Statistical VS Neural Approach

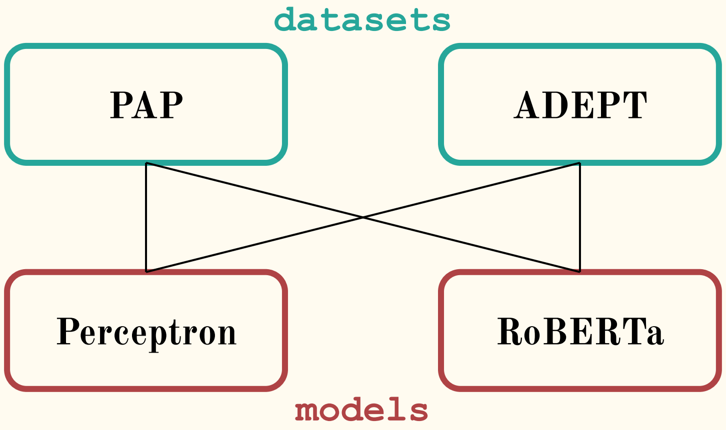
Modeling Semantic Plausibility

Swift Swans 🦢

Nina Vikhrova, Mattalika Intarahom, Martin Wolf

Contents

Overview – Roadmap



How does the performance change between **statistical** and **neural** approaches?

Overview – Datasets

PAP

How plausible is an SVO combination?

`scandal hurts department`

implausible — **plausible**

1 2 3 4 5

1728	216	216
train	dev	test

ADEPT

How does plausibility change with an additional adjective?

`The effect of (additional) sleeping is rejuvenation.`

impossible — **necessarily true**

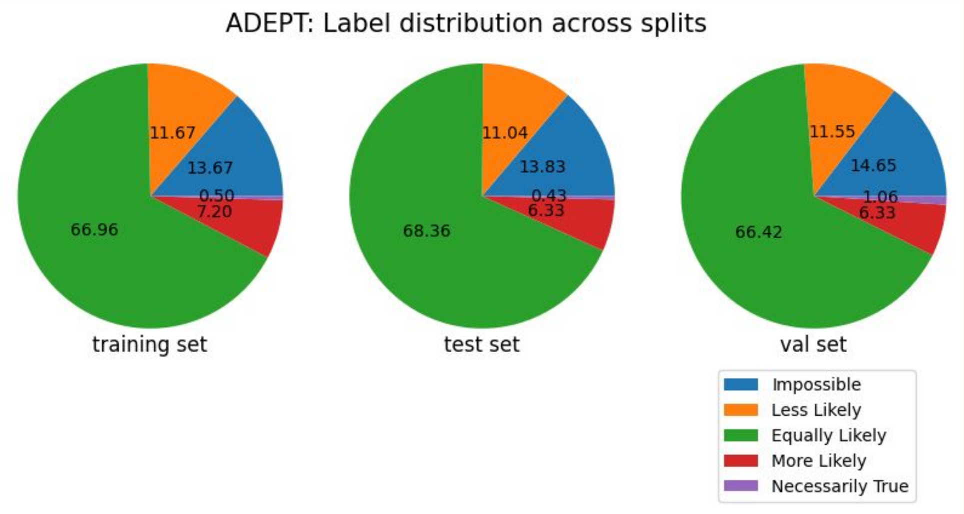
1 2 3 4 5

12 892	1 611	1 612
train	val	test

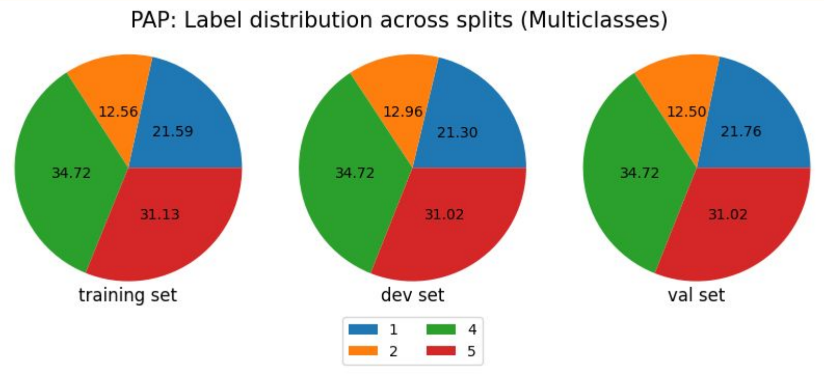
Label distribution (adept)

