Fine-Tuned Neural Models for Propaganda Detection at the Sentence and Fragment Levels

TARIQ ALHINDI | JONAS PFEIFFER | SMARANDA MURESAN

Department of Computer Science, Columbia University
Ubiquitous Knowledge Processing Lab, Technische Universitat Darmstadt

PROBLEM

Propaganda aims at influencing a target audience with a specific group agenda using faulty reasoning and/or emotional appeals.

Automatic detection of propaganda has been studied mainly at the article level. However, in order to build computational models that can explain why an article is propagandistic, the model would need to detect specific techniques present at sentence or even token level.

CONTRIBUTIONS

CUNLP participation at the NLP4IF workshop's shared task for propaganda detection at both the sentence and fragment levels

Our best models:

Sentence-level Classification: Fine-tuned BERT classifier with modified probability thresholds for label prediction.

Fragment-level Classification: BiLSTM-CRF tagger with stacked Flair, Glove, and Urban embeddings along with one-hot encoded features

Sentence Level Classification

METHODS

- Models: fine-tuned BERT and Logistic Regression.
- Features: word embeddings, handcrafted features (LIWC and punctuation features), context: previous and next sentence
- Predict the majority class (non-propaganda) only if it has a prediction probability > 0.70, to deal with class imbalance.

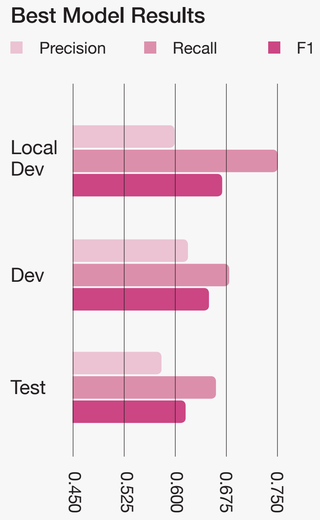
RESULTS

Feature Ablation Experiments

FEATURES	MODEL	DEVELOPMENT		
		P	R	F
text	BERT	0.69	0.55	0.61
text	**BERT**	**0.57**	**0.79**	**0.66**
context	BERT	0.70	0.53	0.60
context	BERT	0.63	0.67	0.65
BERT logits + handcrafted	LR	0.70	0.56	0.61
BERT logits + handcrafted	LR	0.60	0.71	0.65
BERT logits + tagged spans	LR	0.70	0.53	0.60
BERT logits + tagged spans	LR	0.61	0.71	0.66
BERT logits + all	LR	0.71	0.52	0.60
BERT logits + all	LR	0.61	0.71	0.66

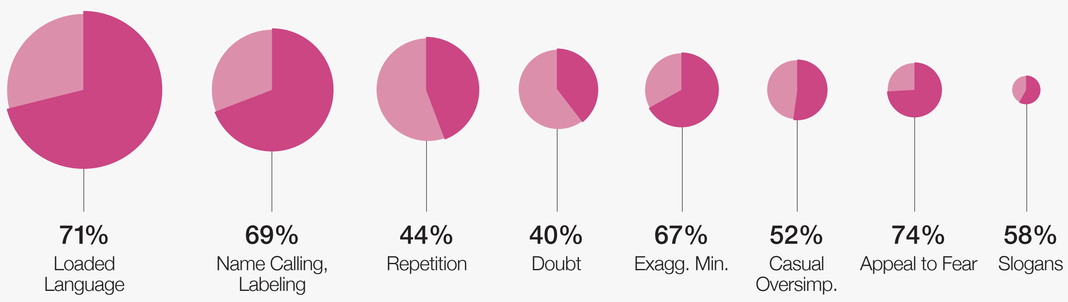
Non-propaganda class is predicted only if its prediction probability is > 0.80

Best Model Results



Non-propaganda class is predicted only if its prediction probability is > 0.70

Accuracy on Frequent Propaganda Techniques



OBSERVATIONS

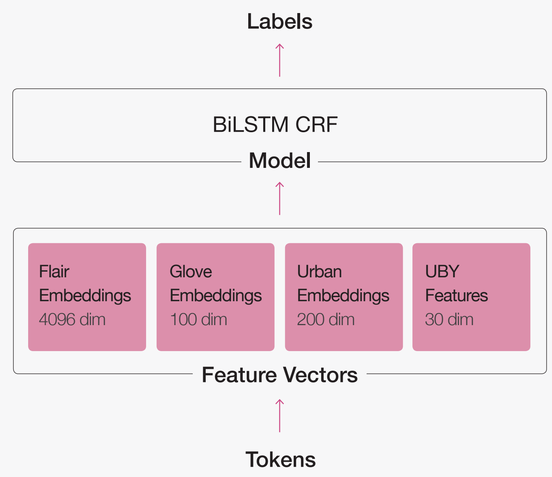
- Negligible effect of handcrafted and context features
- Better performance on frequent propaganda techniques in the dataset, except:
  - Repetition: occurs across sentences
  - Doubt: wide lexical coverage and variant sentence structure.

  "How is it possible the pope signed this decree?" and "I've seen little that has changed"

Fragment Level Classification

METHODS

- Model: BiLSTM-CRF tagger
- Features:
  - Flair Contextualized Embeddings: using both forward and backward embeddings
  - Glove Embeddings: using the 100 dimension embeddings
  - Urban Embeddings: Word2Vec embeddings trained on Urban Dictionary definitions, a crowdsourced online dictionary for slang words and phrases.
  - One-hot encoded features: using 30 concepts from the UBY dictionary that capture words associated with concepts such as offensive, vulgar, coarse, or ethnic slur.



RESULTS

Detailed Scores per Propaganda Technique

TECHNIQUE	DEVELOPMENT			TEST
	P	R	F	F
Appeal to Auth.	0	0	0	0.212
Appeal to Fear	0.285	0.006	0.011	0
Bandwagon	0	0	0	0
B&W Fallacy	0	0	0	0
Causal Oversimp.	0	0	0	0
Doubt	0.007	0.001	0.002	0
Exagg. Min.	0.833	0.085	0.154	0
Flag-Waving	0.534	0.102	0.171	0.195
Loaded Language	0.471	0.16	0.237	0.13

TECHNIQUE	DEVELOPMENT			TEST
	P	R	F	F
Name Calling	0.27	0.112	0.158	0.15
O,IV,C	0	0	0	0
Red Herring	0	0	0	0
Reductio ad hitlerum	0.318	0.069	0.113	0
Repetition	0	0	0	0
Slogans	0.221	0.034	0.059	0.003
Straw Men	0	0	0	0
Thought-terminating	0	0	0	0
Whataboutism	0	0	0	0

Overall Scores

	P	R	F
Dev	0.365	0.073	0.122
Test	0.323	0.082	0.131

OBSERVATIONS

- Some propaganda techniques are easier to classify across both datasets. For example, strong lexical signals ("American People" in Flag-Waving) and punctuation signals (quotes in Slogans).
- We have the highest precision among all teams on both dev and test sets. This could be due to adding UBY one-hot encoded features.

CONCLUSION

For some propaganda techniques, it is not enough to only look at one sentence to make an accurate prediction (e.g. Repetition) and therefore including the whole article as context is needed.

Using a combination of contextualized embeddings and one-hot encoded features improves the results in the FLC task.

Future work: experimenting with multiple methods each for detecting a sub-group of the propaganda techniques.

2nd Workshop on NLP for Internet Freedom (NLP4IF): Censorship, Disinformation, and Propaganda. 2019