UNIVERSITÄT PASSAU

CHAIR OF DISTRIBUTED INFORMATION SYSTEMS

Argument Quality Assessment Over Textual Data

Presented by: Dorra El Mekki
Supervisors: Alaa Alhamzeh & Prof. Dr. Harald Kosch

December 21, 2022

1 Motivation

2[nd] *Contribution:* Investigate model robustness [Alh+22a]. → **Argument Mining**

1[st] *Contribution:* design a model to predict the **Strength** AQ dimension using the FinArgQuality corpus. → **Argument Quality Assessment**

Earnings Conference Call

↓

Argument Mining

↓

Argumentative Structure

↓

Argument Quality Assessment

↓

Argumentation Quality

↓

Prediction

↓

Analyst Post-Call Recommendation

Figure : The steps for the prediction of analyst post-call recommendation.

2 Introduction

Argument Definition

An argument consists of one claim (i.e., conclusion) supported or attacked by at least one premise (i.e., evidence) [DS19].

3

Argument Quality

1st Contribution

Submitted paper: Is it a Reliable Answer? Quality Assessment of Managers' Arguments in Earnings Conference Calls

Earnings Conference Call

↓

Argument Mining

↓

Argumentative Structure

↓

Argument Quality Assessment

↓

Argumentation Quality

↓

Prediction

↓

Analyst Post-Call Recommendation

Argument Quality Assessment

- Rate arguments according to their quality.
- Quality is determined by the **arguments' components** and the **relation** between them.

- Level of Granularity

Argument unit level [SKW21].

Argument level [Gre+20; SG17b].

Argumentation level [Wal+06].

Debate level [Coh+21; Eem+04].

- Approaches

Pair-wise approach [SKW21].

Point-wise approach [Wac+17b; Gre+20; SG17b].

Existing Corpora

Table : Datasets for argument quality.

Dataset	Source	Size	Approach	Quality dimensions
SwanRank [SEW15]	Internet Argument	5.3K arguments	Point-wise	Context Inference
UKPConvArgRank [HG16]	Online debate forum	16K pairs of arguments	Pair-wise	Convincingness
dagstuhl-15512- argquality [Wac+17b]	Online debate forum	320 arguments	Point-wise	15 dimensions [Wac+17b]
IBM Debater® - IBMArgQ-Rank-30kArgs [Gre+20]	Arguments from crowds	30K arguments	Point-wise	15 dimensions [Wac+17b]

FinArgQuality Corpus

- A point-wise annotated corpus, in the **argument** and **argument unit** levels of granularity.
- 5 Quality dimensions:
  - **Strong**, Persuasive, Specific, Objective, TemporalHistory.
- 2184 arguments
  - 2184 claims
  - 4899 Premises

Computational Argument Quality Assessment

Table : Argument Quality Assessment.

Paper	Argument Quality Assessment Dimension
[Gre+20]	Overall quality
[PN15]	Strength
[HG16]	Convincingness
[WSA17]	Relevance
[PN13]	Clarity
[Lau+20]	Cogency, Reasonableness, Effectiveness

Challenges

- **Main Goal:** Design a model to assess argument **Strength** based on FinArgQuality corpus level of granularity.

- **Challenge 1:** Dataset is imbalanced.
- **Challenge 2:** How to incorporate categorical features to improve the prediction of the strength dimension.

Proposed Approach

- Predict the strength score of arguments.
- Multi-class classification problem.

- Bert model: the **state-of-the-art** in the argument quality assessment [Gre+20; GAW21; SKW21].
- XLNet outperforms Bert in 20 benchmark tasks [YAN+19].

- Use the iterative stratified sampling over 10-fold cross-validation [STV11; SK17].

Model	Execution time (3 epochs)	Macro-F1 score	
		Mean	SEM
Bert	50 mins	.48	+- .01
XLNet	150 mins	**.49**	**+- .00**

- No significant improvement.
- The running time of XLNet.
- → Select the **Bert** model.

Baseline

Bert Model

Proposed Approach

- Input = **[cl_text]** claim **[/cl_text] [pr_text]** premise_1 **[/pr_text]** … Premise_n **[/pr_text]**
- Output = 0, 1 or 2

Results & Discussion: Bert with special separator token

Table : Evaluation of the different examined models, on ***FinArgQuality****, where SEM stands for standard error of the mean*

Model	**Macro-F1 score**		**Accuracy**	
	Mean	**SEM**	**Mean**	**SEM**
Bert (Baseline)	.48	+- .01	.74	+- .01
Bert, special separator token	**.50**	+- .00	**.77**	**+- .01**

Proposed Approach

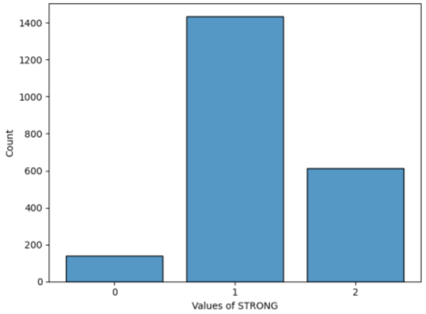
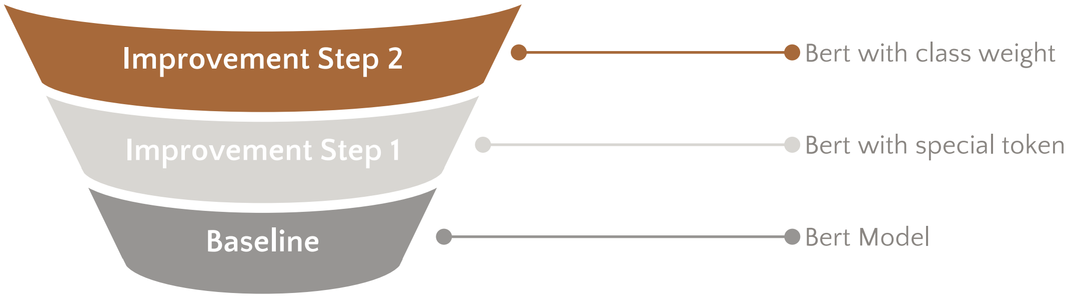


Figure: Distribution of the Strong dimension

Imbalanced classes → Macro-F1 score [LLS09].