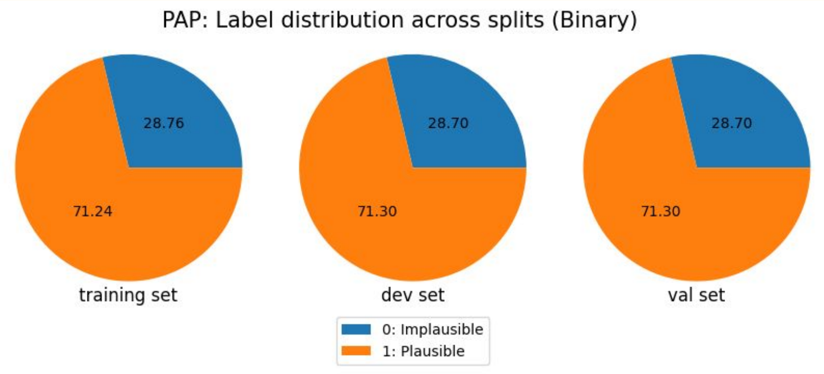
- Huge class imbalance
- ‘Equally likely’ most often

- Fewest examples in the positive direction (more likely & necessarily true)



Label distribution (pap)



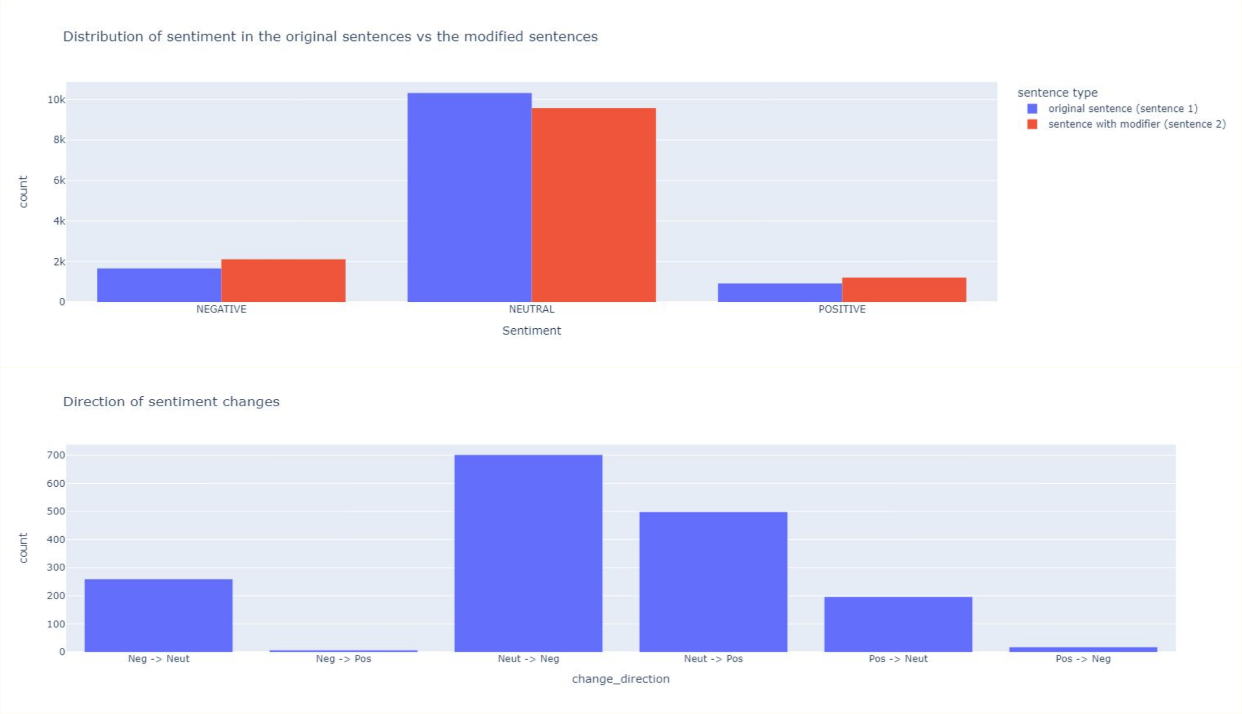


- More balanced
- Skew towards labels 4 & 5

