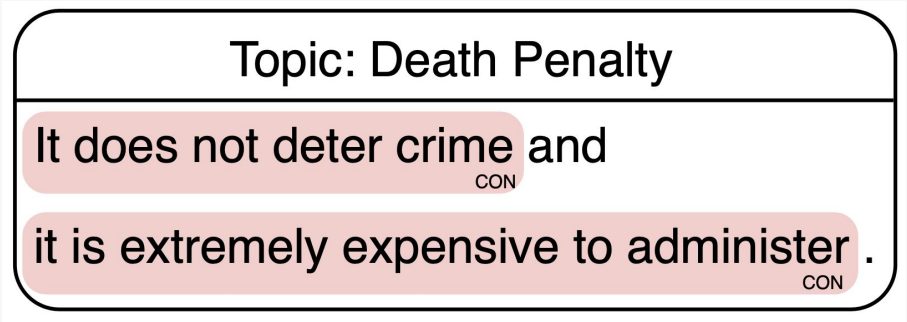
Generalization Across Argument Mining Datasets

Supervisor: Michael Fromm

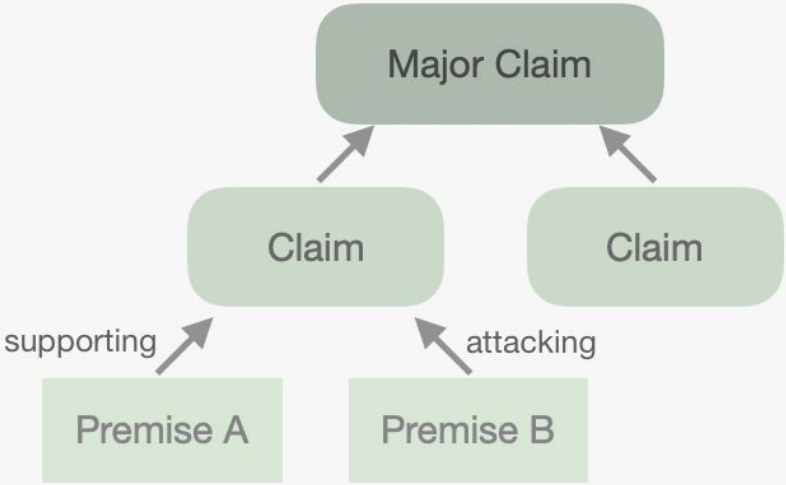
Members: Caroline Frank, Dominik Seeholzer, Elvira Siegel, Stefanie Kunze & Yindong Wang

Argument Mining

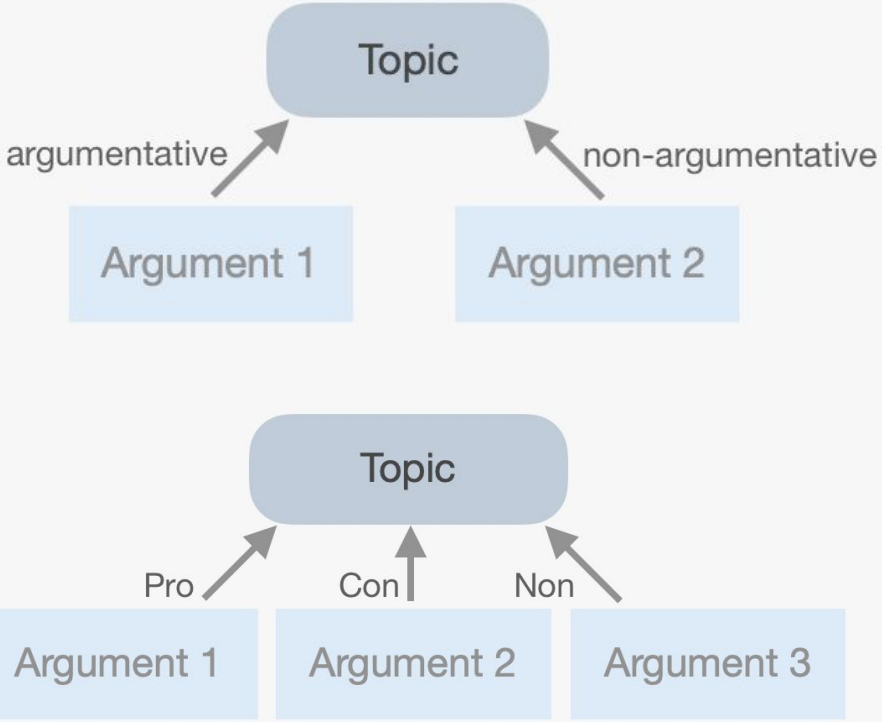


Fine-Grained Argument Unit Recognition and Classification

Examples of Argument Models



Stab and Gurevych, 2014



Argument Mining for Cross Topic Heterogeneous Sources 2018

Motivation & Goals

- Is it beneficial to include data from different argument models to increase accuracy?

Motivation & Goals

- Is it beneficial to include data from different argument models to increase accuracy?

  → Using 5 different annotated datasets and convert the labels

Motivation & Goals

- Is it beneficial to include data from different argument models to increase accuracy?

  → Using 5 different annotated datasets and convert the labels

- How well do natural language processing (NLP) machine learning models generalize across argument mining datasets?

Motivation & Goals

- Is it beneficial to include data from different argument models to increase accuracy?

  → Using 5 different annotated datasets and convert the labels

- How well do natural language processing (NLP) machine learning models generalize across argument mining datasets?

  → Train on dataset X and test on dataset Y

Motivation & Goals

- Is it beneficial to include data from different argument models to increase accuracy?

  → Using 5 different annotated datasets and convert the labels

- How well do natural language processing (NLP) machine learning models generalize across argument mining datasets?

  → Train on dataset X and test on dataset Y

- Is "argumentativeness" similarly captured across datasets?

Motivation & Goals

- Is it beneficial to include data from different argument models to increase accuracy?

  → Using 5 different annotated datasets and convert the labels

- How well do natural language processing (NLP) machine learning models generalize across argument mining datasets?

  → Train on dataset X and test on dataset Y

- Is "argumentativeness" similarly captured across datasets?

