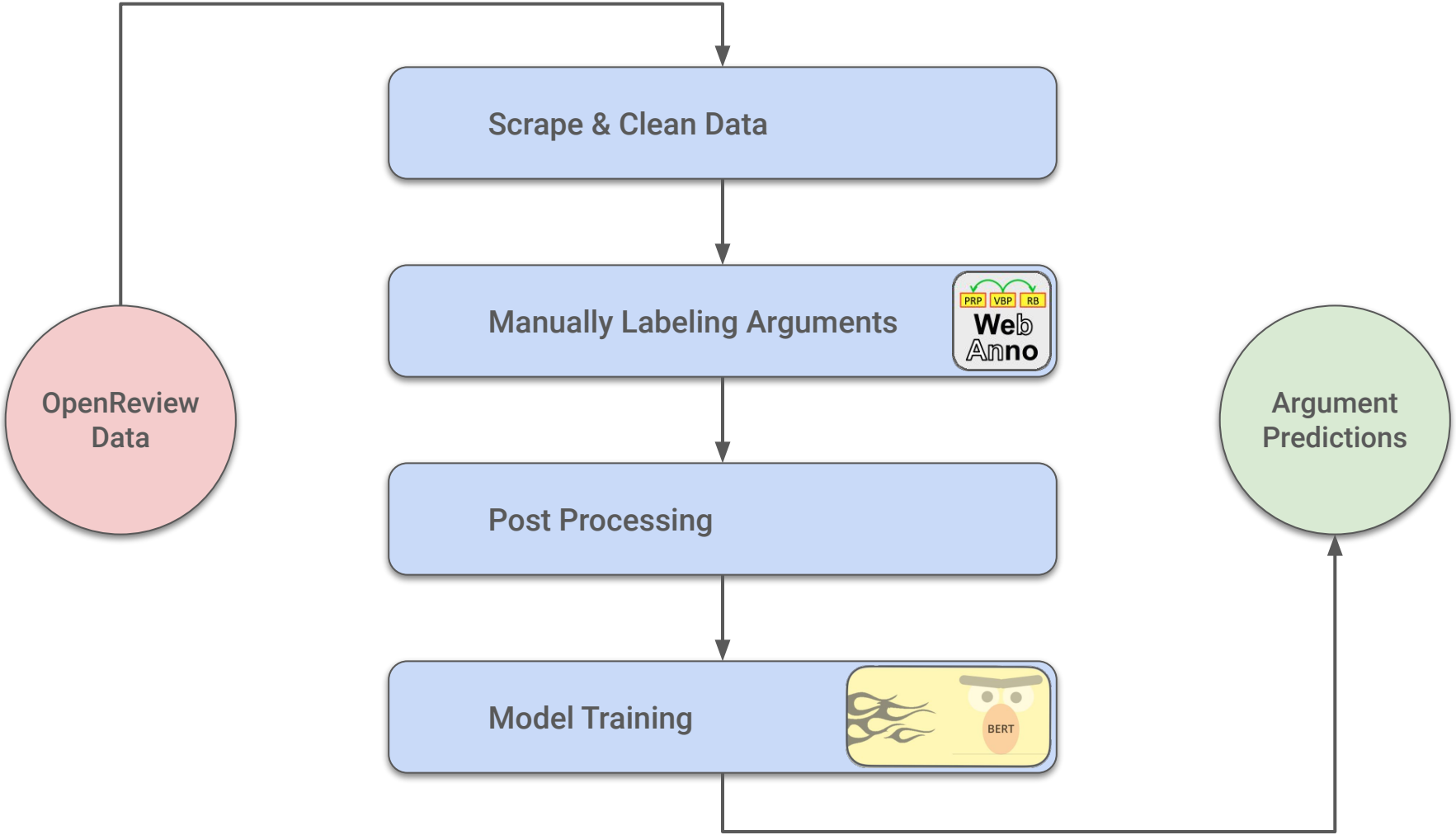


[1]



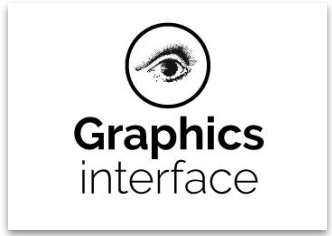
[1]

[1] M. Kovanis et al., The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise, 2016
[2] https://www.uni-muenchen.de/ueber_die_lmu/zahlen_fakten/index.html


Scrape & Clean Data
Manually Labeling Arguments
PRP
VBP
RB
Web
Anno
Post Processing
Model Training
BERT
OpenReview
Data
Argument
Predictions

1. Scrape Data

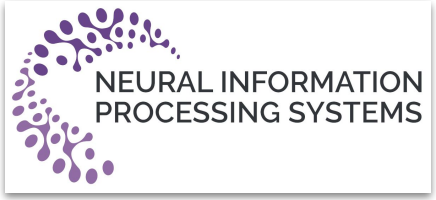
- OpenReview ⇒ better transparency of review process
- Downloaded with OpenReview API
- 6 Conferences, 12144 reviews total



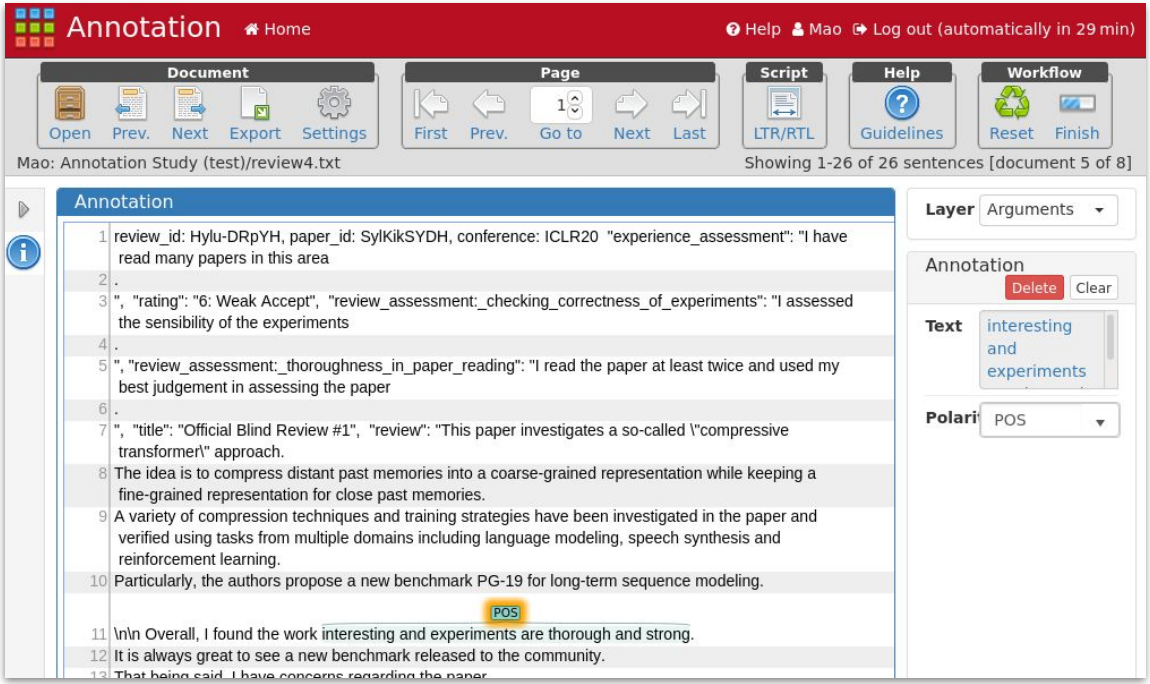
ICLR 2019 & 2020

2. Clean Data

- Remove non-words for BERT
- Universal mapping of ratings



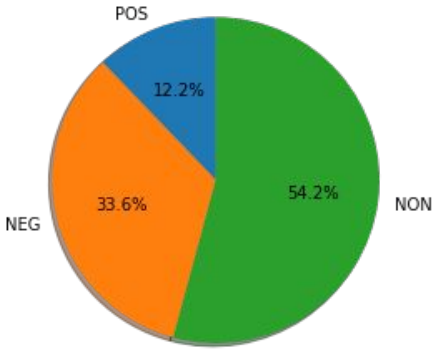
3. Annotation Study



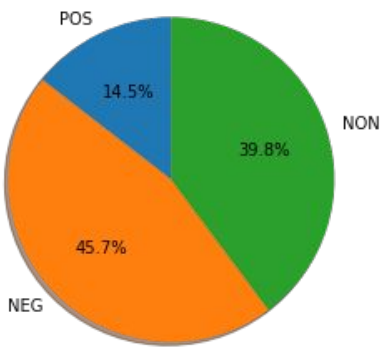
Screenshot of Webanno on annotation page