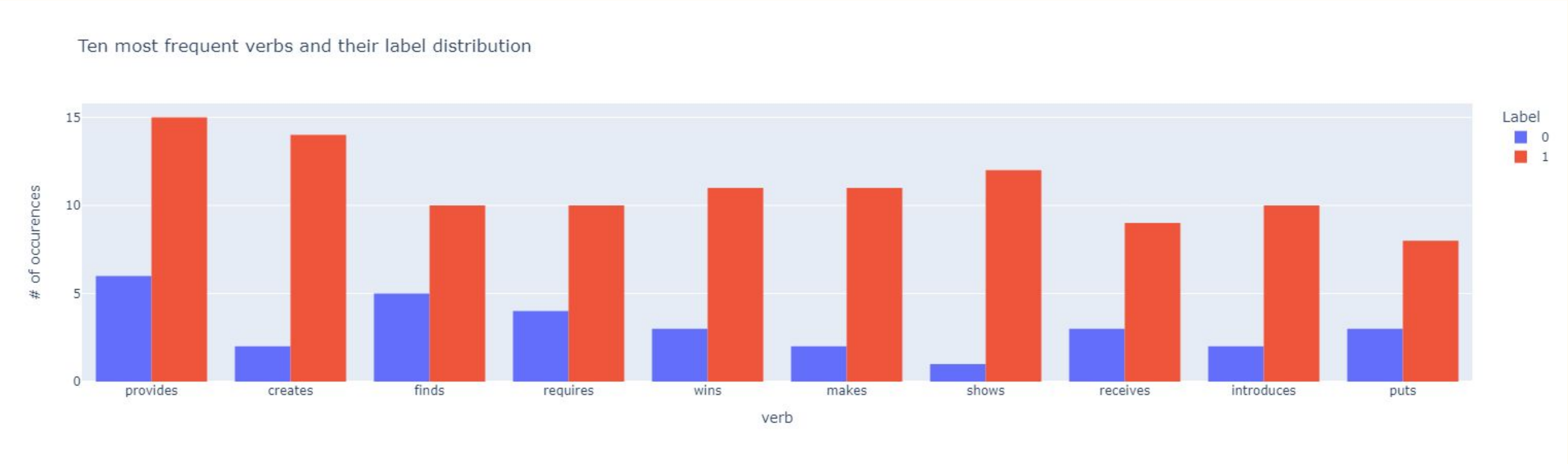
- 3 out of 4 times are plausible

Sentiment analysis (adept) (model: cardiffnlp/twitter-roberta-base-sentiment)

- Modifier adds stronger sentiment
- Most changes from neutral to pos/neg
- Some changes from pos/neg to neutral
- Extreme changes rare