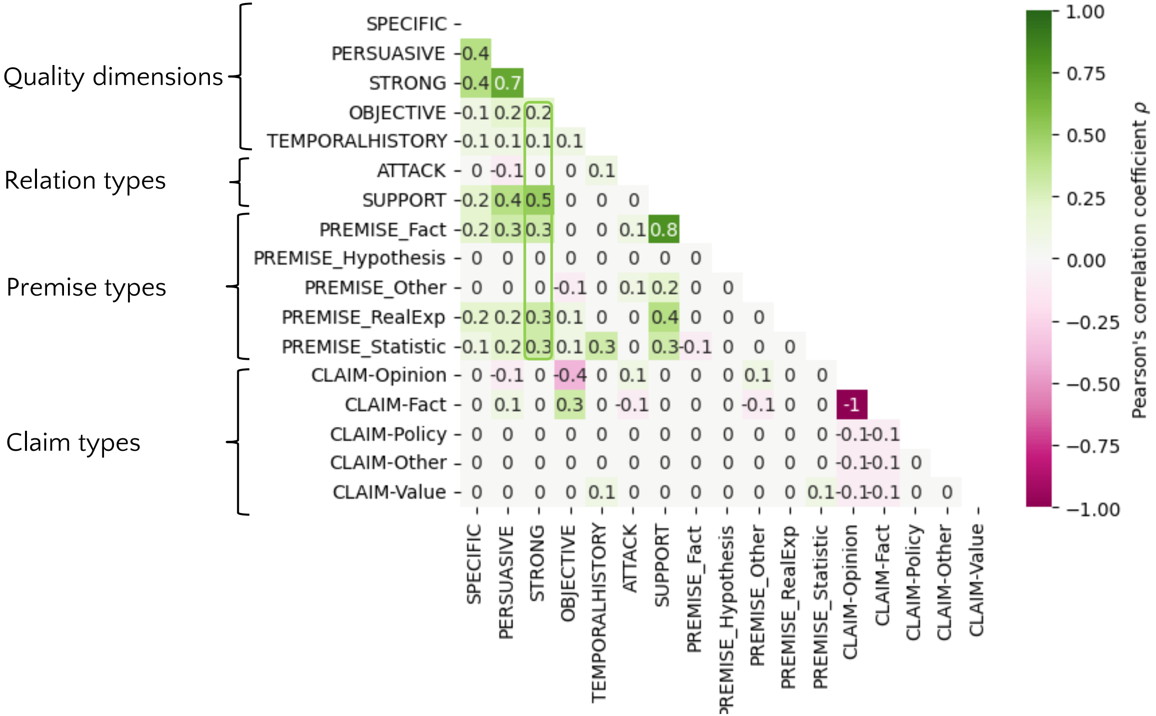
Results & Discussion: Bert with class weight

Table : Results of the macro-F1 score on the different class weights.

Class 0	Class 1	Class 2	macro-F1 score
5	1	1	0.46
6	1	1	0.50
7	1	1	0.53
9	1	1	0.52
10	1	1	0.48
11	1	1	0.48
5	1	2	0.50
6	1	2	0.48
7	**1**	**2**	**0.56**
8	**1**	**2**	**0.55**
9	1	2	0.54
10	1	2	0.48
11	1	2	0.54

- $p=0.15$ (> 0.05) for the weights [7,1,2]
- $p=0.01$ (< 0.05) for the weights [8,1,2]

→ **[8,1,2]**



Strength dimension:
An argument's strength is determined by the number and types of premises that support its claim.

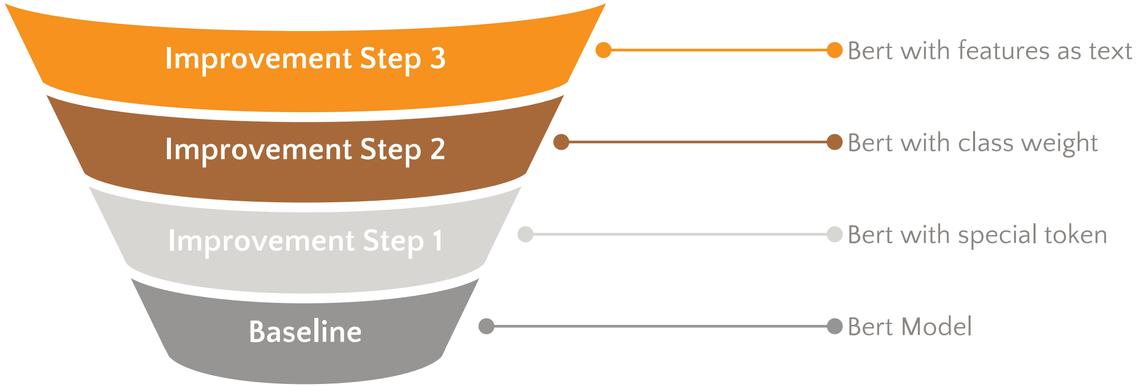
Figure: Correlation between argument quality dimensions

Why Bert with Features?