  → Train with Multitask-Model on all datasets

Dataset Statistics (Comparison)

Dataset	Size	Number of Topics
AURC	8000	8
CTAM	25489	8
CWAM	**30422**	218
PASPE	6547	**402**
PD	29532	0

Example after Preprocessing

1. Fine-Grained Argument Unit Recognition and Classification (AURC)

Topic	Sentence	Label
school uniforms	Uniforms not only save money but also time.	Pro

ID	Topic	Sentence	Label
1	school uniforms	Uniforms not only save money but also time.	supporting-argumentative

Pro → supporting-argumentative
Con → attacking-argumentative
Non → non-argumentative

Example after Preprocessing

2. Cross-topic Argument Mining from Heterogeneous Sources (CTAM)

Topic	Sentence	Annotation
nuclear energy	Nevertheless, the problems of nuclear waste, safety and proliferation still remain to be solved.	Argument_against

ID	Topic	Sentence	Label
1	nuclear energy	Nevertheless, the problems of nuclear waste, safety and proliferation still remain to be solved.	attacking-argumentative

Argument_for → supporting-argumentative
Argument_against → attacking-argumentative
NoArgument → non-argumentative

Example after Preprocessing

3. Corpus Wide Argument Mining - a Working Solution (CWAM)

Wikipedia ID	MotionText	DomainConcept	Evidence	acceptance Rate
1345	Casinos should be **banned**	Casino	"Activist groups argued that a casino could also lead to undesirable activities often … "	**0.95**

ID	Topic	Sentence	Label
1345	Casino	"Activist groups argued that a casino could also lead to undesirable activities often … "	**attacking-argumentative**

acceptanceRate < 0.6 non-argumentative
acceptanceRate ≥ 0.6 attacking/supporting (depends on Motion Text)

Example after Preprocessing

4. Parsing Argumentation Structures in Persuasive Essays (PASPE)

First row (File1)	Sentence (File 2)	Label (File 2)
Should students be taught to compete or to cooperate?	we should attach more importance to cooperation during primary education.	MajorClaim

ID	Topic	Sentence	Label
1	Should students be taught to compete or to cooperate?	we should attach more importance to cooperation during primary education.	non-argumentative

402 Essays with 2 files per essay
File 1: topic + all sentences in essay | File 2: sentence + label

Premise → argumentative
Claim → non-argumentative
MajorClaim → non-argumentative

Example after Preprocessing

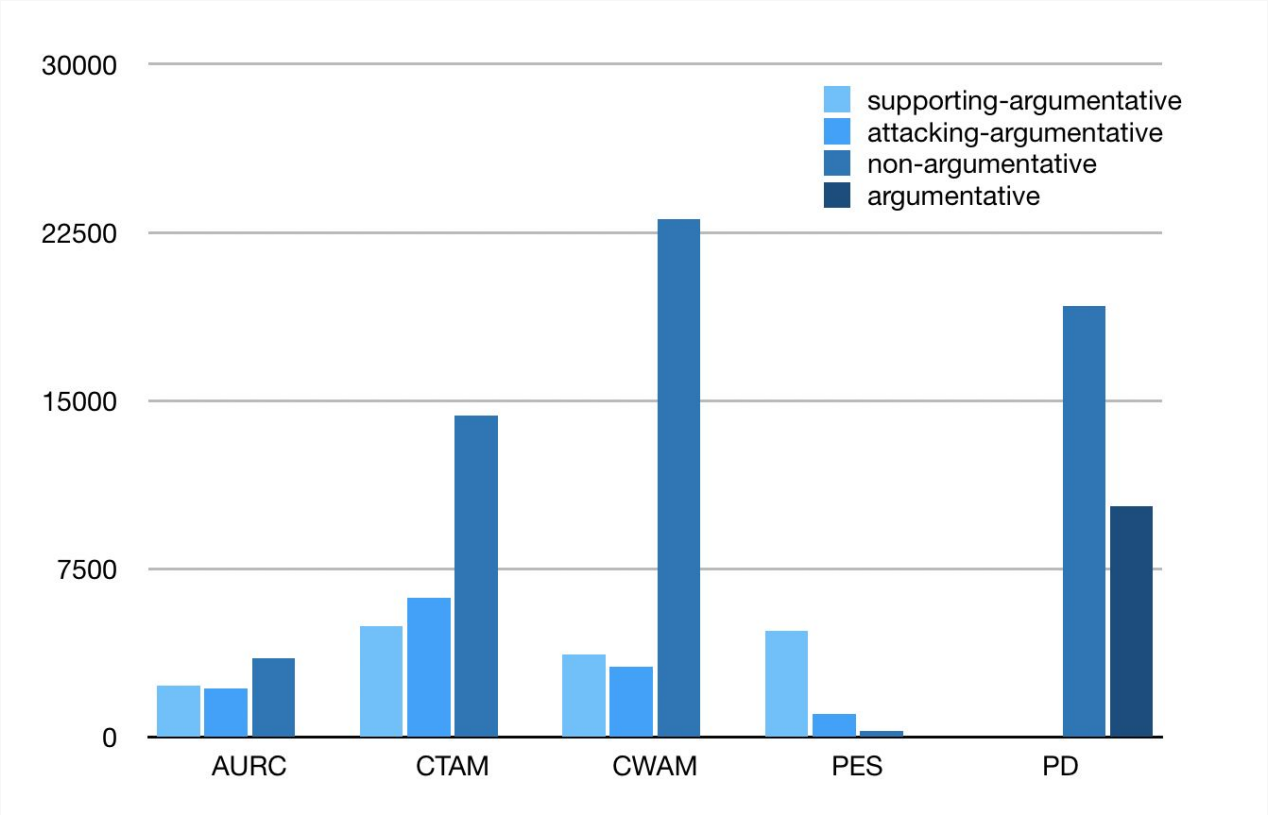
5. Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates (PD)

ID	Sentence	Label
9369	It's been a time, therefore, of illusion and false hopes, and the longer it continues, the more dangerous it becomes.	Premise

ID	Topic	Sentence	Label
9369	None	It's been a time, therefore, of illusion and false hopes, and the longer it continues, the more dangerous it becomes.	argumentative

