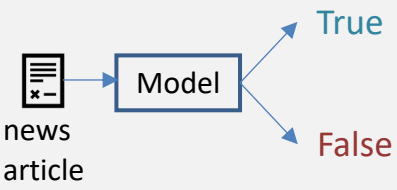
The Sally Smedley Hyperpartisan News Detector at SemEval-2019 Task 4: Learning Classifiers with Feature Combinations and Ensembling

Kazuaki Hanawa※ [1,2], Shota Sasaki※ [1,2], Hiroki Ouchi [1,2], Jun Suzuki [2,1], Kentaro Inui [2,1]
(* equal contribution) [1] RIKEN AIP, [2] Tohoku University

Task

Input & Output

A news article is **hyperpartisan**? or not?



Resources

1. Manually Annotated

, True
⋮
, False

High quality but Small scale (645 articles)

2. Automatically Annotated

All in *publisher A* are hyperpartisan
All in *publisher B* are not hyperpartisan

, True
⋮
, False

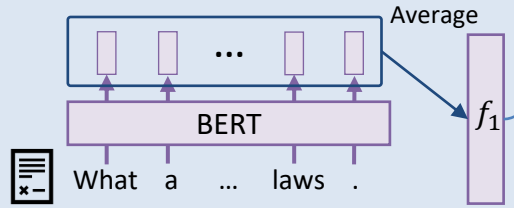
Low quality but Large scale (750k articles)

Method

BERT Feature

Motivation

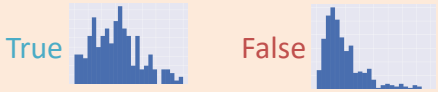
Representations based on language models ,e.g., BERT [Devlin et al., 2018] , contribute to performance improvements in a variety of NLP tasks



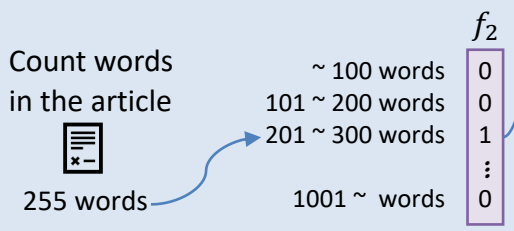
Article Length Feature

Motivation

Bias in the length of the article



Histogram of article length



Informative Phrase Feature

Motivation

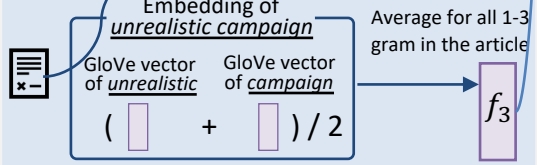
Some phrases predominantly appear in True/False articles

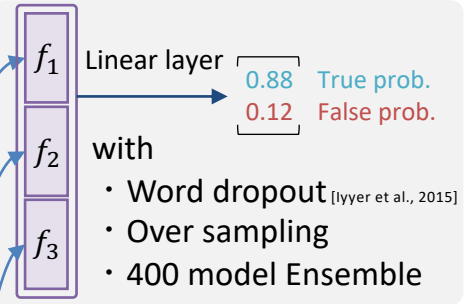
"unrealistic campaign" in True articles =45%

"unrealistic campaign" in False articles =2%

Preprocess: Extract informative 1-3 gram based on chi-squared value

Informative phrase list: *absurd* ⋮ *unrealistic campaign*





Results

Main Leaderboard

#	TEAM	ACCURACY
1	bertha-von-suttner	0.822
2	vernon-fenwick	0.820
3	sally-smedley	0.809
4	tom-jumbo-grumbo	0.806

Ablation Test

in cross validation manner

Features	Ensemble	Accuracy
f_1, f_2, f_3	☑	0.788
f_1, f_2, f_3		0.777
f_1, f_2		0.769
f_1		0.760