- For small datasets, when more features are incorporated into a model, simple models may **outperform** complex ones [PM20].
- For the Stance Detection task, Prakash and Madabushi highlight the advantage of including Count-Based features in the pre-trained models [PM20].

Proposed Approach

Input =
[cl_text] *It's a very rapidly expanding country.* [/cl_text]
[cl_type] *Claim-Opinion* [/cl_type]
[r_type] SUPPORT [/r_type]
[pr_text] *Constant currency growth was 48%.*[/pr_text]
[pr_type] *Premise-Statistic* [/pr_type]
Output = 1



Results & Discussion: Bert with features as text

Table: The results of Bert model with categorical features as text, where SEM stands for standard error of the mean

Feature Format	Included features			Metric	
	Claim type	**Premise type**	Relation type	Macro-F1	
				Mean	SEM
Baseline	x	x	x	.55	+- .02
Feature as test	✓	x	x	.50	+- .04
	x	**✓**	**x**	**.56**	**+- .01**
	x	x	✓	.54	+- .01
	✓	✓	x	.55	+- .02
	x	✓	✓	.54	+- .01
	✓	x	✓	.49	+- .04
	✓	✓	✓	.53	+- .02

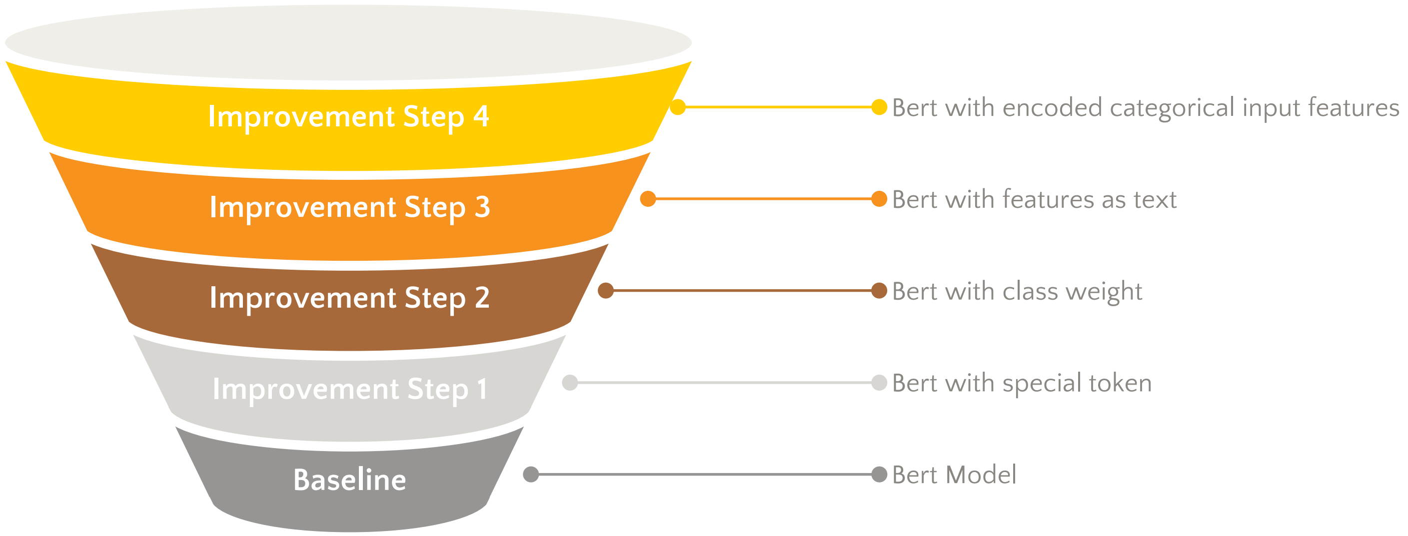
- Shapley values [SHA53].

Claim type: -3%

Relation type: -1.5%

Premise type: +2.5%

Proposed Approach



[CLS] is the special symbol for classification output [Dev+18].

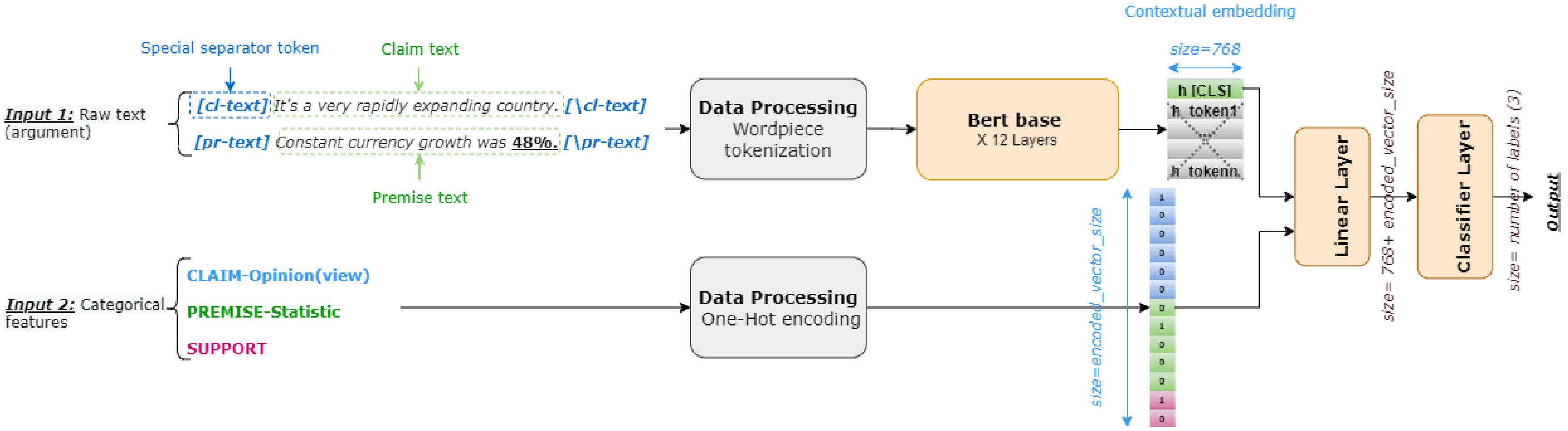


Figure: Model architecture for Bert with encoded categorical input features

Results & Discussion: Bert with Encoded Categorical Input Features

Table: The results of Bert model with categorical features as text, where SEM stands for standard error of the mean