Premise → argumentative | Claim → non-argumentative | Others → non-argumentative

Histogram Placeholder

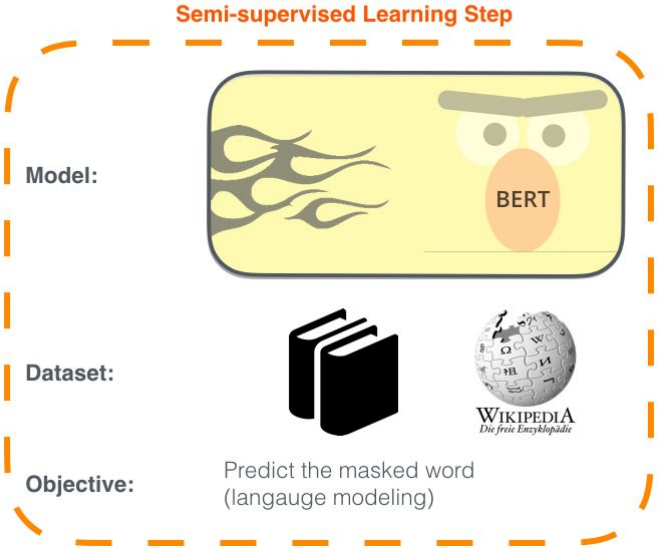


What is BERT?

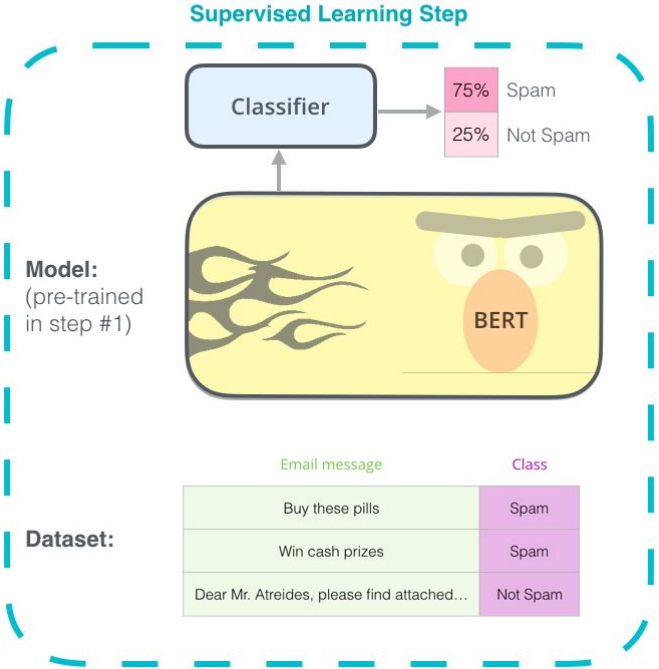
- Stands for:
  - Bidirectional Encoder Representations from Transformers
- key technical innovation:
  - Apply the bidirectional training on **Transformers**
  - In contrast to previous models: left to right, right to right or combined training of both
- Transfer Learning
  - Pre-training model for new purpose-specific task

What is BERT?

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).



2 - Supervised training on a specific task with a labeled dataset.

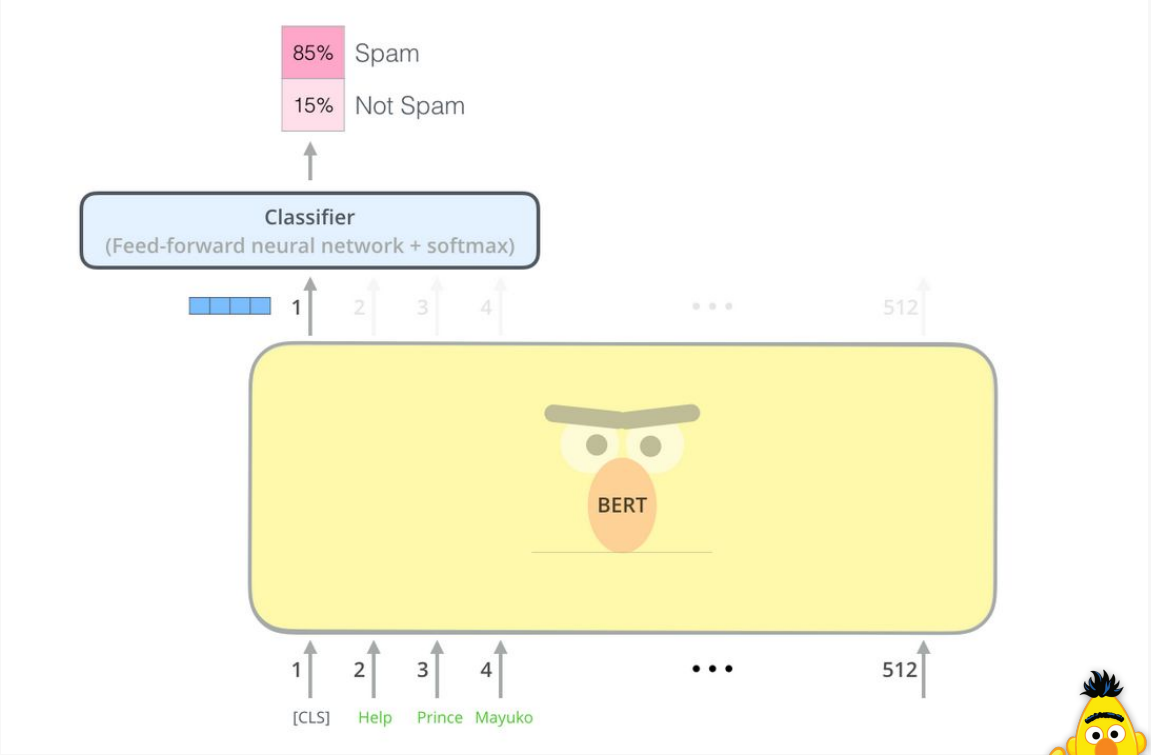


https://jalammar.github.io/illustrated-bert/

Model illustration

Example: Sentence Classification

- Binary classifier:
  - ‘Spam’,
  - ‘Not Spam’
- Multi-label classifier:
  - ‘Spam’,
  - ‘Not Spam’,
  - ‘Promotion’
  - ‘Social’