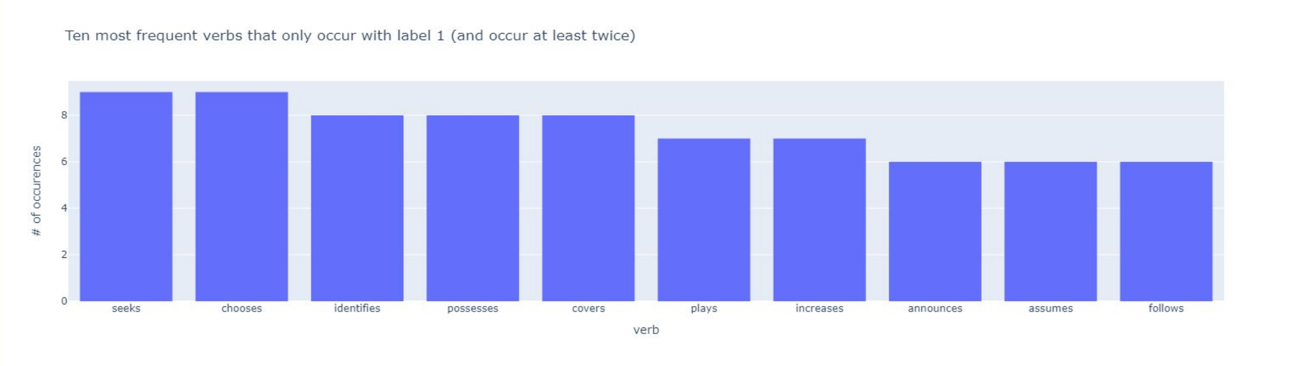
Occurrence analysis (pap)



- Frequently occurring verbs have more often label 1 than label 0

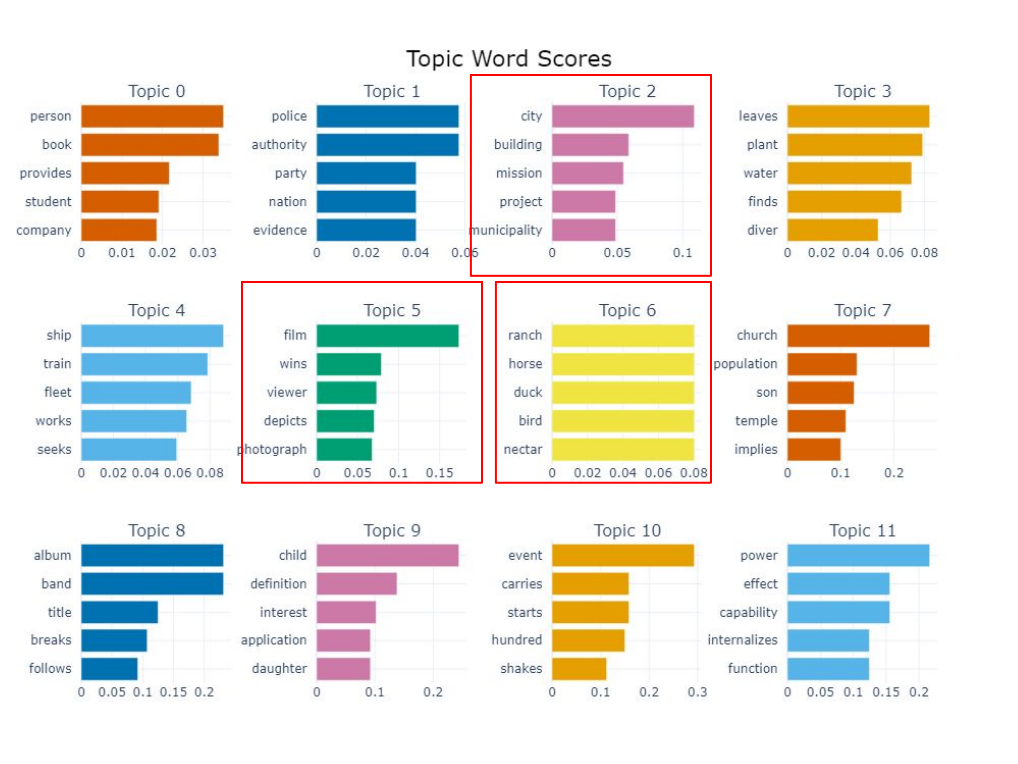
Occurrence analysis (pap)

- Overview of the most frequent verbs for each label



Topic analysis (pap)

- BERTopic
- Some topics plausible
- Because of the short text examples, topic modelling is generally hard