Feature Format	Included features			Metric	
	Claim type	**Premise type**	Relation type	Macro-F1	
				Mean	SEM
Baseline	x	x	x	.55	+- .02
One-Hot Encoding	✓	x	x	.57	+- .01
	x	**✓**	**x**	**.59**	**+- .02**
	x	x	✓	.57	+- .01
	✓	✓	x	.58	+- .01
	x	✓	✓	.57	+- .01
	✓	x	✓	.57	+- .01
	✓	✓	✓	.55	+- .01

- Shapley values [SHA53].

Claim type: -0.17%

Relation type: -0.67%

Premise type: +0.83%

Results & Discussion

Table : Evaluation of the different examined models, on ***FinArgQuality****, where SEM stands for standard error of the mean*

Model	**Macro-F1 score**		**Accuracy**	
	Mean	**SEM**	**Mean**	**SEM**
Bert (Baseline)	.48	+- .01	.74	+- .01
Bert, special separator token	.50	+- .00	**.77**	**+- .01**
Bert, class weight	.55	+- .02	.67	+- .02
Bert, features as text	.56	+- .01	.71	+- .01
Bert, One-Hot Encoding	**.59**	**+- .02**	.71	+- .01

UNIVERSITÄT PASSAU

4 Model Robustness

2nd Contribution

Alhamzeh, Alaa, Előd Egyed-Zsigmond, Dorra El Mekki, Abderrazzak El Khayari, Jelena Mitrović, Lionel Brunie, and Harald Kosch. "Empirical Study of the Model Generalization for Argument Mining in Cross-Domain and Cross-Topic Settings." In Transactions on Large-Scale Data-and Knowledge-Centered Systems LII, pp. 103-126. **Springer**, Berlin, Heidelberg, 2022.

Earnings Conference Call

↓

Argument Mining

↓

Argumentative Structure

↓

Argument Quality Assessment

↓

Argumentation Quality

↓

Prediction

↓

Analyst Post-Call Recommendation

Challenges

- The lack of labelled data.
- The domain dependency performance of the existing models.

→ Investigate model robustness.

Research Goals

Baseline

Ensemble learning approach.

Task: Argument identification

Datasets: Student Essays and Web Discourse.

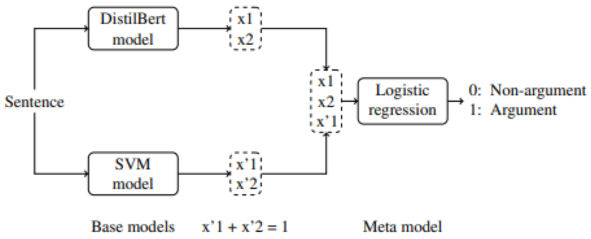


Figure: Stacked Model Architecture [Alh+21].

Research Goals

RG1: Extend the ensemble learning approach.

Tasks: argument identification and argument unit classification.

Datasets: Student Essays, Web Discourse and IBM corpus.

RG2: Conduct an empirical study of the model generalization in cross-domain and cross-topic settings.

RG3: Study the trade-off between the number of topics #T and the number of samples per topic #S/T to enhance the model generalization.

Model Generalization

	Paper	Task(s)	Algorithm(s)	Dataset(s)	Approach	Comments
1	Schiller et al. [SDG21]	Stance detection	Bert	10 heterogeneous datasets	SDL vs. MDL	-Different task -Same approach
2	Ajjour et al. [Ajj+17]	argument unit segmentation	NN model	3 corpora	Cross Domain	-Features on the token level. -Our features on the sentence level.
3	Bouslama et al. [BAA19].	extract the argument from the web and classify its components	CNN SVM Naïve Bayes	3 corpora	Cross Domain	-Different task -Same approach
4	Stab et al. [SMG18].	Argument identification with respect to the topic	LSTM	25,000 instances over eight topics	Cross Topic	-Different setup -We investigate the effect of diversity sampling.

Proposed Approach

- Extend the ensemble learning stacking approach proposed in [Alh+21] on the **argument unit classification (premise/claim)** and train it on the **IBM corpus** [Aha+14].
- Investigate the model robustness over **unseen data**.

Experimental Setup

Task	**Algorithms**	**Datasets**	**Epochs**	**Optimizer**	**Loss**	**Seeds**	**Train-test**
Argument identification	SVM DistilBert (Stacking approach)	Student Essays [SG14a] Web Discourse [HG17]	3	AdamW optimizer	Cross-Entropy	5	5-cross validation
Argument unit classification	SVM DistilBert (Stacking approach)	Student Essays [SG14a] Web Discourse [HG17] **IBM [Aha+14]**	3	AdamW optimizer	Cross-Entropy	5	5-cross validation

Corpora description

Table: Class distributions for all used datasets.