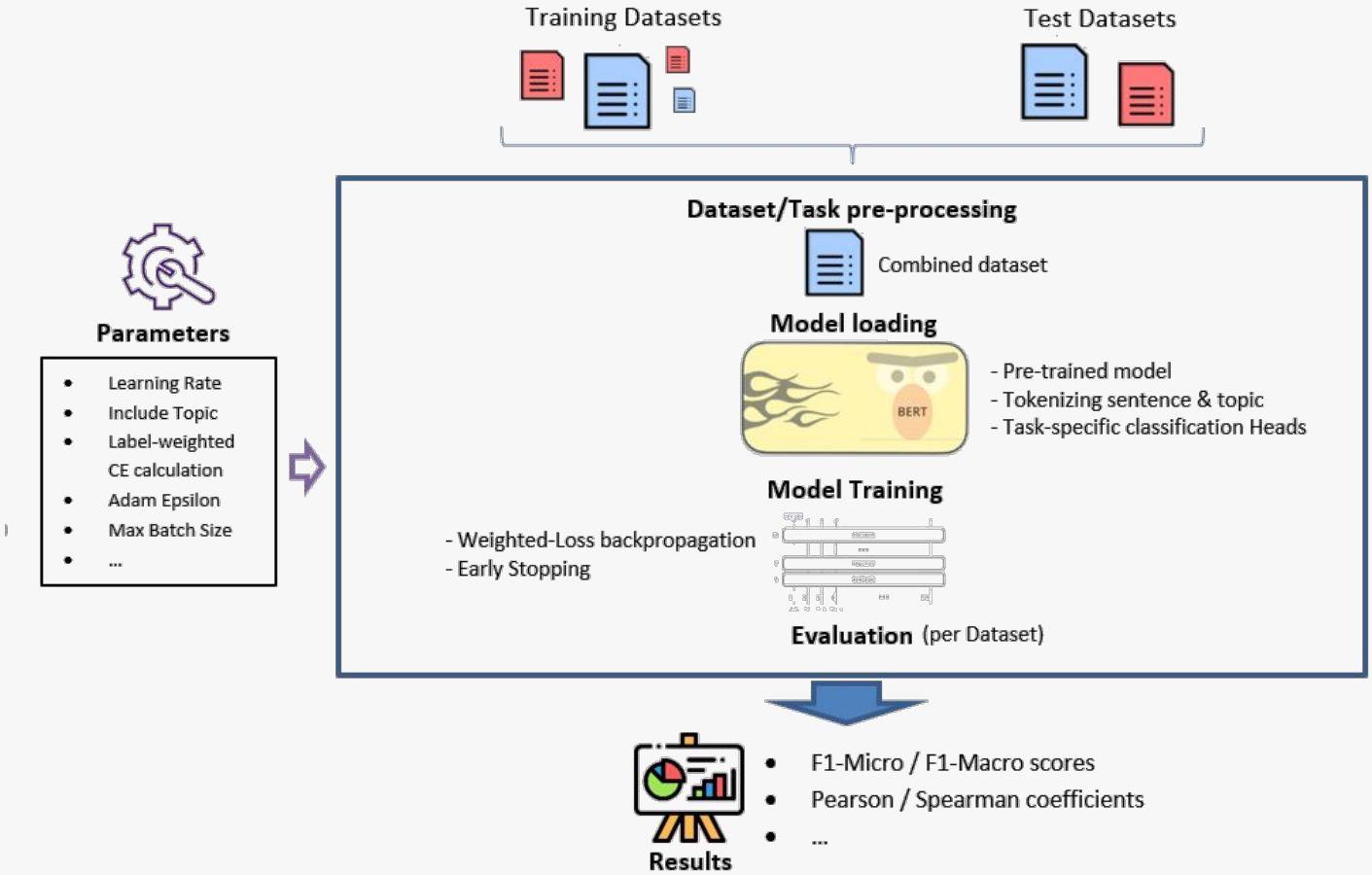


https://jalammar.github.io/illustrated-bert/

Possible downstream tasks of BERT

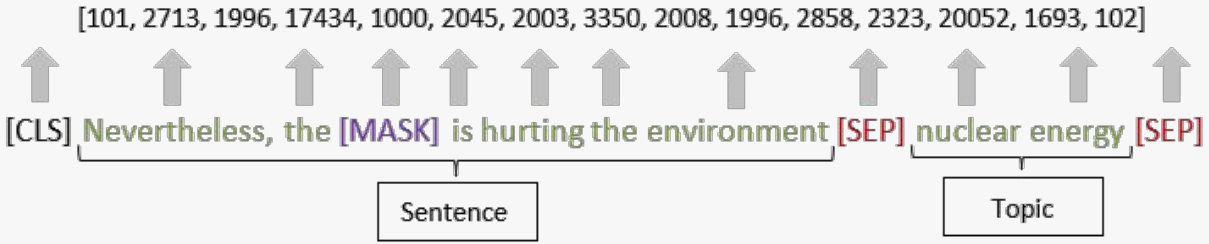
- Classification tasks
  - sentiment analysis
  - **argument mining**
- Question Answering
- Named Entity Recognition
- **...**

Our Framework



Our Framework - Details

- **Topic Tokenizer:** Include topic information in sentence tokenization



- **Percentile Max Sequence Length:** Determine max length of tokenized sentences based on percentile (improves runtime)