- Annotation Software: Webanno
- Labels: ‘POS’ and ‘NEG’

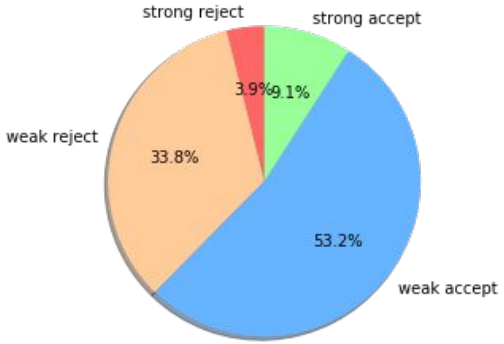
3. Summary of our 77 Annotated Reviews



Position of Segments

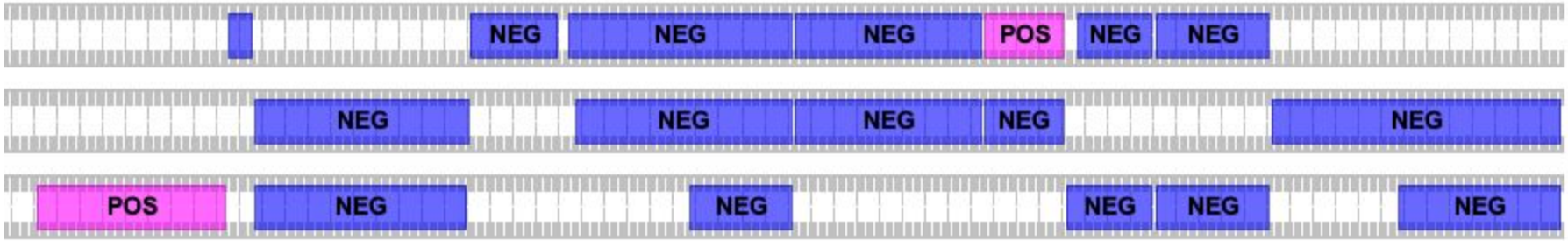


Position of Sentences



Rating of the Review

3. Inter Annotator Agreement



- Generally high agreement on POS/NEG, low agreement on Arg/Non-Arg
- No big difference when removing 1 annotator
- Krippendorf's Alpha: measure for the reliability of unitizing textual continua, i.e. annotate units without fixed boundaries.

4. Post-Processing: our new dataset “AMSR”

Input

	A	B	C	D	E	F	G	H
1		unit_id	rater	tag	token_start	token_end	offset_start	offset_end
2	0	1	Mao	POS	2-10	2-23	182	266
3	1	2	Mao	POS	3-1	3-68	267	592
4	2	3	Ruoxia	POS	2-10	2-22	182	265
5	3	4	Ruoxia	POS	3-4	3-68	282	592

Output

	A	B	
1	graph20_25_2_0	The submission presents evaluation of BendyPass,	('(0,128);', 'NA;')
2	graph20_25_2_1	The prototype is a simplified version of Bend Passw	('(0,108);', 'NA;')
3	graph20_25_2_2	The evaluation consisted of two sessions (taking pl	('(0,153);', 'NA;')
4	graph20_25_2_3	The experiment compared BendyPass with standar	('(0,92);', 'NA;')
5	graph20_25_2_4	The results show that although it took longer for pa	('(0,182);', 'NA;')
6	graph20_25_2_5	This submission contributes new knowledge about	('(0,103);', 'POS;')
7	graph20_25_2_6	The main strength of the paper is the experimental	('(0,105);', 'POS;')
8	graph20_25_2_7	It is particularly important to evaluate technology wi	('(0,76);', 'POS;')
9	graph20_25_2_8	The paper is well written: the work is motivated well	('(0,25);(25,2);(27,196);', 'POS;NA;POS;')
10	graph20_25_2_9	However, there are two main weaknesses: 1) the su	('(0,9);(9,151);', 'NA;NEG;')
11	graph20_25_2_10	The paper never justifies why Bend Passwords [33]	('(0,113);', 'NEG;')

	A	B	C
1	graph20_25_2_0	The submission presents evaluation of BendyPass,	NA
2	graph20_25_2_1	The prototype is a simplified version of Bend Passw	NA
3	graph20_25_2_2	The evaluation consisted of two sessions (taking pla	NA
4	graph20_25_2_3	The experiment compared BendyPass with standarc	NA
5	graph20_25_2_4	The results show that although it took longer for part	NA
6	graph20_25_2_5	This submission contributes new knowledge about h	POS
7	graph20_25_2_6	The main strength of the paper is the experimental u	POS
8	graph20_25_2_7	It is particularly important to evaluate technology wit	POS
9	graph20_25_2_8	The paper is well written: the work is motivated well,	POS
10	graph20_25_2_9	However, there are two main weaknesses: 1) the su	NEG

5. Experimental Setup (Datasets)

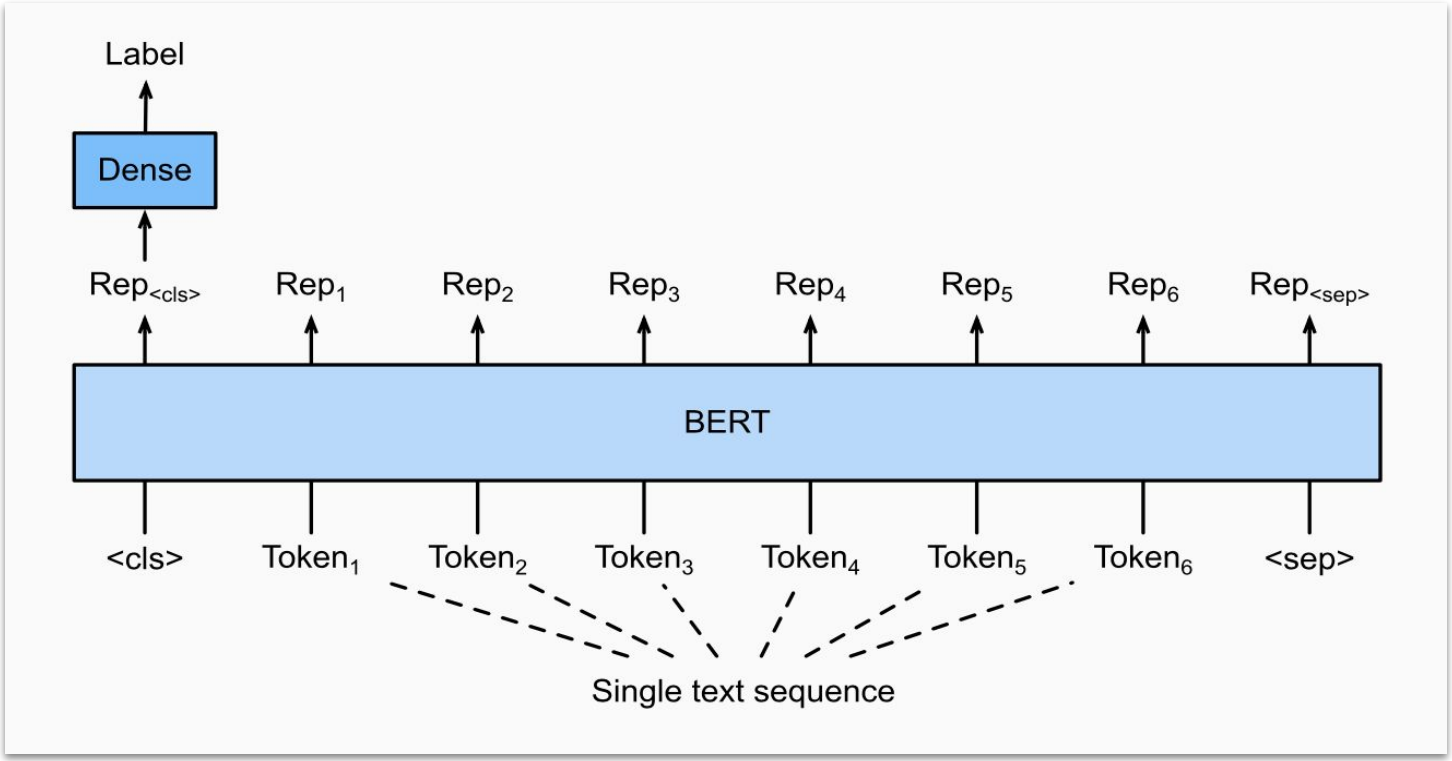
		Pos	Neg	NonArg	SUM
AMSR	Token-level	3259 (11%)	10559 (37%)	14689 (51%)	28507
	Segment-level	257 (12%)	711(33%)	1145 (54%)	2113
AURC	Token-level	36902 (20%)	35116 (19%)	109908 (60%)	181926
	Segment-level	2190 (16%)	2072 (15%)	9522 (69%)	13784