	Student Essays [SG14a]		**Web Discourse [HG17]**		**IBM [Aha+14]**		**Total**
	Count	**[%]**	**Count**	**[%]**	**Count**	**[%]**	
Argument	**5459**	**60**	1025	11	2683	29	9167
Non-Argument	**1358**	**77**	411	23	0	0	1769
Premise	**3510**	**62**	830	15	1291	23	5631
Claim	**1949**	**55**	195	6	1392	39	3536
Topic	**372**	**91**	6	1	33	8	411

Approach: SDL vs MDL

- For the SDL, a single dataset is used to train and test the model.
- For the MDL, we train on all datasets and test on the test set of one dataset.

Results & Discussion: SDL vs MDL

Table: SDL vs. MDL argument identification and argument unit classification using the stacked model.

		Argument identification		**Argument unit classification**	
		Macro-F1 score		**Macro-F1 score**	
		Mean	**Std**	**Mean**	**Std**
SDL	SE	**.864**	+- .004	**.808**	+- .004
	WD	**.715**	+- .020	**.803**	+- .016
	IBM	-	-	**.987**	+- .002
MDL	SE	.776	+- .007	.693	+- .141
	WD	.661	+- .024	.670	+- .035
	IBM	-	-	.894	+- .008

Discussion

- SDL gives a better performance, while MDL can improve the model robustness and stability over unseen data.
- Detecting argumentative text proved to be an intrinsically more generalized task than determining premises and claims.

Approach: Cross Domain

- Testing on completely unseen datasets.
- The model is trained on n – 1 corpora and tested on the remaining one.

Results & Discussion: Cross Domain

Table: Evaluation of the cross-domain argument identification task.

Training	Testing	Model	Macro-F1 score	
			Mean	**Std**
SE	**WD**	Stacked model	.436	+- .009
		DistilBert	**.571**	+- .005
		[Al +16]	.524	-
WD	**SE**	Stacked model	**.599**	+- .009
		DistilBert	.580	+- .015
		[Al +16]	.128	

Table: Evaluation of the cross-domain argument identification task.

Training	Testing	Model	Macro-F1 score	
			Mean	**Std**
SE,WD	**IBM**	Stacked model	**.554**	+- .079
		DistilBert	.469	+- .023
SE, IBM	**WD**	Stacked model	.455	+- .196
		DistilBert	**.602**	+- .012
WD, IBM	**SE**	Stacked model	**.526**	+- .060
		DistilBert	.366	+- .057

Discussion

- A drop in performance of the stacked model compared to in-domain settings (0.784 for argument identification and 0.869 for argument unit classification).
- For limited dataset, incorporating the features might play a crucial role to achieve reliable results.

Experimental Setup : Cross Topic

RQ: Given a fixed size of data, would it be better to include more topics with fewer sentences per topic or fewer topics with more sentences per topic?

$$\boldsymbol{N = \#T.\#S/T}$$

- $\boldsymbol{\#T}$ is the number of topics (variable).
- $\boldsymbol{\#S/T}$ is the number of sentences per topic (variable).
- **N** is the fixed size of data.

Fix N in a way that we can have multiple pairs of (#S/T, #T) satisfying the Equation, this implies that a higher #T leads to a lower #S/T.

UNIVERSITÄT PASSAU

Results: Cross Topic

Table: Cross-topic experiments for argument identification task.

#S/T	#T	Model	Macro-F1 score	
			Mean	Std
4	300	Stacked model	**.806**	+- .029
		DistilBert	.566	+- .095
6	200	Stacked model	.766	+- .017
		DistilBert	.487	+- .056
24	50	Stacked model	.660	+- .038
		DistilBert	.439	+- .012

Table: Cross-topic experiments for argument unit classification task.

#S/T	#T	Model	Macro-F1 score	
			Mean	Std
4	300	Stacked model	.804	+- .036
		DistilBert	.748	+- .031
6	200	Stacked model	.817	+- .020
		DistilBert	.779	+- .018
24	50	Stacked model	**.846**	+- .070
		DistilBert	.828	+- .091

#S/T: *number of Sentences/Topic*, **#T**: *number of Topics, Std: standard deviation*

Discussion: Cross Topic

- The effect of diversity sampling

Observation:

For argument identification: higher #T (train) → higher macro-F1

For argument unit classification: higher #T (train) → lower macro-F1

Discussion:

The argument structure may vary depending on the topic (law vs. finance-related topics).

The argument unit classification, separating premise from claim is determined by the grammatical structure of sentences, which is independent of the employed vocabulary or the topic (claim vs. premise keywords).

5 Conclusion

Conclusion: Argument Quality Assessment

- A linear relation between the **Specific**, **Persuasive**, and **Strong** quality dimensions proved by the highest inter-correlation.
- We identify a positive correlation between the **Strong** quality dimension and the **premise/relation types**.

- The **premise type** contributes positively to the prediction of the **Strength** quality dimension.
- The proposed model architecture incorporating encoded categorical features (premise type) improves the macro-F1 score by 11% over the Bert baseline model.
- The model can be adjusted to predict the **Specific** and **Persuasive** quality dimensions.

UNIVERSITÄT PASSAU

Conclusion: Model Robustness

- We extend the stacking approach from the argument identification to the argument unit classification task and we enlarge the size of the training set.
- We investigate the model robustness over unseen data: **SDL vs. MDL**, **Cross-domain** settings, **Cross-topic** settings.

- Despite the drop in performance compared to in-domain settings, the model is still able to generalize.
- Detecting argumentative text (**argument identification**) proved to be an intrinsically more generalized task than determining premises and claims (**argument unit classification**).
- For limited dataset, incorporating the features might play a crucial role to achieve reliable results.

Future Outlook

- Feature analysis study for the argument quality assessment model.
- Data valuation, removing outliers.
- A Leave-one-company-out experiment to study the similarities across data sources.
- End-to-end pipeline.