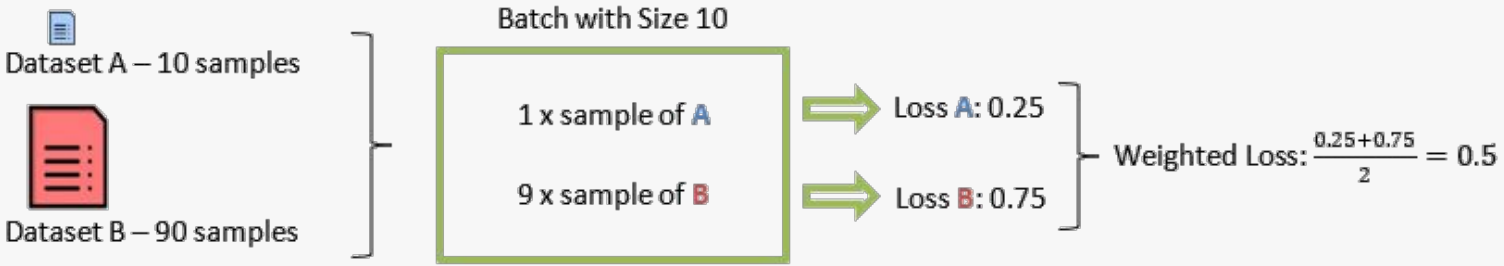
[CLS] Very **long sentences** are cut [MASK] after a certain amount of [SEP] nuclear energy [SEP]

[CLS] **Short sentences** are padded [MASK] [padding] ... [SEP] nuclear energy [SEP]

Our Framework - Details

- Maximize Batch Size: Faster convergence due to maximal variance

- Weighted Batch Sampling & Loss Calculation: Prevent big datasets from dominating training



Our Framework - Details

- Early Stopping: Stop training after validation loss increases consecutively

- Weighted cross entropy calculation: Useful for unbalanced training datasets

Our Framework - Details

- Other:
  - Monitoring with **MlFLow**
  - Model/Dataset loading & remote artifact logging
  - Various **Scenarios** (Cross-topic, In-topic, Cross-val-topic)
  - Caching Features, Max-Token-Length
  - Extendable for more tasks

Training Arguments

- Adam Epsilon = 1e-08
- Learning Rate = 2e-05
- Patience = 5 (Validation after half epoch)
- Max Sequence Length Percentile = 0.95
- Scenario = Cross-Topic
- Pre-trained Model = Bert-base-uncased

Motivation Recap

1. Is it beneficial to include data from different argument models to increase accuracy?
2. How well do NLP machine learning models generalize across datasets?
3. Is "argumentativeness" similarly captured across datasets?

Results and Experimental Setup

1. Is it beneficial to include data from different argument models to increase accuracy?

- Annotations are how the data is labeled by argument structure
  - Pro-Con-Non (AURC, CTAM)
  - Acceptance Score (CWAM)
  - MajorCLaim, Claim, Premise (PASPE)
  - Claim-Premise-Other (PD)

- First: Old labels vs New labels
- Second: Single vs Multi-task model:
  - Single Task: Training and testing on one dataset
  - Multi-task: Training on all datasets then evaluating each dataset separately

F1 Macro Results Old vs New Label

Is it beneficial to include data from different argument models to increase accuracy?

DATASET	Single Task Old Labels	Single Task New Labels
AURC	**0.7642**	-
CTAM	**0.6796**	-
CWAM	0.6543 (Spearman)	0.7194 (F1 Macro)
PASPE	**0.6192**	0.4745
PD	0.8492	**0.984**

F1 Macro Results Single vs Multi-task

Is it beneficial to include data from different argument models to increase accuracy?

DATASET	Single Task New Labels	Multi Task New Labels
AURC	**0.6994**	0.5426
CTAM	**0.6468**	0.5359
CWAM	**0.5195**	0.4989
PASPE	0.3392	**0.3867**
PD	**0.9827**	0.749

F1 Macro Results Single vs Multi-task

Is it beneficial to include data from different argument models to increase accuracy?

DATASET	Single Task Old Labels	Multi Task Old Labels
AURC	**0.6994**	0.6495
CTAM	**0.6468**	0.6065
CWAM	**0.6378 (Spearman)**	0.5894 (Spearman)
PASPE	0.5452	**0.5604**
PD	**0.8477**	0.6812

Experimental Setup

2. How well do natural language processing (NLP) machine learning models generalize across datasets?

- Trained model on one dataset
- Evaluated on other datasets
- No topic information included when training

F1 Macro Results Cross Training

How well do natural language processing (NLP) machine learning models generalize across datasets?

EVALUATED / TRAINED	AURC	CTAM	CWAM	PASPE	PD
AURC	0.668	**0.6038**	0.4733	0.2148	0.4280
CTAM	**0.644**	0.6472	0.4871	0.2086	0.3982
CWAM	**0.4396**	0.4202	0.5144	0.2199	0.4268
PASPE	0.1846	0.3065	0.2003	0.3315	**0.4515**
PD	0.2622	0.2806	0.2384	0.2124	0.9853

F1 Macro Results Cross Training

How well do natural language processing (NLP) machine learning models generalize across datasets?

EVALUATED / TRAINED	AURC	CTAM	CWAM	PASPE	PD
AURC	0.668	0.6038	0.4733	0.2148	0.4280
CTAM	0.644	0.6472	0.4871	0.2086	0.3982
CWAM	0.4396	0.4202	0.5144	0.2199	0.4268
PASPE	0.1846	0.3065	0.2003	0.3315	0.4515
PD	0.2622	0.2806	0.2384	0.2124	0.9853

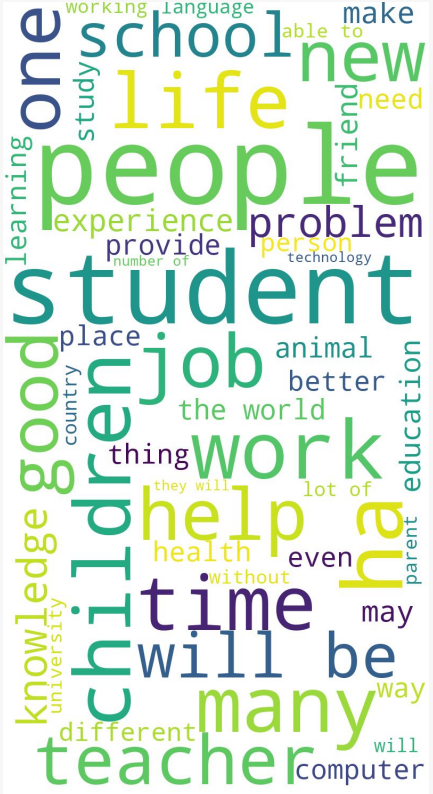
F1 Macro Results Cross Training

How well do natural language processing (NLP) machine learning models generalize across datasets?