Token: This paper should be rejected, because the research question is not clearly articulated.

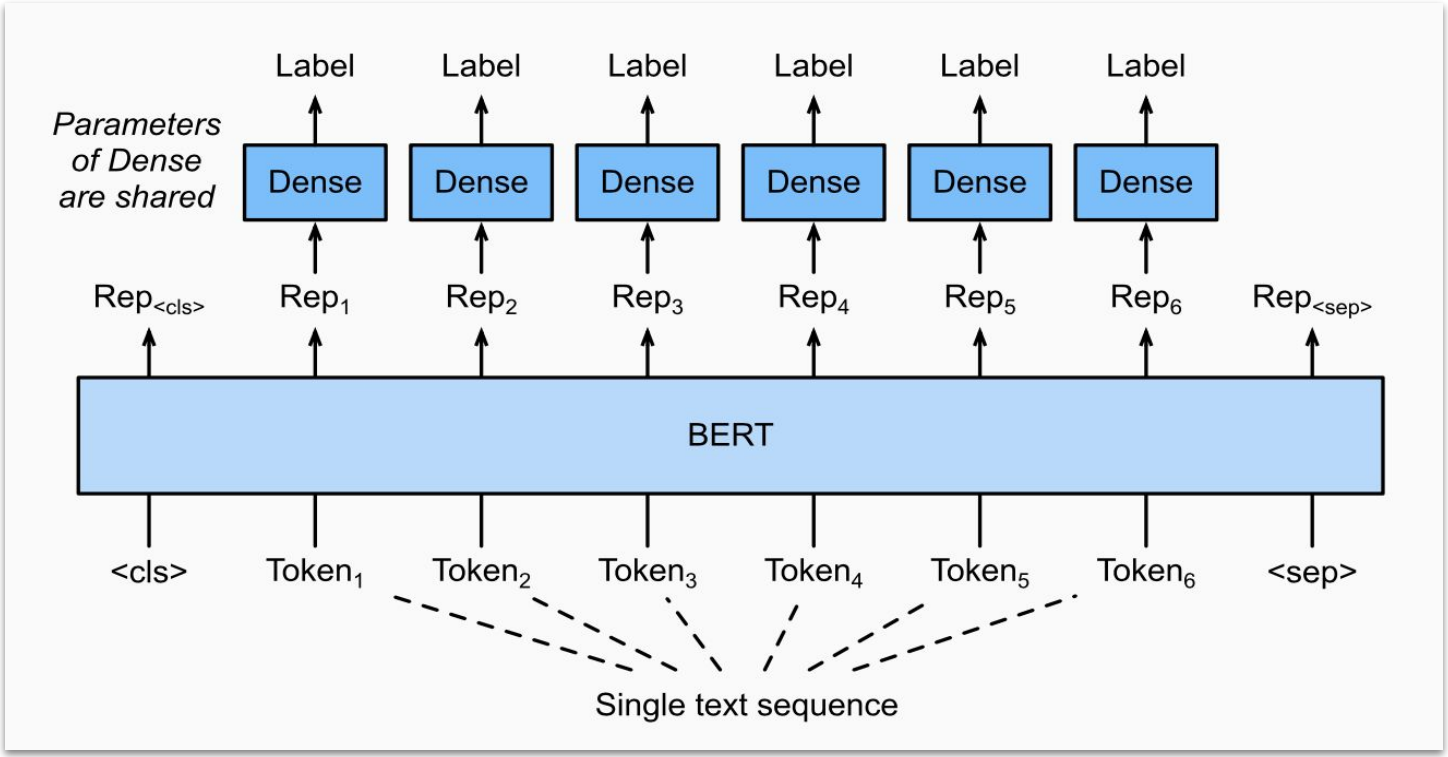
Segment: This paper should be rejected, because the research question is not clearly articulated.

Sentence: This paper should be rejected, because the research question is not clearly articulated.

5. Experimental Setup (Sentence-Level)



5. Experimental Setup (Token-Level)



5. Experimental Setup (Tasks)

	Description (Classify token/sentence into...)
Recognition '**Recog**'	**2** classes: **Arg** **vs.** **NonArg**
Stance Detection '**Stance**'	**2** classes: **Pos** **vs.** **Neg**
Classification '**Classify**'	**3** classes: **Pos** **vs.** **Neg** **vs.** **NonArg**

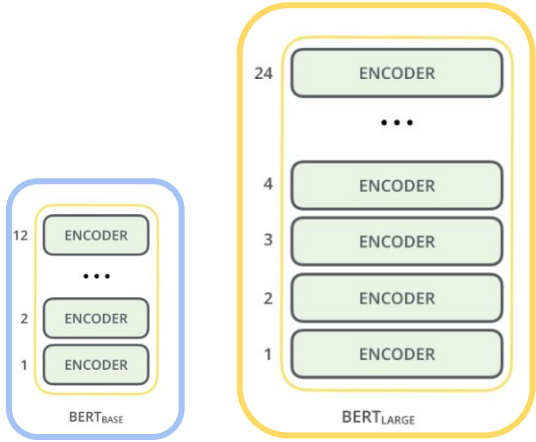
"The platform was nicely designed."

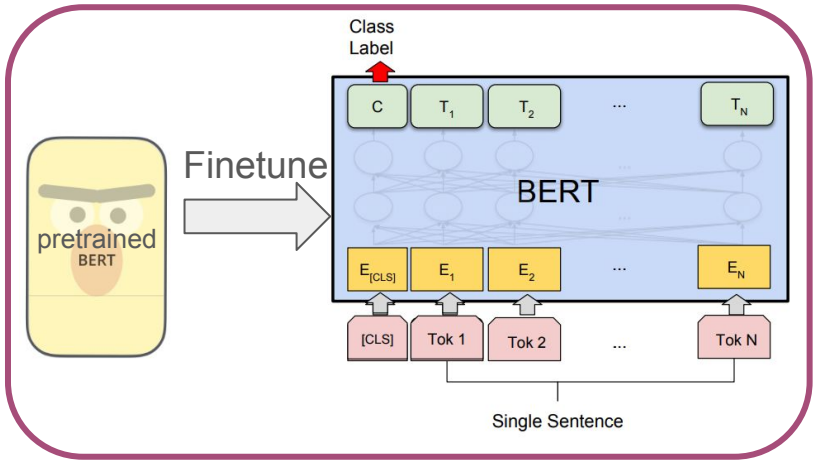
"The research question is not clearly articulated."

"Did they record data during the studies?"