UNIVERSITÄT PASSAU

References (1)

[Aha+14] Ehud Aharoni et al. “A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics”. In: Proceedings of the first workshop on argumentation mining. 2014, pp. 64–68 (cit. on p. 28).

[Ajj+17] Yamen Ajjour et al. “Unit segmentation of argumentative texts”. In: Proceedings of the 4th Workshop on Argument Mining. 2017, pp. 118–128(cit. on pp. 1, 17, 21, 27, 61).

[Al +16] Khalid Al Khatib et al. “Cross-domain mining of argumentative text through distant supervision”. In: Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies. 2016, pp. 1395–1404 (cit. on pp. 17, 52, 53, 80).

[AML22] Alaa Alhamzeh, Dorra El Mekki, and Kürsad Lacin. “Is it a Reliable Answer? Quality Assessment of Managers’ Arguments in Earnings Conference Calls”. In: 2022 (cit. on pp. 21, 22).

[Alh+21] Alaa Alhamzeh et al. “A Stacking Approach for Cross-Domain Argument Identification”. In: International Conference on Database and Expert Systems Applications. Springer. 2021, pp. 361–373 (cit. on pp. 2, 4, 7, 8, 21, 24, 27, 45, 70).

[Alh+22a] Alaa Alhamzeh et al. “Empirical Study of the Model Generalization for Argument Mining in Cross-Domain and Cross-Topic Settings”. In: Transactions on Large-Scale Data-and Knowledge-Centered Systems LII. Springer, 2022, pp. 103–126 (cit. on pp. ii, 5, 30, 64, 66, 71).

[Alh+22b] Alaa Alhamzeh et al. “Query Expansion, Argument Mining and Document Scoring for an Efficient Question Answering System”. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer. 2022, pp. 162–174 (cit. on p. 42).

[Atk+17] Katie Atkinson et al. “Towards artificial argumentation”. In: AI magazine 38.3 (2017), pp. 25–36 (cit. on p. 1).

[BAM09] Trevor Bench-Capon, Katie Atkinson, and Peter McBurney. “Altruism and agents: an argumentation based approach to designing agent decision mechanisms”. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2. 2009, pp. 1073–1080 (cit. on p. 1).

[BN06] Christopher M Bishop and Nasser M Nasrabadi. Pattern recognition and machine learning. Vol. 4. 4. Springer, 2006 (cit. on p. 14).

[BAA19] Rihab Bouslama, Raouia Ayachi, and Nahla Ben Amor. “Using Convolutional Neural Network in Cross-Domain Argumentation Mining Framework”. In: International Conference on Scalable Uncertainty Management. Springer. 2019, pp. 355–367 (cit. on p. 17).

[Car+18] Winston Carlile et al. “Give me more feedback: Annotating argument persuasiveness and related attributes in student essays”. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018, pp. 621–631 (cit. on p. 19).

[CW17] Nicholas Carlini and David Wagner. “Towards evaluating the robustness of neural networks”. In: 2017 ieee symposium on security and privacy (sp). Ieee. 2017, pp. 39–57 (cit. on p. 16).

References (2)

[CT15] Lucas Carstens and Francesca Toni. “Towards relation based argumentation mining”. In: Proceedings of the 2nd Workshop on Argumentation Mining. 2015, pp. 29–34 (cit. on pp. 1, 22).

[CHM19] Tuhin Chakrabarty, Christopher Hidey, and Kathleen McKeown. “IMHO fine-tuning improves claim detection”. In: arXiv preprint arXiv:1905.07000 (2019) (cit. on pp. 16, 18).

[CHC21] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. From Opinion Mining to Financial Argument Mining. Springer Nature, 2021 (cit. on pp. 19, 22, 23, 36).

[Cra18] Belinda Crawford Camiciottoli. “Persuasion in earnings calls: A diachronic pragmalinguistic analysis”. In: International Journal of Business Communication 55.3 (2018), pp. 275–292 (cit. on p. 36).

[Coh+21] Cohen, Daniel H. "Evaluating arguments and making meta-arguments." Informal Logic 21.2 (2001).

[Dax+17] Johannes Daxenberger et al. “What is the essence of a claim? cross-domain claim identification”. In: arXiv preprint arXiv:1704.07203 (2017) (cit. on pp. 18, 65).

[De 13] Joost CF De Winter. “Using the Student’s t-test with extremely small sample sizes”. In: Practical Assessment, Research, and Evaluation 18.1 (2013), p. 10 (cit. on p. 49).

[Dev+18] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: arXiv preprint arXiv:1810.04805 (2018) (cit. on pp. 9, 33, 39, 42).

[Gou+14] Theodosis Goudas et al. “Argument extraction from news, blogs, and social media”. In: Hellenic Conference on Artificial Intelligence. Springer. 2014, pp. 287–299 (cit. on pp. 1, 21).

[Gre+20] Shai Gretz et al. “A large-scale dataset for argument quality ranking: Construction and analysis”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. 05. 2020, pp. 7805–7813 (cit. on pp. 2, 11, 20, 22, 24, 25, 34, 39, 68).

[GAW21] Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. “Assessing the Sufficiency of Arguments through Conclusion Generation”. In: arXiv preprint arXiv:2110.13495 (2021) (cit. on pp. 24, 25).

[HG16] Ivan Habernal and Iryna Gurevych. “Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm”. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, pp. 1589–1599 (cit. on pp. 20, 24).

[HG17] Ivan Habernal and Iryna Gurevych. “Argumentation mining in user-generated web discourse”. In: Computational Linguistics 43.1 (2017), pp. 125–179 (cit. on pp. 7, 28).