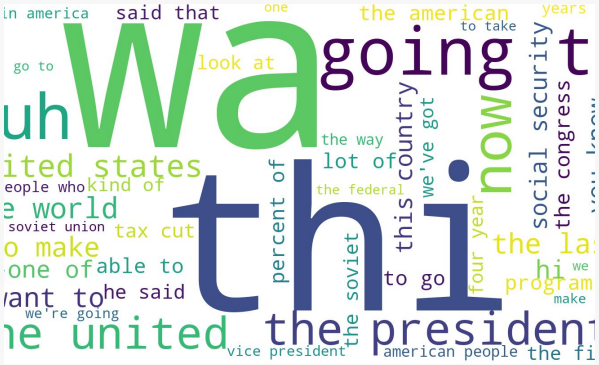
EVALUATED / TRAINED	AURC	CTAM	CWAM	PASPE	PD
AURC	0.668	**0.6038**	0.4733	0.2148	0.4280
CTAM	**0.644**	0.6472	0.4871	0.2086	0.3982
CWAM	0.4396	0.4202	0.5144	0.2199	0.4268
PASPE	0.1846	0.3065	0.2003	0.3315	0.4515
PD	0.2622	0.2806	0.2384	0.2124	0.9853

Word Cloud

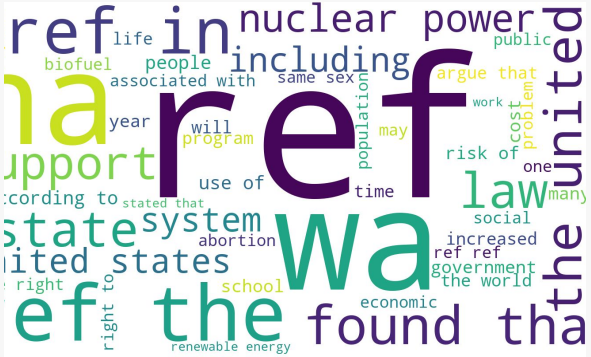
With topic words:



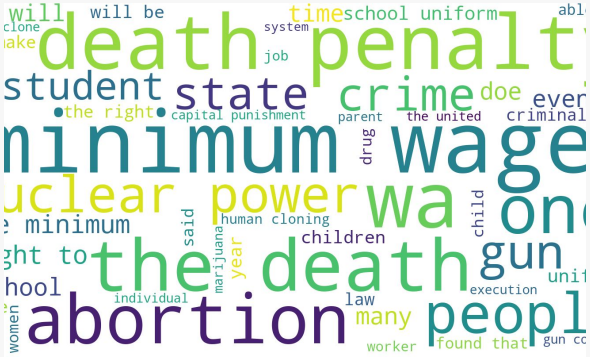
PES



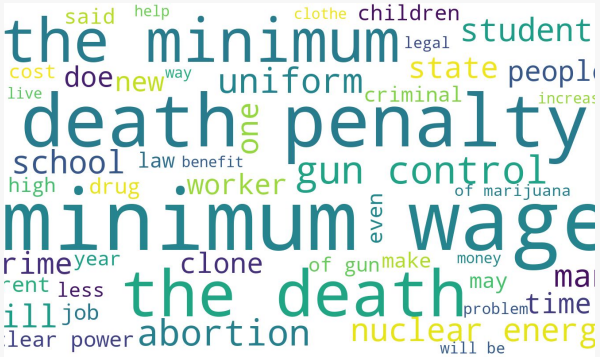
PD



CWAM



CTAM



AURC

Results

3. Is "argumentativeness" similarly captured across datasets?

- Single-task model: Training and testing on one dataset
- Multi-task model: Training on all datasets then evaluating each dataset separately on trained model

F1 Macro Results

Is "argumentativeness" similarly captured across datasets?

DATASET	Single Task		Multi Task	
	Old Labels	New Labels	Old Labels	New Labels
AURC	**0.6994**	-	0.6495	0.5426
CTAM	**0.6468**	-	0.6065	0.5359
CWAM	**0.6378**	0.5195	0.5894	0.4989
PASPE	0.5452	0.3392	**0.5604**	0.3867
PD	0.8477	**0.9827**	0.6812	0.749

F1 Macro Results Cross Training

Is "argumentativeness" similarly captured across datasets?

EVALUATED / TRAINED	AURC	CTAM	CWAM	PASPE	PD
AURC	0.668	**0.6038**	0.4733	0.2148	0.4280
CTAM	**0.644**	0.6472	0.4871	0.2086	0.3982
CWAM	0.4396	0.4202	0.5144	0.2199	0.4268
PASPE	0.1846	0.3065	0.2003	0.3315	0.4515
PD	0.2622	0.2806	0.2384	0.2124	0.9853

Results

Does including topic while training increase accuracy?

- With and without topic
- New Labels
- Single-task model: Training and testing on one dataset
- Multi-task model: Training on all datasets then evaluating each dataset separately on trained model

F1 Macro Results

Does including topic while training increase accuracy?

DATASET	Single Task		Multi Task	
	Without Topic	With Topic	Without Topic	With Topic
AURC	**0.6994**	0.6746	**0.6495**	0.4544
CTAM	0.6468	**0.6618**	**0.6065**	0.4353
CWAM	0.5195	**0.5643**	**0.5894**	0.5233
PASPE	0.3392	**0.3918**	**0.5604**	0.364
PD	0.9827	-	0.6812	**0.7359**

Conclusion

Is it beneficial to include data from different argument models to increase accuracy?

- No, did not improve results except for the small dataset PASPE

Conclusion

Is it beneficial to include data from different argument models to increase accuracy?

- No, did not improve results except for the small dataset PASPE

How well do natural language processing (NLP) machine learning models generalize across datasets?

- Datasets do not generalize well with the BERT model
  - Argument structure important
  - AURC and CTAM performed well but have same topics and argument structure

Conclusion

Is it beneficial to include data from different argument models to increase accuracy?

- No, did not improve results except for the small dataset PASPE

How well do natural language processing (NLP) machine learning models generalize across datasets?