Models – Perceptron

Input → **Tokenization** → **Word2Vec** → **Perceptron** → **Classifications**

Input

PAP:

- SVO

ADEPT:

- Sentence1 + [SEP] + Sentence2
- Only Sentence 2

Tokenization

- word_tokenizer
- MWETokenizer

Word2Vec

- CBOW model
- Average word embeddings
- Zero vector for unknown words

Models – Perceptron

	PAP	**ADEPT**
`Embedding size`	100	100
`Max Epochs:`	1000	1000
`Early stopping:`	False	True
`Stop Criteria:`	loss > previous_loss - tol	loss > previous_loss - tol

Models – RoBERTa

(PAP)

`Subject Verb Object`

Input

`[CLS] + Sentence1`
`+ [SEP] + Sentence2`

(ADEPT)

Tokenization

`AutoTokenizer`

- Model-dependent
- Special tokens defined here

roberta-base

Trainer

`AutoModelForClassification`

- Multi-class setting
- Finetune directly
- Trainer used for the whole process

Pipeline

`text-classification`

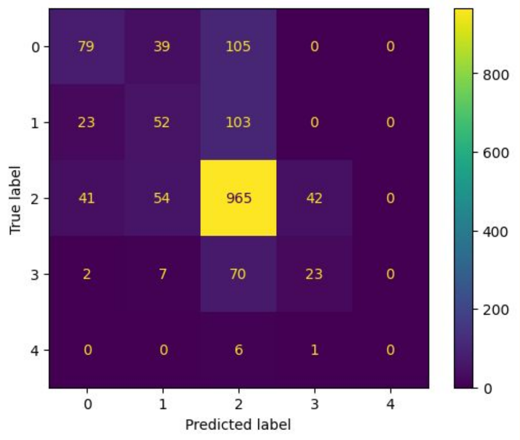
Classifications on test sets

Models – RoBERTa

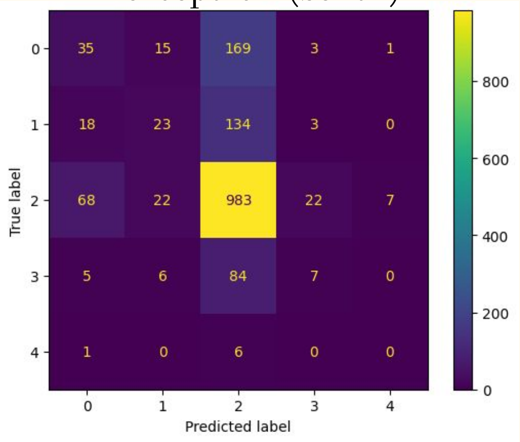
	PAP	ADEPT
Initial Accuracy: (no fine-tuning)	(dev) 0.13	(val) 0.12
Epochs:	5	3
Training Rate:	$2e^{-5}$	$3e^{-5}$
Final Accuracy:	0.407	0.698

Results adept

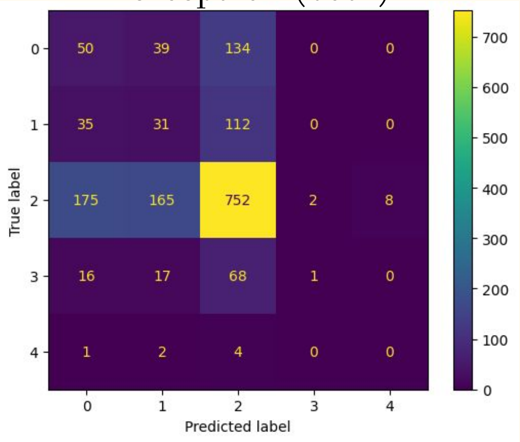
RoBERTa



Perceptron (sent2)