5. Experimental Setup (Methods)

1. Majority Baseline
2. $Bert_{base}$ -> finetune on **AURC**
3. $Bert_{base}$ -> finetune on **AURC** -> finetune on **AMSR**
4. $Bert_{base}$ -> finetune on **AMSR**
5. $Bert_{large}$ -> finetune on **AMSR**
6. Human performance





6. Results

Recognition Model without Topic Information:

Ground Truth:

While the submission is hard to read in some places and some details about the system and study are missing, I think it is above the bar and should be accepted.

Sentence Level Prediction:

While the submission is hard to read in some places and some details about the system and study are missing, I think it is above the bar and should be accepted.

Token Level Prediction:

While the submission is hard to read in some places and some details about the system and study are missing, I think it is above the bar and should be accepted.

6.1 Based on Training Method

	Recognition		Stance Detection		Classification	
	sentence-setup[1]	token-setup[2]	sentence-setup[1]	token-setup[2]	sentence-setup[1]	token-setup[2]
1. Majority Baseline	0.351	0.35	0.423	0.434	0.234	0.233
2. Bert_BASE -> finetune on AURC	0.316 (T: 0.308)	0.353 (T: 0.355)	0.719 (T: 0.735)	0.644 (T: 0.627)	0.203	0.241 (T: 0.246)
3. Bert_BASE -> finetune on AURC -> finetune on AMSR	0.720 (T: 0.707)	0.877 (T: 0.878)	0.858 (T: 0.846)	0.862 (T: 0.868)	**0.700 (T: 0.660)**	0.796 (T: 0.807)
4. Bert_BASE -> finetune on AMSR	0.730 (**T: 0.713**)	0.886 (T: 0.896)	0.890 (T: 0.868)	0.853 (T: 0.849)	0.698 (T: 0.517)	0.814 (T: 0.808)
5. Bert_LARGE -> finetune on AMSR	**0.755** (T: 0.702)	**0.890 (T: 0.900)**	**0.905 (T: 0.867)**	**0.942 (T: 0.930)**	0.678 (T: 0.554)	**0.831 (T: 0.839)**
6. Human Performance	0.885	0.873	0.978	0.98	0.881	0.86

Table: Evaluation Table showing the F1 Macro values for different training methods used for fine-tuning our models, averaged over 10 seeds.

[1]: F1 Macro calculated based on unweighted loss function
[2]: F1 Macro calculated based on weighted loss function
“T”: F1 Macro calculated based on model with with topic information incorporated

6.2 Based on Topic Information

Incorporating Topic Information:

[CLS] {sentence tokens} [SEP] {**topic info**}

Topic for AMSR chosen: “paper quality”

Inference:

- Token Level Setup
  ⇒ no effect with topic information

- Sentence Level Setup
  ⇒ slight reduction of performance

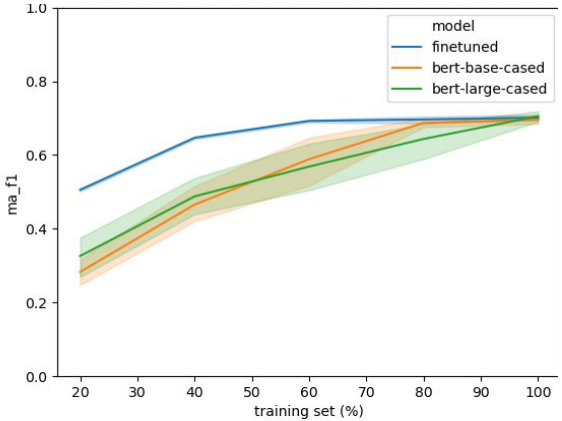
	Sentence-setup	Token-setup
Bert_BASE -> finetune on AURC	0.203	0.241 (**T: 0.246**)
Bert_BASE -> finetune on AURC -> finetune on AMSR	0.700 (**T: 0.660**)	0.796 (**T: 0.807**)
Bert_BASE -> finetune on AMSR	0.698 (**T: 0.517**)	0.814 (**T: 0.808**)
Bert_LARGE -> finetune on AMSR	0.678 (**T: 0.554**)	0.831 (**T: 0.839**)

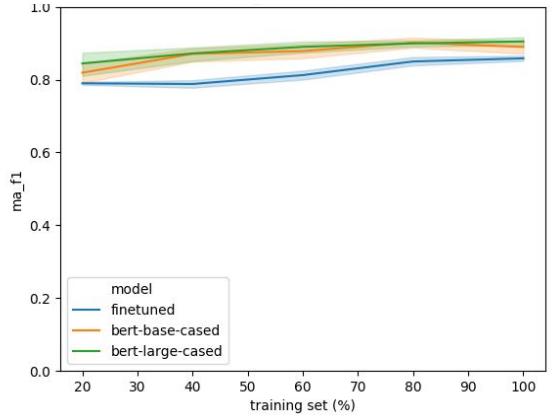
Table: F1 Macro values for Classification task without topic information and with topic information (in brackets).

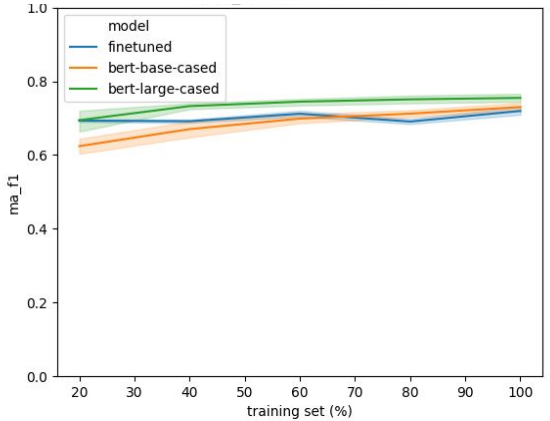
6.3 Based on Training Set Size

Classification | Stance Detection | Recognition

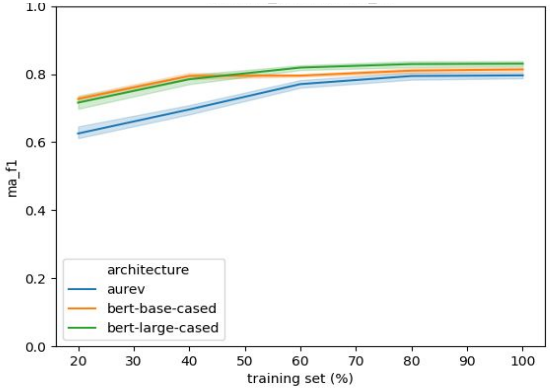
Sentence

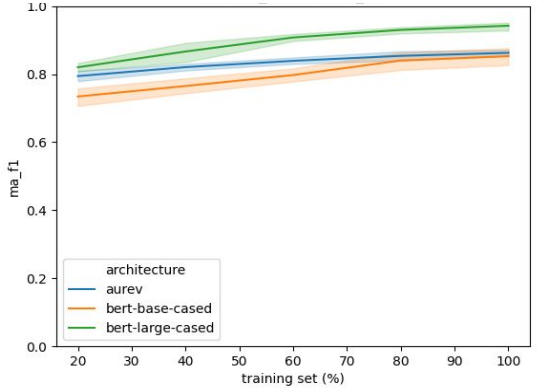


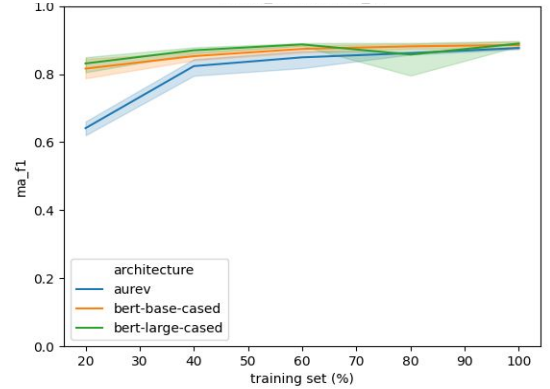




Token







Figures: Sentence/Token Level Setup without Topic 1) Classification 2) Stance Detection 3) Recognition

6.4 Multi-Task Experiment

Multi-Task Model

- Classification Task
- Without Topic Information
- Sentence Level

Results:

- F1 Macro for $Bert_{Base}$ on AURC and AMSR = 0.7
- F1 Macro for $Bert_{Base}$ on MTL and AMSR = 0.5

Datasets	Source	Topic
AURC	Fine-Grained Argument Unit Recognition and Classification	8 topics
CTAM	Cross-topic Argument Mining from Heterogeneous Sources	8 topics
CWAM	Corpus Wide Argument Mining - a Working Solution	213 topics
PASPE	Parsing Argumentation Structures in Persuasive Essays	79 topics
Debates	Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates	0 topic
AMSR	OpenReview.net	‘Paper Quality’

7. Application

Goal: Model correlation between arguments and review scores

Step 1: Confidence Prediction

- Apply best-performance recog model on all unlabeled reviews
- Got confidence score + sentence representation ([CLS])
- Sentence Level:

[CLS]	The results are not stellar, but certainly a worthy investigation.	0.999

- Token Level:

[CLS]	The	results	are	not	stellar	,	but	Certainly	a	worthy	Investigation	.
	0.81	0.95	0.90	0.69	0.96	0.03	0.19	0.78	0.94	0.94	0.78	0.04

7. Application (Future Work)

Token Level Score review_id: graph_20_1_1		
sen	avg_score	arg%
sen_3	0.68	0.75
sen_2	0.65	0.67
sen_9	0.45	0.52
sen_1	0.31	0.39
...	...	...
sen_n	0.01	0.02

(3 args expected)	l=0.4
reviews%	71.6%

	k=4
avg(arg%)	0.603

Step 2: Argument Extraction

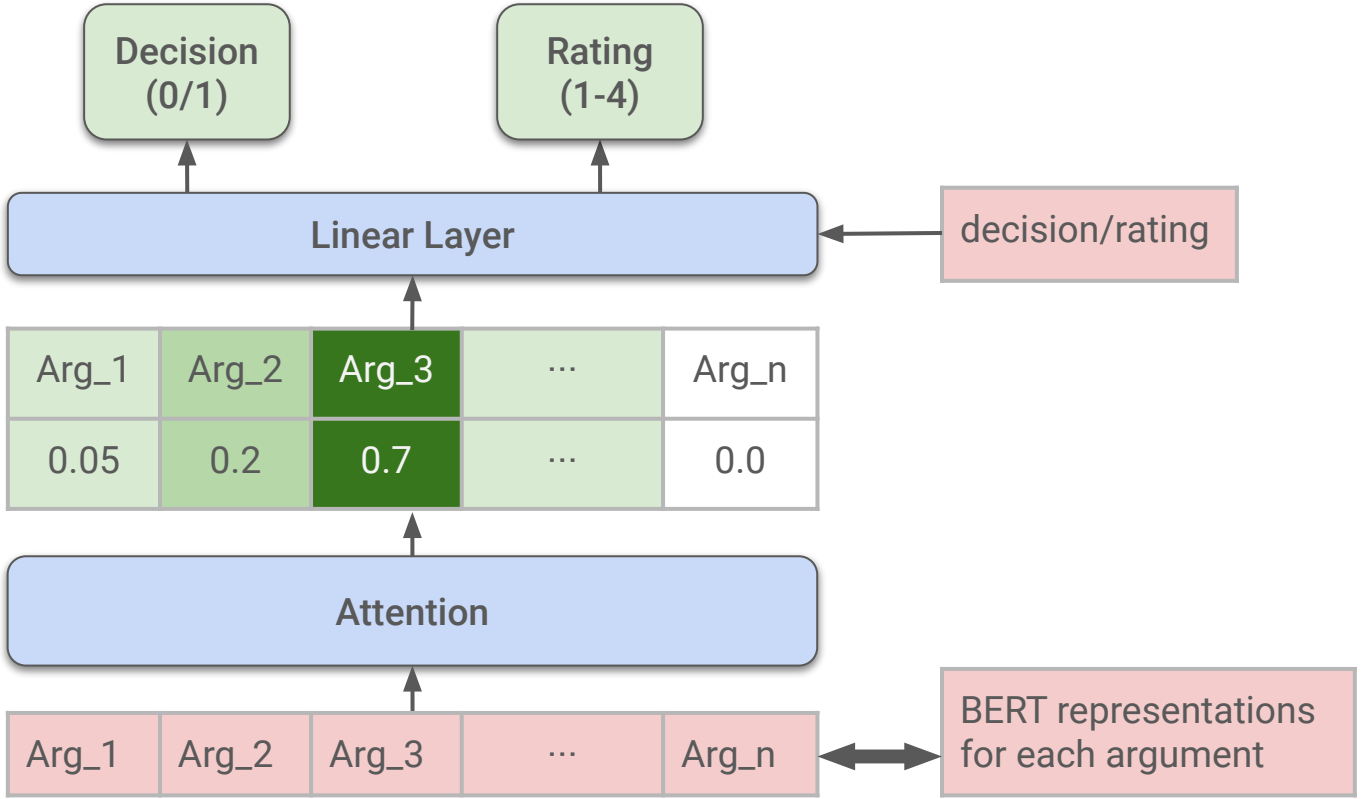
- Sentence Level: confidence
- Token Level:

$$\text{Avg\_score} = \text{avg}(\text{confidence of token})$$

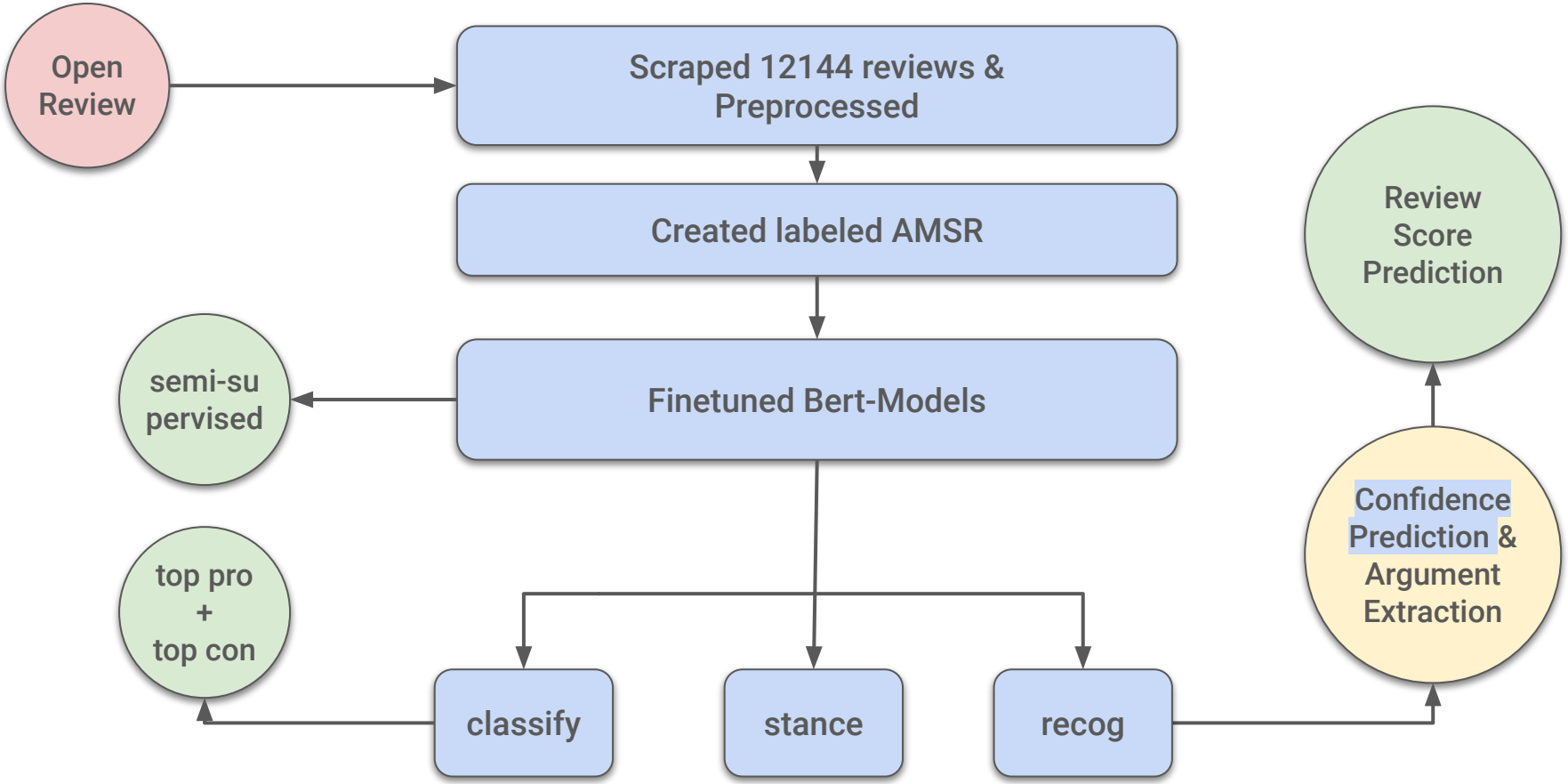
$$\text{Arg\_percentage} = \frac{|\text{arguments}|}{|\text{tokens}|_{\text{confidence(arg)}>0.}}$$

