News Bias Detection
Team 12

- Mudit Malpani
- Shobhit Gautam

The Problem Statement

Given a News article, predict whether the article it’s Hyper-partisan or not.

One of the shared tasks of SEMEVAL-2019, with 41 participating teams.

Two Datasets: ‘byarticle’ and ‘bypublisher’.

Byarticle: Manually annotated by crowdsourcing, Comprised of 1273 articles.

Bypublisher: Annotated w.r.t. To publisher’s bias as reported from some fact-check websites. Comprised of over 750k articles.

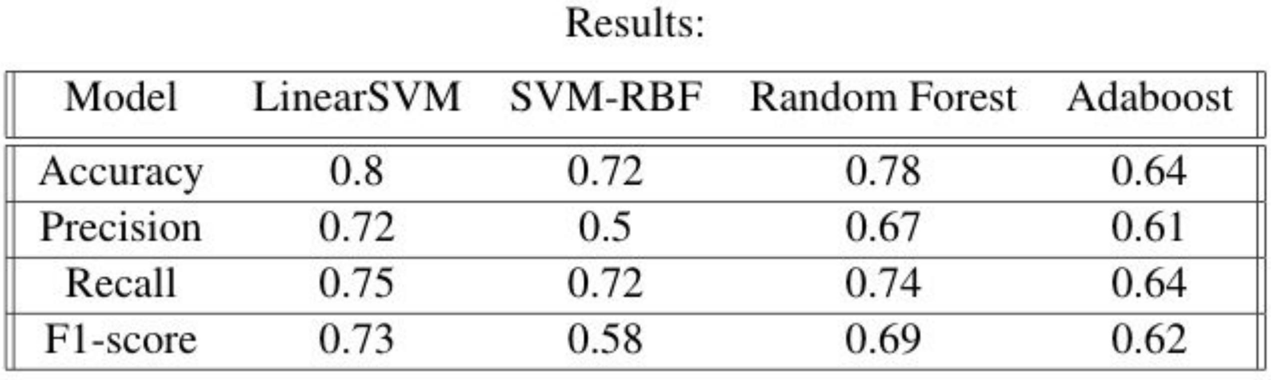
Models tried

- Classical Models: SVM with Linear and RBF kernel, Random Forest, Adaboost. (baseline)
- LSTM based Models: BiLSTM with/without Attention
- BERT based Models: With content and content+title as input.

Results and Analysis

Classical Models on ‘byarticle’ dataset

First we get TFIDF features, then test with each of the given models, on 2 label classification (partisan or not).



Results:

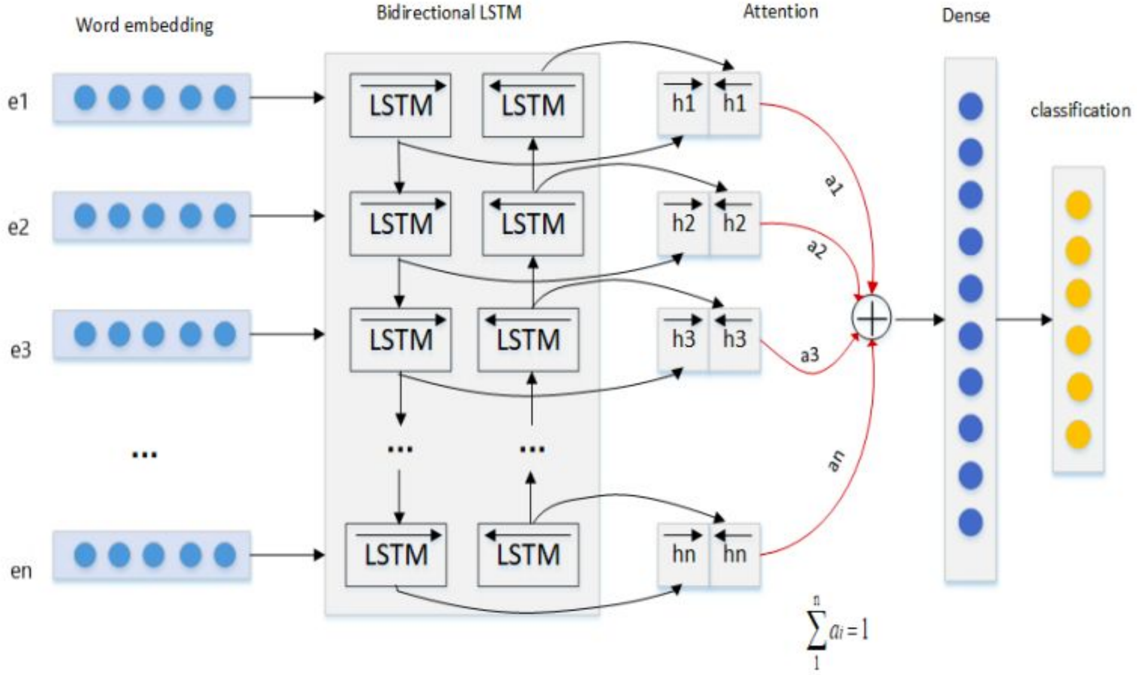
Model	LinearSVM	SVM-RBF	Random Forest	Adaboost
Accuracy	0.8	0.72	0.78	0.64
Precision	0.72	0.5	0.67	0.61
Recall	0.75	0.72	0.74	0.64
F1-score	0.73	0.58	0.69	0.62

Best performing Model: SVM with Linear Kernel

LSTM based Models on ‘byarticle’

With and Without

Attention


Word embedding
Bidirectional LSTM
Attention
Dense
classification
e1
e2
e3
...
en
LSTM
$\overrightarrow{h1}$ $\overleftarrow{h1}$
$\overrightarrow{h2}$ $\overleftarrow{h2}$
$\overrightarrow{h3}$ $\overleftarrow{h3}$
$\overrightarrow{hn}$ $\overleftarrow{hn}$
a1
a2
a3
an
$\sum_{1}^{n} a_i = 1$

With Attention

LSTM Results

Without Attention - 69% accuracy

With Attention - 73% accuracy

The bad performance can be attributed to the lack of data and pretrained model, as we just trained with ‘byarticle’

BERT Model on ‘byarticle’ and ‘bypublisher’

Used Hugging Face bert with a classification head.

Pre-Trained over by ‘bypublisher’ and finetuned on ‘byarticle’ (for more epocs).

Idea: use ‘bypublisher’ for domain adaptation and ‘byarticle’ for classification.

Results:

76% Accuracy pretrained on ‘byarticle’ tested on ‘byarticle’

86% Accuracy pretrained on ‘bypublisher’ tested on ‘‘bypublisher’

64% Accuracy pretrained on ‘bypublisher’ tested on ‘byarticle’

77% Accuracy pretrained on ‘bypublisher’ + finetuned on ‘‘byarticle’’ tested on ‘byarticle’

Final Thoughts

The model although seemingly performs well on ‘bypublisher’ with random test-train split, but performs bad when tested with articles from new publishers.

This suggests model learned the source of article rather than the article.

Model performs better on this data if just fine-tuned with ‘byarticle’.

In conclusion, inclusion of ‘bypublisher’ in any way seemingly lowers performance.