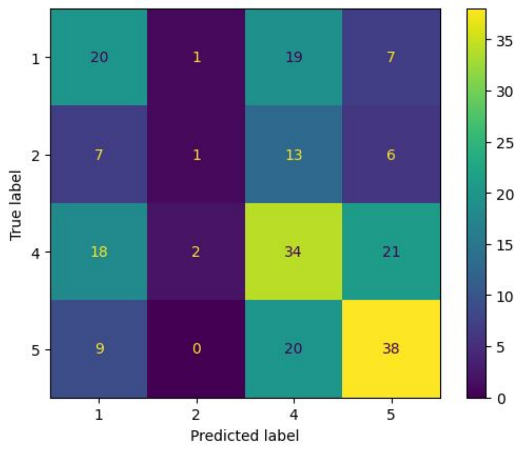
Perceptron (both)



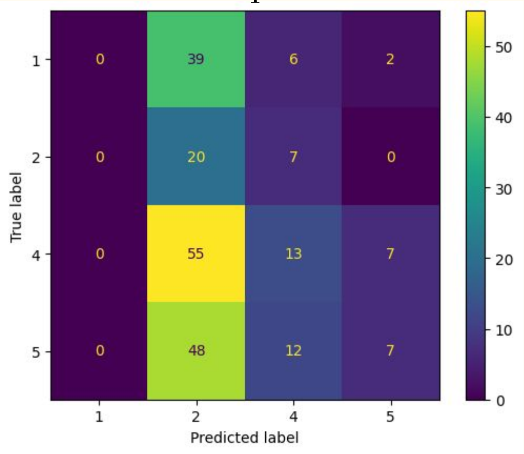
	RoBERTa	Perceptron (sent2)	Perceptron (both)
Micro F1	0.694	0.650	0.517
Macro F1	0.367	0.256	0.211
ROC-AUC	0.604	0.535	0.514

Results pap

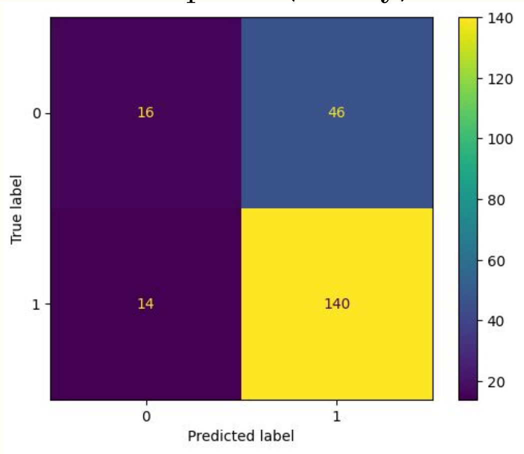
RoBERTa



Perceptron



Perceptron (binary)



	RoBERTa	Perceptron	Perceptron (binary)
Micro F1	0.430	0.185	0.722
Macro F1	0.357	0.152	0.585
ROC-AUC	0.583	0.503	0.583

Conclusion

- All models are skewed towards the most-occurring class

Adept Micro F1	0	1	2	3	4
RoBERTa	0.429	0.315	**0.820**	0.273	0
Perceptron (sent2)	0.2	0.188	**0.793**	0.102	0

- Seeing the whole sentence helps both approaches
- Statistical method performs worse but not by as much as expected
- Perceptron best on binary classification

Future work:

- Pre-trained word2vec for Perceptron
- Augmentation of training data to balance classes
- Find better fine-tuning hyperparameters
- LORA approaches

References

Emami Ali, Porada Ian, Olteanu Alexandra, Suleman Kaheer, Trischler Adam, Cheung Jackie Chi Kit. *ADEPT: An Adjective-Dependent Plausibility Task*. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 7117–7128.

Eichel Annerose, Schulte im Walde Sabine. *A Dataset for Physical and Abstract Plausibility and Sources of Human Disagreement*. Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII). Toronto, Canada: Association for Computational Linguistics, VII 2023. 31–45.

Links to models used:

Sentiment analysis: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Topic modeling: https://maartengr.github.io/BERTopic/

Perceptron: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Perceptron.html

RoBERTa: https://huggingface.co/docs/transformers/model_doc/roberta