- Method 1: threshold (l), e.g. l=0.4
- Method 2: topK, e.g. k=4
- Select l and k based on statistical analysis (ongoing)

7. Model Training and Evaluation



8. Conclusion



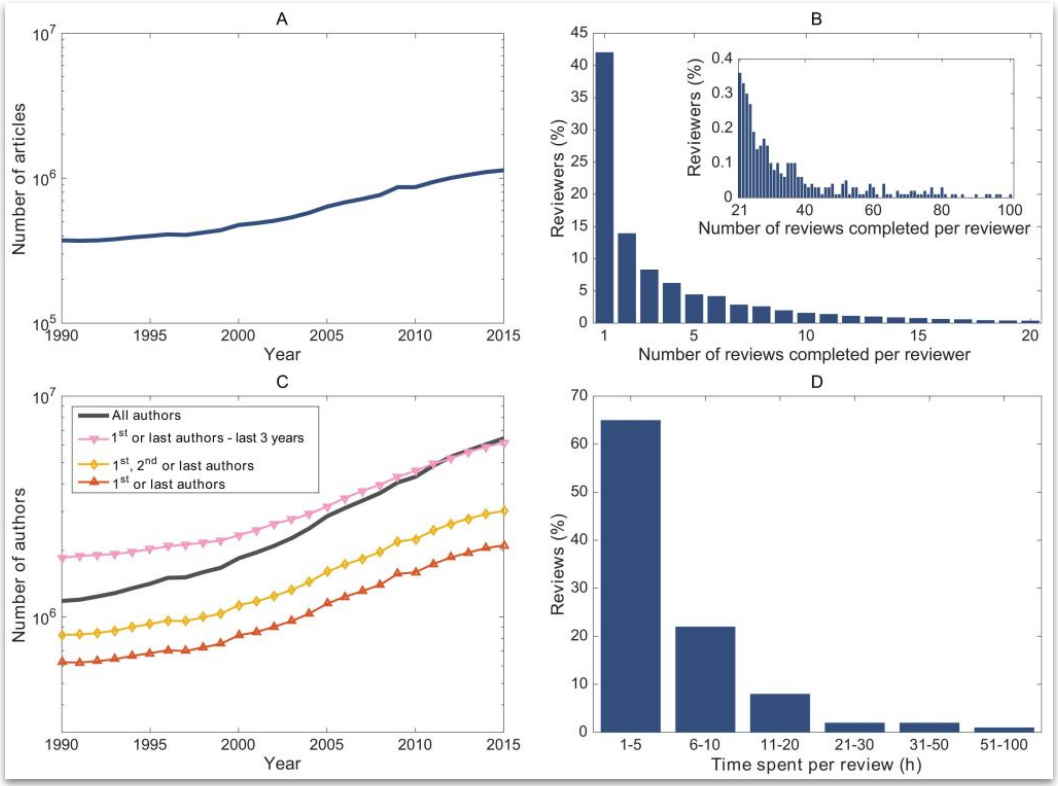
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thank you, for your attention.

Team **five**

Project Supervisor: Michael Fromm
Lukas Dennert, Ruoxia Qi, Siddharth Bhargava,
Sophia Selle, Yang Mao, Yao Zhang

Backup Slides: Statistics about Reviews



M. Kovanis et al., The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise, 2016

Backup Slides: Mapping of Ratings

1.3. Ratings Definition

The following observations have been made regarding the ratings format for each of the chosen conference,

The ratings scheme can be seen using the command: `conference_name["rating"].unique()`

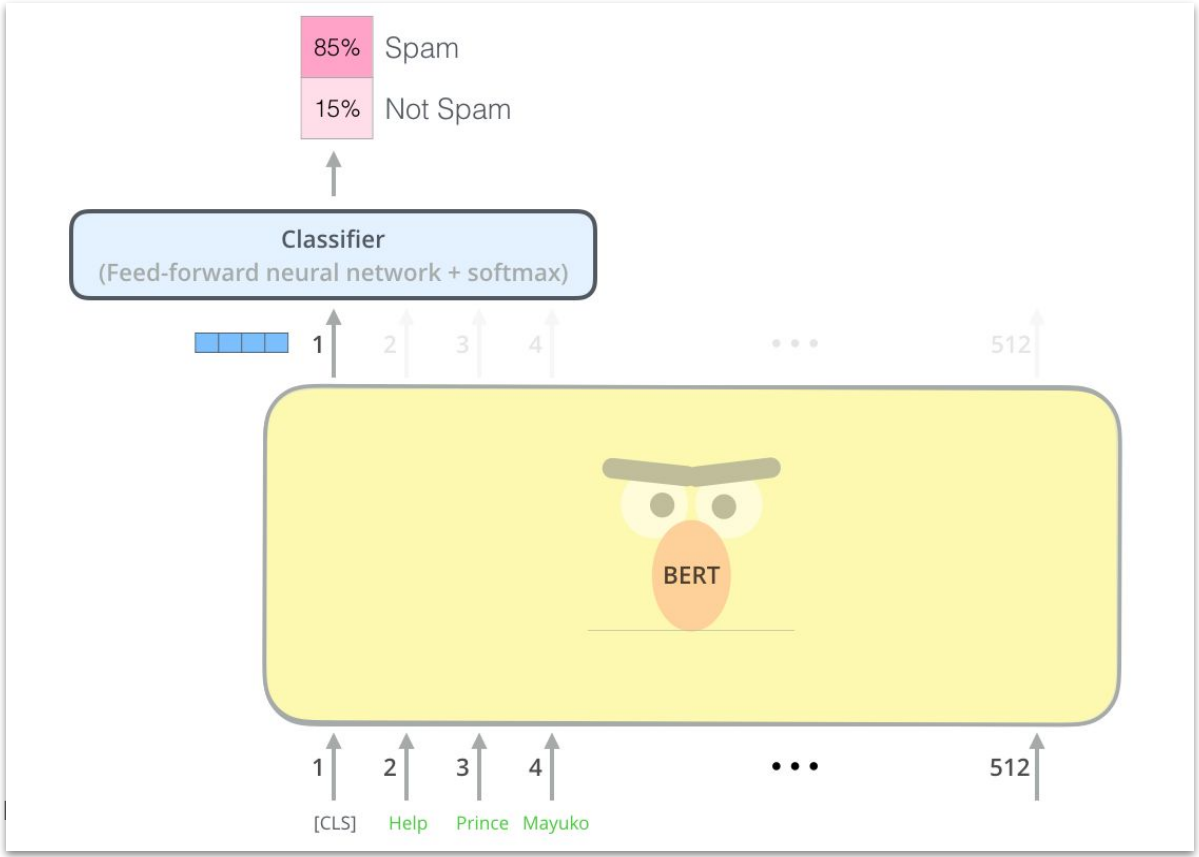
- ICLR19 follows a ratings distribution ranging from 1 - 10 with "1" meaning "trivial or Wrong" and "10" meaning "Top 5% of accepted papers, seminal paper".
- ICLR20 follows a ratings distribution ranging from {1,3,6,8} with "1" meaning "Reject" and "8" meaning "Accept".
- MIDL19 and MIDL20 follow a ratings distribution ranging from 1 - 4 with "1" meaning "Strong Reject" and "4" meaning "Strong Accept".
- NeuroAI19 follows a ratings distribution ranging from 1 - 5 with "1" meaning "Very Poor" and "5" meaning "Excellent".
- Graphics20 follows a ratings distribution ranging from 2 - 9 with "2" meaning " Strong rejection" and "9" meaning "Top 15% of accepted papers, strong accept".