- Datasets do not generalize well with the BERT model
  - Argument structure important
  - AURC and CTAM performed well but have same topics and argument structure

Is "argumentativeness" similarly captured across datasets?

- Not similarly captured across datasets
  - If it were, then they would generalize well
  - Only similar datasets achieved similar results

Future work

- Test more datasets and see which generalize well with each other
  - Do some argument structures work better with other argument structures?
- Knowledge distillation (a teacher-student model)
  - Has previously shown to improve multi-task
- Different models/tokenizers other than BERT/Hugging-Face

Thanks for your Attention!

Backup

F1 Micro Results Single vs Multi-task

Is it beneficial to include different annotated data to increase accuracy?

DATASET	F1 Micro - Single Task Old Labels	F1 Micro - Multi Task Old Labels
AURC	**0.7642**	0.7622
CTAM	**0.6796**	0.669
CWAM	**0.6543**	0.5896
PASPE	0.6192	**0.6327**
PD	**0.8492**	0.6812

F1 Micro Results Old vs New Label

Is it beneficial to include different annotated data to increase accuracy?

DATASET	F1 Micro - Single Task Old Labels	F1 Micro - Single Task New Labels
AURC	0.7642	0.7642
CTAM	0.6796	0.6796
CWAM	0.6543	**0.7194**
PASPE	**0.6192**	0.4745
PD	0.8492	**0.984**

F1 Micro Results Single vs Multi-task

Is it beneficial to include different annotated data to increase accuracy?

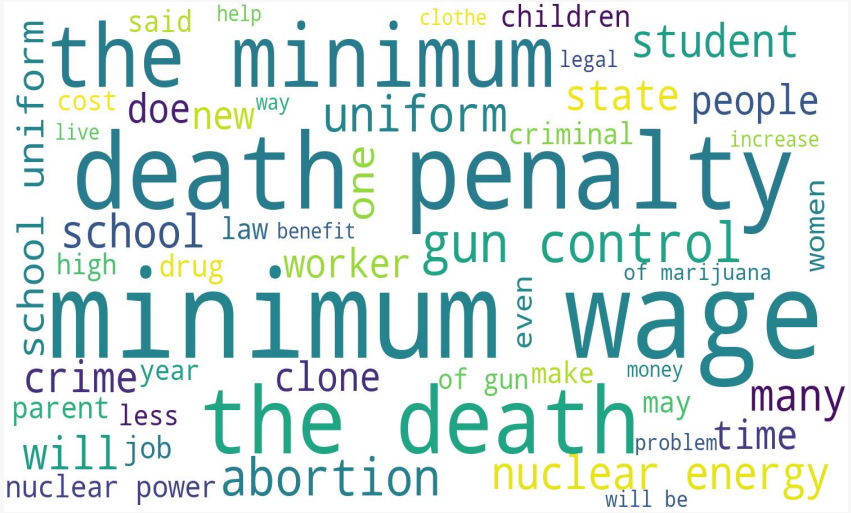
DATASET	F1 Micro - Single Task New Labels	F1 Micro - Multi Task New Labels
AURC	**0.7642**	0.6969
CTAM	**0.6796**	0.6201
CWAM	0.7194	**0.7225**
PASPE	0.4745	**0.6962**
PD	**0.984**	0.7594

F1 Micro Results

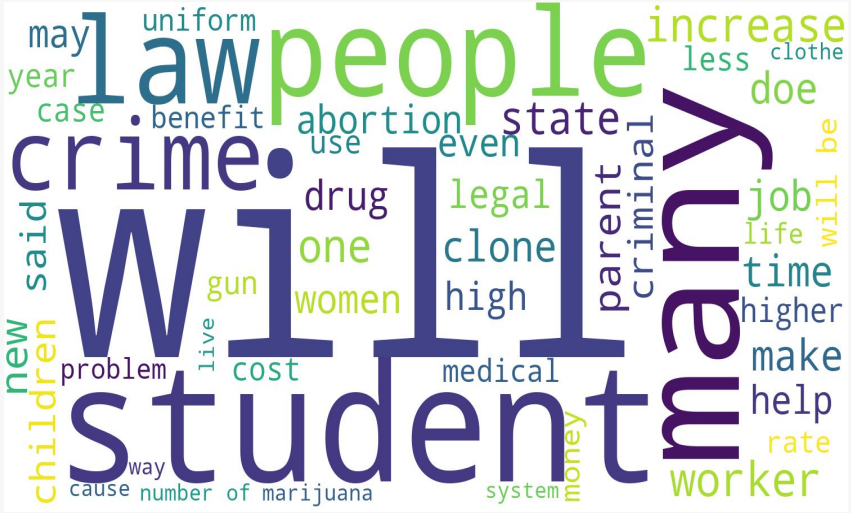
Is "argumentativeness" similarly captured across datasets?

DATASET	Single Task		Multi Task	
	Old Labels	New Labels	Old Labels	New Labels
AURC	**0.7642**	-	0.7622	0.6969
CTAM	**0.6796**	-	0.669	0.6201
CWAM	0.6543	0.7194	0.5896	**0.7225**
PASPE	0.6192	0.4745	0.6327	**0.6962**
PD	0.8492	**0.984**	0.6812	0.7594