[Hos+19] MD Zakir Hossain et al. “A comprehensive survey of deep learning for image captioning”. In: ACM Computing Surveys (CsUR) 51.6 (2019), pp. 1– 36 (cit. on p. 8).

[Kar+19] Shigeki Karita et al. “A comparative study on transformer vs rnn in speech applications”. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE. 2019, pp. 449–456 (cit. on p. 8).

References (3)

[Yan+19a] Wei Yang et al. “End-to-end open-domain question answering with bertserini”. In: arXiv preprint arXiv:1902.01718 (2019) (cit. on p. 42). [Yan+20] Zekun Yang et al. “Bert representations for video question answering”. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020, pp. 1556–1565 (cit. on p. 39).

[Eem+04] Frans H. van Eemeren and Rob Grootendorst. 2004. A Systematic Theory of Argumentation: The PragmaDialectical Approach. Cambridge University Press, Cambridge, UK.

[EG19] Vlad Eidelman and Brian Grom. “Argument identification in public comments from eRulemaking”. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law. 2019, pp. 199–203 (cit. on pp. 1, 21).

[FPP07] David Freedman, Robert Pisani, and Roger Purves. “Statistics (international student edition)”. In: Pisani, R. Purves, 4th edn. WW Norton & Company, New York (2007) (cit. on p. 10).

[GZ19] Amirata Ghorbani and James Zou. “Data shapley: Equitable valuation of data for machine learning”. In: International Conference on Machine Learning. PMLR. 2019, pp. 2242–2251 (cit. on p. 12).

[Din+20] Ming Ding et al. “Cogltx: Applying bert to long texts”. In: Advances in Neural Information Processing Systems 33 (2020), pp. 12792–12804 (cit. on p. 36).

[DS19] Lorik Dumani and Ralf Schenkel. “A systematic comparison of methods for finding good premises for claims”. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019, pp. 957–960 (cit. on p. 1).

[DLC20] Esin Durmus, Faisal Ladhak, and Claire Cardie. “The role of pragmatic and discourse context in determining argument impact”. In: arXiv preprint arXiv:2004.03034 (2020) (cit. on p. 39). [EDG17] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. “Neural end-toend learning for computational argumentation mining”. In: arXiv preprint arXiv:1704.06104 (2017) (cit. on p. 16).

[Wu+16] Yonghui Wu et al. “Google’s neural machine translation system: Bridging the gap between human and machine translation”. In: arXiv preprint arXiv:1609.08144 (2016) (cit. on p. 9).

[Lau+20] Anne Lauscher et al. “Rhetoric, logic, and dialectic: Advancing theorybased argument quality assessment in natural language processing”. In: arXiv preprint arXiv:2006.00843 (2020) (cit. on pp. 19, 22, 24).

[LR20] John Lawrence and Chris Reed. “Argument mining: A survey”. In: Computational Linguistics 45.4 (2020), pp. 765–818 (cit. on p. 27). [LKB20] Qi Liu, Matt J Kusner, and Phil Blunsom. “A survey on contextual embeddings”. In: arXiv preprint arXiv:2003.07278 (2020) (cit. on p. 42).

[PN15] Isaac Persing and Vincent Ng. “Modeling argument strength in student essays”. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, pp. 543– 552 (cit. on pp. 19, 24, 27, 32).

References (4)

[LLS09] Ying Liu, Han Tong Loh, and Aixin Sun. “Imbalanced text classification: A term weighting approach”. In: Expert systems with Applications 36.1 (2009), pp. 690–701 (cit. on p. 56).

[Liu+19] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: arXiv preprint arXiv:1907.11692 (2019) (cit. on p. 10).

[Lop+20] Luis Enrico Lopez et al. “Transformer-based end-to-end question generation”. In: arXiv preprint arXiv:2005.01107 4 (2020) (cit. on p. 39).

[LL17] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: Advances in neural information processing systems 30 (2017) (cit. on p. 12).

[Ma+20] Zhiqian Ma et al. “Utilization of deep learning to mine insights from earning calls for stock price movement predictions”. In: Proceedings of the First ACM International Conference on AI in Finance. 2020, pp. 1–8 (cit. on pp. 21, 23).

[MML20] R. Thomas McCoy, Junghyun Min, and Tal Linzen. “BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance”. In: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Online: Association for Computational Linguistics, Nov. 2020, pp. 217– 227. DOI: 10 . 18653 / v1 / 2020 . blackboxnlp - 1 . 21. URL: https : //aclanthology.org/2020.blackboxnlp-1.21 (cit. on p. 66).

[Moe+07] Marie-Francine Moens et al. “Automatic detection of arguments in legal texts”. In: Proceedings of the 11th international conference on Artificial intelligence and law. 2007, pp. 225–230 (cit. on p. 7).

[MKH19] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. “When does label smoothing help?” In: Advances in neural information processing systems 32 (2019) (cit. on p. 14).

[Nag+15] Kenji Nagata et al. “An exhaustive search and stability of sparse estimation for feature selection problem”. In: IPSJ Online Transactions 8 (2015), pp. 25–32 (cit. on p. 41).

[NB07] Todd Neideen and Karen Brasel. “Understanding statistical tests”. In: Journal of surgical education 64.2 (2007), pp. 93–96 (cit. on p. 56).

[PM09] Raquel Mochales Palau and Marie-Francine Moens. “Argumentation mining: the detection, classification and structure of arguments in text”. In: Proceedings of the 12th international conference on artificial intelligence and law. 2009, pp. 98–107 (cit. on p. 23).

[PN13] Isaac Persing and Vincent Ng. “Modeling thesis clarity in student essays”. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013, pp. 260–269 (cit. on p. 24).