To ensure uniformity in the data, we propose to convert the above ratings distribution into a uniform distribution ranging from 1 - 4, with "1" meaning "Strong Rejection" and "4" meaning "Strong Acceptance", as described in the table below.

Rating	Proposed Meaning	Raw Meaning
1	Strong Reject	Trivial or wrong / Strong Rejection / Clear Rejection / Very Poor
2	Weak Reject	Ok but not good enough - rejection / Marginally below acceptance threshold / Poor
3	Weak Accept	Marginally above acceptance threshold / Good paper, accept
4	Strong Accept	Top 50% of accepted papers, clear accept / Top 15% of accepted papers, strong accept / Top 5% of accepted papers, seminal paper / Very good / Excellent

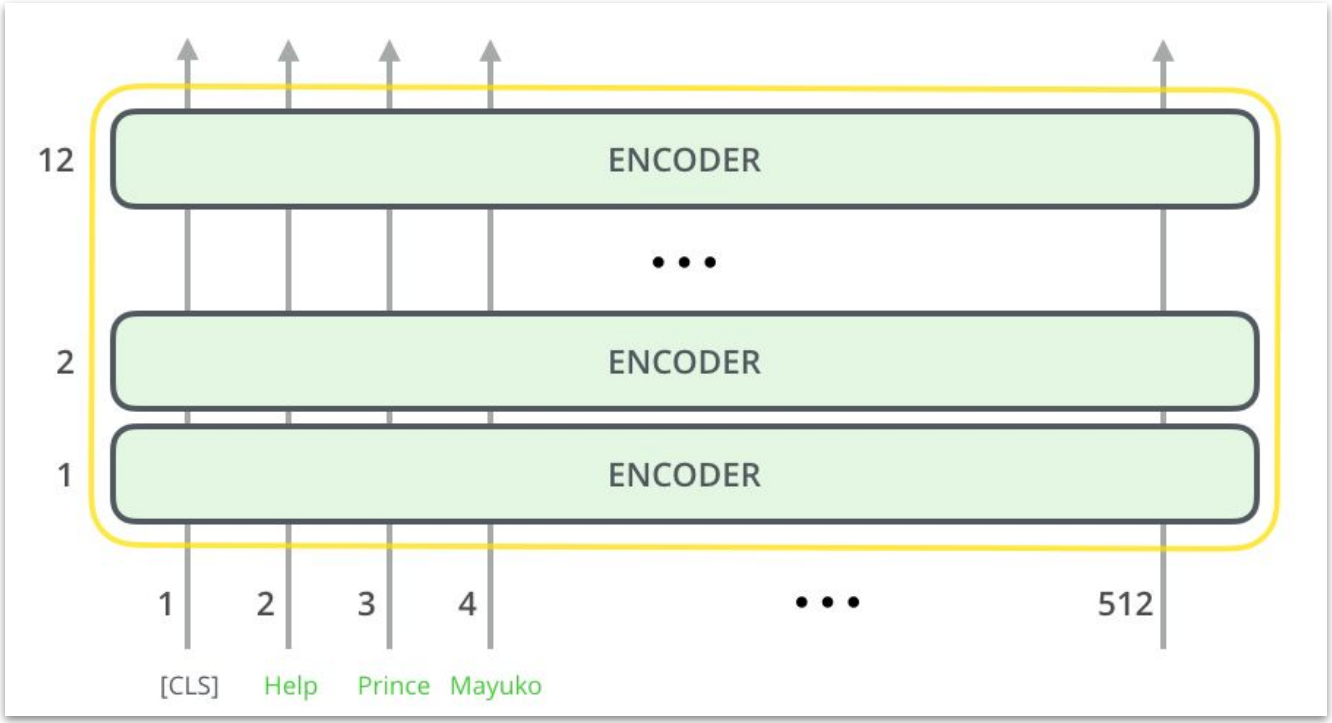
Table 2: Proposed Ratings System

Backup Slides: BERT + Classification Layer



Source: http://jalammar.github.io/illustrated-bert/

Backup Slides: BERT Structure (Encoders)



Source: http://jalammar.github.io/illustrated-bert/

Backup Slides: Krippendorff Alphas

With reference to the coincidences defined by (1) and (4) and following (2), the *agreement coefficient* ${}_u\alpha$ can take advantage of its limitation to the nominal metric and becomes defined by:

$$ {}_u\alpha_{\text{nominal}} = 1 - \frac{{}_uD_o}{{}_uD_e} = 1 - \frac{\frac{1}{\ell_{..}}\sum_{c=\phi}^{v}\sum_{k=\phi}^{v}\ell_{ck\ \text{nominal}}\delta^2_{ck}}{\frac{1}{\ell_{..}}\sum_{c=\phi}^{v}\sum_{k=\phi}^{v}\varepsilon_{ck\ \text{nominal}}\delta^2_{ck}} = 1 - \frac{\ell_{..} - \sum_{c=\phi}^{v}\ell_{cc}}{\ell_{..} - \sum_{c=\phi}^{v}\varepsilon_{cc}} \qquad (5a) $$

Replacing the reference to the expected coincidences ε_{ck} in the last expression of (5a) with its definition (4) yields another form of ${}_u\alpha_{\text{nominal}}$:

$$ {}_u\alpha_{\text{nominal}} = 1 - \frac{{}_uD_o}{{}_uD_e} = 1 - \left(\ell_{..} - \frac{1}{\ell_{..}}\sum_{i}^{m}\sum_{g}\left\{ \begin{matrix} L\left(S_{ig\ \text{valued}=\phi}\right) \\ \left(L\left(S_{ig\ \text{valued}\neq\phi}\right)\right)^2 \end{matrix} \right. \right) \frac{\ell_{..} - \sum_{c=\phi}^{v}\ell_{cc}}{\ell^2{}_{..} - \sum_{c=\phi}^{v}\ell^2_{c.}} \qquad (5b) $$

For our numerical example in Figure 1, which was constructed to highlight features that are easily overlooked when evaluating complex unitizations, the observed coincidences (1) and expected coincidences (4) are found in Figure 4:

Klaus Krippendorff et al., On the reliability of unitizing textual continua: Further developments, 2015
Source: https://link.springer.com/article/10.1007/s11135-015-0266-1

Backup Slides: Krippendorf's Alpha

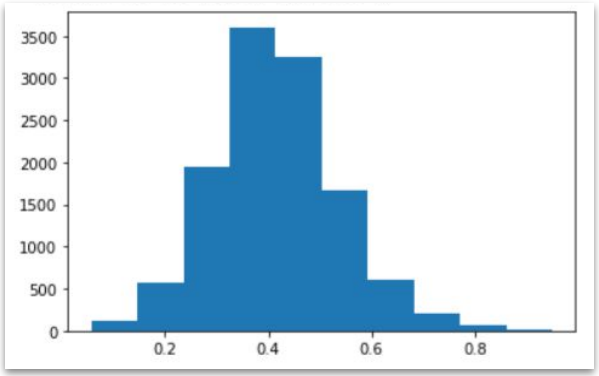
- General scores: uAlpha = 0.579 (NON/POS/NEG), biAlpha = 0.537 (NON-AU/AU), cuAlpha = 0.861 (POS/NEG)
- Verify quality of annotations per rater: uAlpha_leave_one in the range of 0.531-0.611
- Assign same weight to gaps: cuAlphaNON = 0.596
- Reduce the impact of skewed distribution:
  - uAlpha per tag: uAlphaPOS = 0.669, uAlphaNEG = 0.6
  - Merged Alpha: uAlpha = cuAlphaNON = 0.568, biAlpha = 0.521, cuAlpha = 0.903