Word Cloud



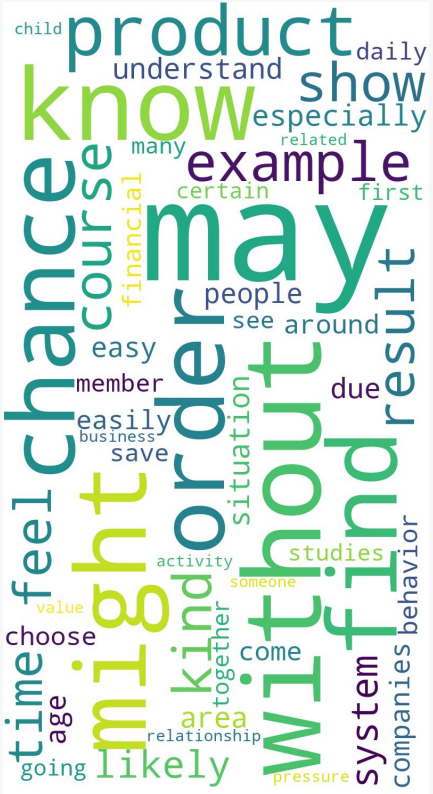
With topic words



Without topic words

Word Cloud

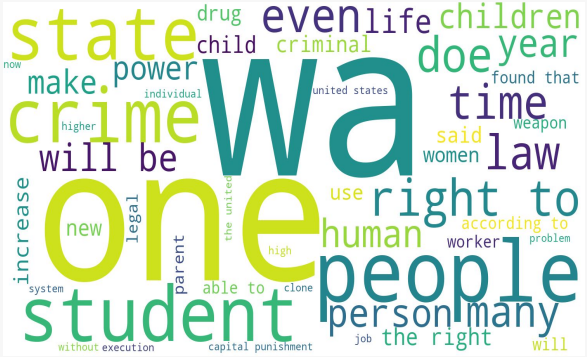
Without topic words:



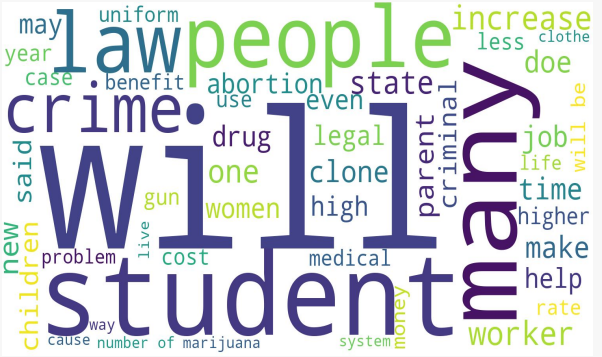
PES

PD

CWAM



CTAM



AURC

Argument vs Argumentativeness Definition

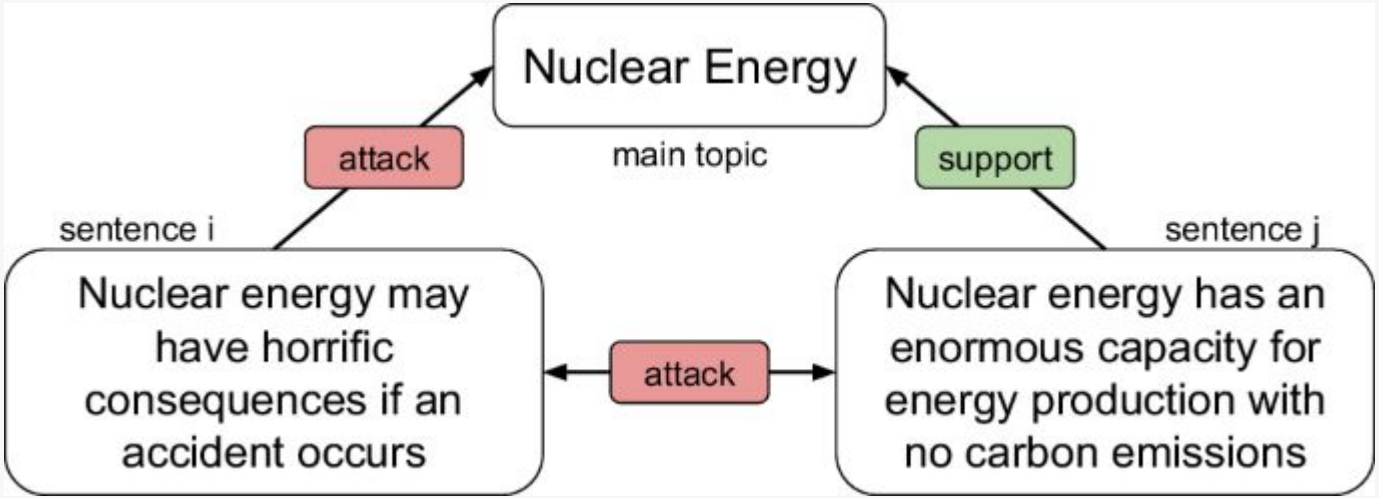
Components of an argument are claims and premises.

- The simplest structure of an argument is a combination of premises which are conveyed to justify a certain claim. A crucial part of an argument is a claim or the conclusion.
- Asserting a claim is the main goal of an argument made by someone, which can be true or false.
- The claim is usually supported by at least one evidence or premise. These premises are stated such that the claim of the argument is reasonable. One cannot expect the claim of their argument to be accepted by others if they do not have strong premises to justify the claim.
- Thus, another important component of the argument are premises for backing the claim.
- There are some structures that premises and claims may follow to form an argument.
- These structures are not definitive proof of existence of an argument and are not as common in oral debates than in written persuasive essays. However, they can be helpful to trace arguments. After detection of these structures we should investigate from the context, if the structure is used to develop an argument

"We define an argument as a span of text expressing evidence or reasoning that can be used to either support or oppose a given topic. An argument need not be “direct” or self-contained—it may presuppose some common or domain knowledge, or the application of commonsense reasoning—but it must be unambiguous in its orientation to the topic."

Introduction

Ex. Claim: Nuclear Energy is good



Preprocessing

Old labels vs. New labels

Dataset	Old Label	New Label
AURC/ CTAM	Con, Pro, Non	Attacking-argumentative, Supporting-argumentative, Non-argumentative
CWAM	acceptanceRate [0.0, 1.0]	Attacking-argumentative, Supporting-argumentative, Non-argumentative
PASPE	MajorClaim, Claim, Premise	Attacking-argumentative, Supporting-argumentative, Non-argumentative
PD	Premise, Claim, Other	Argumentative, Non-argumentative

Scores

Accuracy is the number of correctly predicted data points out of all the data points

$$\text{Precision}_c = \frac{\text{True Pos}_c}{\text{True Pos}_c + \text{False Pos}_c}$$

$$\text{Recall}_c = \frac{\text{True Pos}_c}{\text{True Pos}_c + \text{False Neg}_c}$$

$$\text{F1 Score}_c = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

$$\text{Macro F1 Score} = \frac{1}{n}\sum_{c=i}^{n} \text{F1 Score}_c$$