[Yan+19b] Zhilin Yang et al. “Xlnet: Generalized autoregressive pretraining for language understanding”. In: Advances in neural information processing systems 32 (2019) (cit. on p. 10).

References (5)

[PM20] Anushka Prakash and Harish Tayyar Madabushi. “Incorporating countbased features into pre-trained models for improved stance detection”. In: arXiv preprint arXiv:2010.09078 (2020) (cit. on p. 24).

[Rad+19] Alec Radford et al. “Language models are unsupervised multitask learners”. In: OpenAI blog 1.8 (2019), p. 9 (cit. on p. 10)

[SKW21] Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. “Learning from revisions: Quality assessment of claims in argumentation at scale”. In: arXiv preprint arXiv:2101.10250 (2021) (cit. on pp. 2, 24, 25).

[Sta18] Christian Stab. Argumentative writing support by means of natural language processing. Gesellschaft für Informatik eV, 2018 (cit. on p. 1).

[SG14a] Christian Stab and Iryna Gurevych. “Annotating argument components and relations in persuasive essays”. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers. 2014, pp. 1501–1510 (cit. on pp. 7, 20, 27, 28).

[SG14b] Christian Stab and Iryna Gurevych. “Identifying argumentative discourse structures in persuasive essays”. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014, pp. 46–56 (cit. on p. 7).

[SG17a] Christian Stab and Iryna Gurevych. “Parsing argumentation structures in persuasive essays”. In: Computational Linguistics 43.3 (2017), pp. 619– 659 (cit. on pp. 16, 21, 40).

[SG17b] Christian Stab and Iryna Gurevych. “Recognizing insufficiently supported arguments in argumentative essays”. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. 2017, pp. 980–990 (cit. on pp. 2, 19, 24, 32).

[SMG18] Christian Stab, Tristan Miller, and Iryna Gurevych. “Cross-topic argument mining from heterogeneous sources using attention-based neural networks”. In: arXiv preprint arXiv:1802.05758 (2018) (cit. on pp. 18, 66).

[Ste20] Manfred Stede. “Automatic argumentation mining and the role of stance and sentiment”. In: Journal of Argumentation in Context 9.1 (2020), pp. 19– 41 (cit. on p. 1).

[Sun+19] Chi Sun et al. “How to fine-tune bert for text classification?” In: China national conference on Chinese computational linguistics. Springer. 2019, pp. 194–206 (cit. on p. 42).

[SEW15] Reid Swanson, Brian Ecker, and Marilyn Walker. “Argument mining: Extracting arguments from online dialogue”. In: Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue. 2015, pp. 217–226 (cit. on p. 20).

[SK17] Piotr Szymanski and Tomasz Kajdanowicz. “A Network Perspective on ´ Stratification of Multi-Label Data”. In: Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications. Ed. by Luís Torgo et al. Vol. 74. Proceedings of Machine Learning Research. ECML-PKDD, Skopje, Macedonia: PMLR, 2017, pp. 22– 35 (cit. on p. 55).

References (6)

[Tol+19] Assaf Toledo et al. “Automatic Argument Quality Assessment–New Datasets and Methods”. In: arXiv preprint arXiv:1909.01007 (2019) (cit. on p. 39). [TE11] Antonio Torralba and Alexei A Efros. “Unbiased look at dataset bias”. In: CVPR 2011. IEEE. 2011, pp. 1521–1528 (cit. on pp. 30, 65).

[Tou03] Stephen E Toulmin. The uses of argument. Cambridge university press, 2003 (cit. on p. 28). [Vas+17] Ashish Vaswani et al. “Attention is all you need”. In: Advances in neural information processing systems 30 (2017) (cit. on pp. 8, 13).

[WSA17] Henning Wachsmuth, Benno Stein, and Yamen Ajjour. ““PageRank” for argument relevance”. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. 2017, pp. 1117–1127 (cit. on p. 24).

[WW20] Henning Wachsmuth and Till Werner. “Intrinsic quality assessment of arguments”. In: arXiv preprint arXiv:2010.12473 (2020) (cit. on p. 25).

[Wac+17a] Henning Wachsmuth et al. “Building an argument search engine for the web”. In: Proceedings of the 4th Workshop on Argument Mining. 2017, pp. 49–59 (cit. on p. 1).

[Wac+17b] Henning Wachsmuth et al. “Computational argumentation quality assessment in natural language”. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. 2017, pp. 176–187 (cit. on pp. 2, 19, 20, 22, 25).

[WMS20] Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. “Unlocking transfer learning in argumentation mining: A domain-independent modelling approach”. In: 15th International Conference on Wirtschaftsinformatik. 2020 (cit. on pp. 1, 21).

[Wan+21] Qian Wan et al. “Automated Claim Identification Using NLP Features in Student Argumentative Essays.” In: International Educational Data Mining Society (2021) (cit. on p. 40).

[WL16] Lu Wang and Wang Ling. “Neural network-based abstract generation for opinions and arguments”. In: arXiv preprint arXiv:1606.02785 (2016) (cit. on p. 1).

[Wan+19] Zhiguo Wang et al. “Multi-passage bert: A globally normalized bert model for open-domain question answering”. In: arXiv preprint arXiv:1908.08167 (2019) (cit. on p. 39).

[Xu+19] Hu Xu et al. “BERT post-training for review reading comprehension and aspect-based sentiment analysis”. In: arXiv preprint arXiv:1904.02232 (2019) (cit. on p. 42).

[YCZ21] Chenyu You, Nuo Chen, and Yuexian Zou. “Self-supervised contrastive cross-modality representation learning for spoken question answering”. In: arXiv preprint arXiv:2109.03381 (2021) (cit. on p. 39).

Thank you for your attention!

Any questions ?