# CyberWallE at SemEval-2020 Task 11: An Analysis of Feature Engineering for Ensemble Models for Propaganda Detection

Verena Blaschke, Maxim Korniyenko and Sam Tureski

Eberhard Karls Universität Tübingen

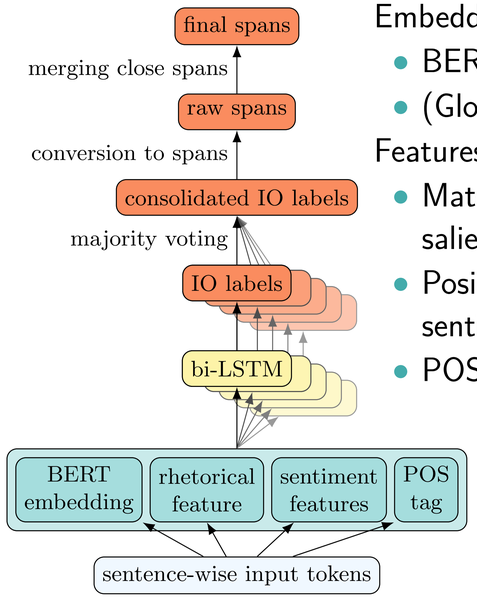
https://github.com/cicl-iscl/CyberWallE-propaganda-detection first.last@student.uni-tuebingen.de

## Introduction

Shared task [1] to automatically

- find propagandistic snippets in news articles
- determine which of 14 propaganda techniques is used in each such fragment

## Finding propagandistic spans




Embeddings

- BERT base uncased [2]
- (GloVe, small) [3]

Features

- Match with rhetorically salient phrase patterns [4]
- Positive/negative sentiment [5]
- POS tag [6]

### Results

Character-level F1: 43.9% (rank 8/35); recall > precision
(Corrected evaluation:* 43.6%; 12/35)

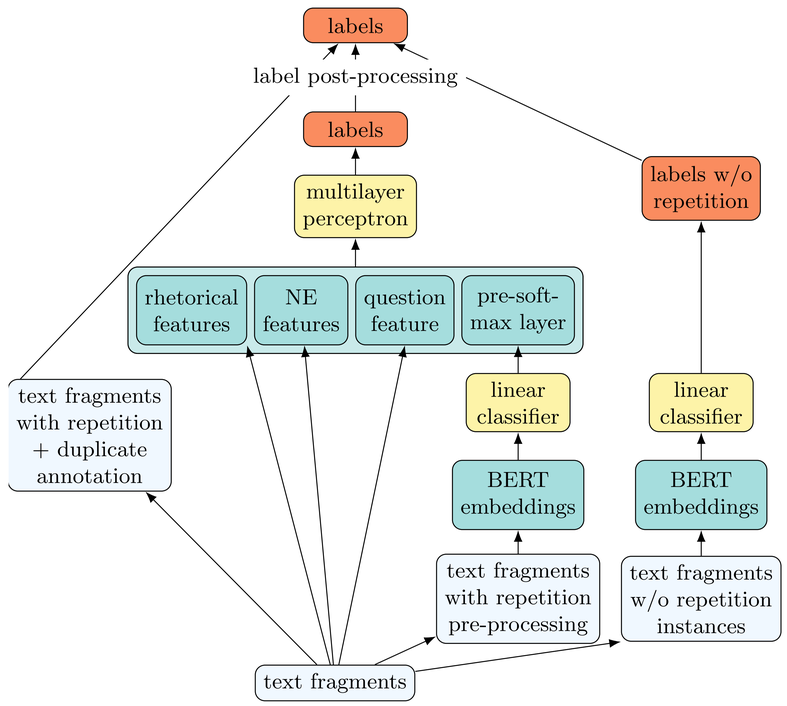
- GloVe (recall↑, precision↓) vs. BERT (recall↓, precision↑, F1↑)
- Features: precision↗ (at the cost of some recall)
- Majority voting: precision↗, more stable performance
- Span merging: recall↗

Repetition

it's inevitable that this bacterial infection [...] will become resistant to antibiotics. [...] But experts warn that [...] antibiotics resistance is inevitable and making this disease much more terrifying.

Appeal to fear/prejudice

## Identifying propaganda techniques




- Repetition pre-processing: Enhance fragments with 'repetition' label (train) or that appear elsewhere in the same article (dev/test) with duplicate
- Repetition post-processing: Classify **all** duplicates of a fragment within an article as 'repetition'
- Additional model predicts extra label for instances with 2+ classes

*The shared task's evaluation script contained a bug that was fixed after the competition ended.

## Identifying propaganda techniques cont'd

- Embeddings: BertForSequenceClassification [7]
- Classifier: MLP (SVM, linear classifier)

Features

- Phrase pattern in rhetorical lexicon [4]
- Nationality/religion, country/city [6]
- Question mark
- (More named entities, emotion [8], sequence length, # of repetitions)

### Results

Micro F1: 57.4% (rank 8/31)
(Corrected evaluation:* 58.9%; 6/31)

- Repetition pre- and postprocessing: F1↑
- Classifier choice: minor effect
- (Some) feature combinations: stability↗, F1↗

## References

[1] Giovanni Da San Martino et al. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval 2020, Barcelona, Spain, Dec. 2020.

[2] Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Jeffrey Pennington et al. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–43, 2014.

[4] Swapna Somasundaran et al. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, vol. 6, 2007.

[5] Stefano Baccianella et al. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, vol. 10, pp. 2200–2204, 2010.

[6] Matthew Honnibal and Ines Montani. spaCy 2: `spacy.io/`.

[7] Thomas Wolf et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[8] `https://cloud.ibm.com/catalog/services/natural-language-understanding`.