Backup Slides: Krippendorf's Alpha

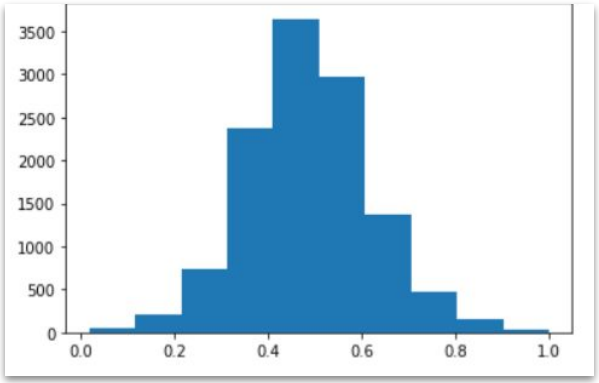
- uAlpha applies to all segments, including annotated units and gaps between them. Since we regard non-annotated parts as non-argumentative (sub)sentences, which will also be used as input of our model, it is necessary to take them into consideration.
- biAlpha measures reliability of a binary distinction between annotated unites taken together and gaps.
- cuAlpha focuses only on annotated units (ignoring gaps). It indicates the level of confidence in annotating argumentative (sub)sentences.
- Since non-annotated parts might be as important as annotated ones, we remove gaps with offset length < 3 (e.g. single punctuation/stop words...), assign 'NON' to remaining gaps, and calculate cuAlpha among three labels (cuAlphaNON).

Backup Slides: EDA at Token Level

mean(avg_score) = 0.415



mean(arg%) = 0.486



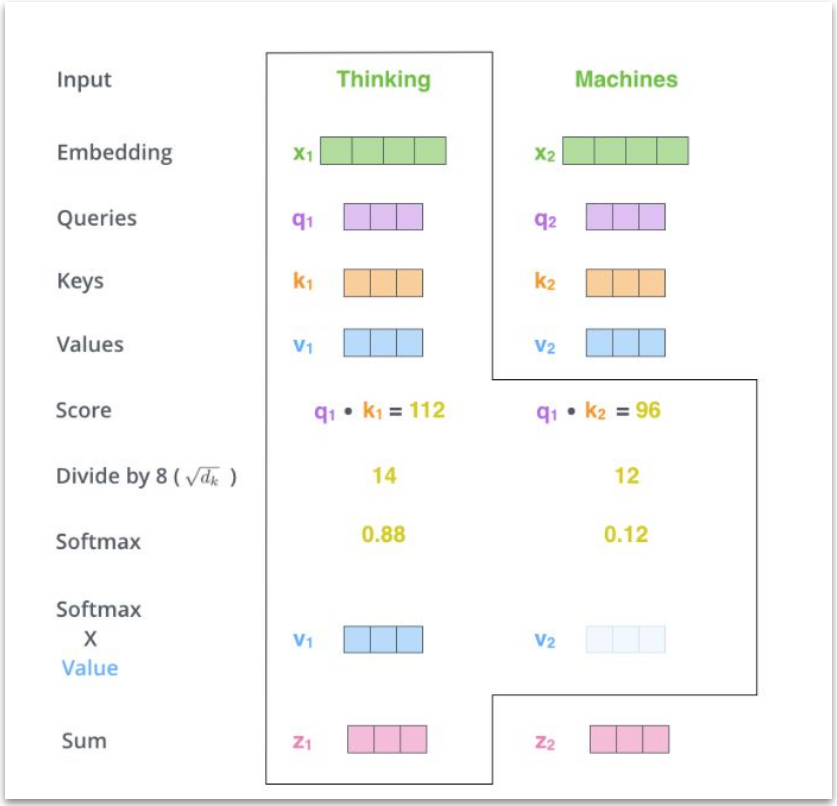
Different thresholds with 3 arguments per review

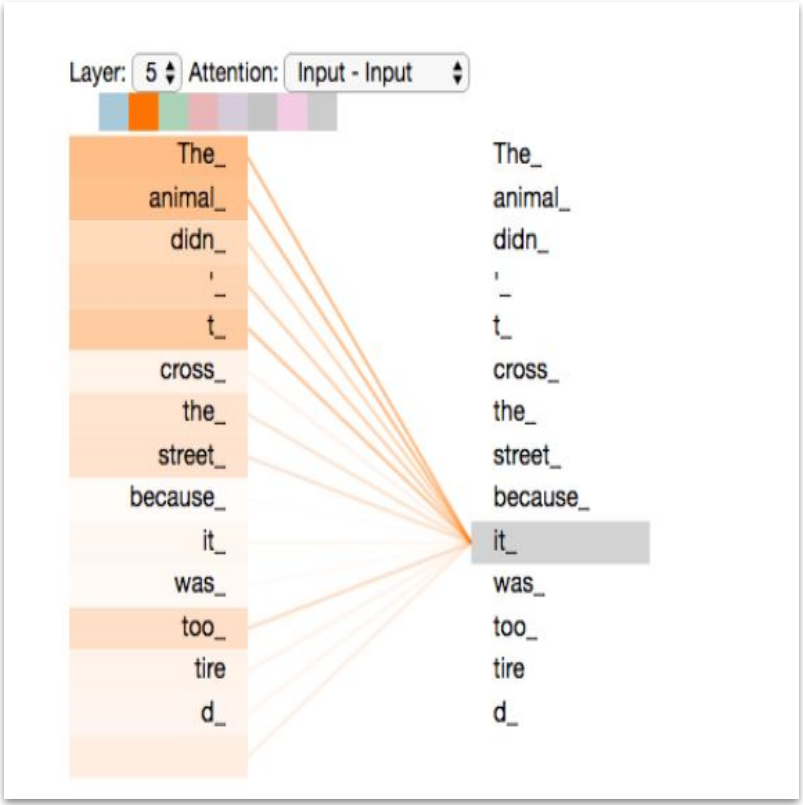
arg%	0.4	0.5	0.6	0.7
reviews%	71.6%	63%	50.6%	34.7%

Different topKs

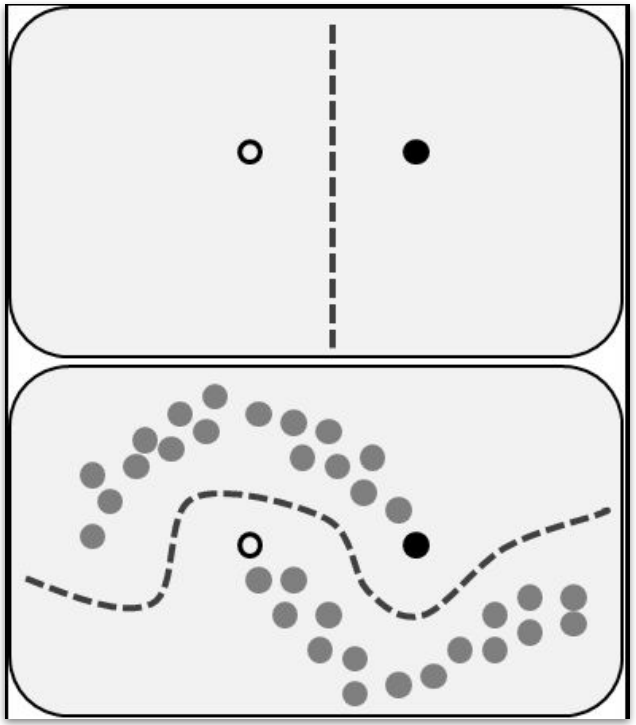
k=3		k=4	
avg(avg_score)	avg(arg%)	avg(avg_score)	avg(arg%)
0.637	0.665	0.580	0.603

Backup Slides: Attention Layer (Example)





Backup Slides: Semi-supervised Learning



Small amount labeled data +

Large amount unlabeled data

-> classification + clustering

-> improvement of performance