Biased News Data Influence on Classifying Social Media Posts

3rd Int'l Workshop on Recent Trends in News Information Retrieval (NewsIR 2019)
Collocated with 42nd Int'l ACM SIGIR, Paris, July, 2019

Marija Stanojevic, Jumanah Alshehri, Eduard Dragut, Zoran Obradovic

Temple University, Philadelphia, USA
Presented by: Zoran Obradovic

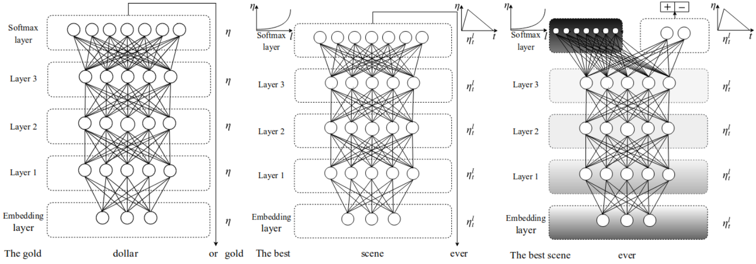
July 25, 2019

TEMPLE UNIVERSITY®

Introduction

- Motivation: short-text classification models typically require large labeled data
- **Hypothesis 1:** language models (LM) trained by self-supervised learning fine-tuned by domain-specific data require less labeled samples.
- **Hypothesis 2:** type and bias of additional data, used for self-supervised learning, can also hurt the performance.
- Objectives:
  (a) add news data to twitter posts on US elections and use them for self-supervised learning to test hypothesis 1.
  (b) test influence of news bias and additional data characteristics on the performance.

Model - ULMFiT



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

Universal language model fine-tuning for text classification (ULMFiT)[1]

[1] J. Howard, S. Ruder. Universal language model fine-tuning for text classification. arXiv:1801.06146, 2018.

Data

- Task is to classify **twitter data** (244,320 distinct posts) on US midterm elections 2018 into one of the three categories: left, right or neutral.
- To increase size of the corpus, additional data is used from six **news outlets** (Table 1).
- News articles discuss US election 2016 with different bias.

Table 1: Outlets

Outlet	Bias	#Words
CNN News (CNN)	left	426,778
Washington Post (WP)	left-center	9,229,176
BBC News (BBC)	neutral-left	1,247,437
MarketWatch (MW)	neutral-right	1,505,107
Wall Street Journal (WSJ)	right-center	547,548
FoxNews (FN)	right	3,082,912

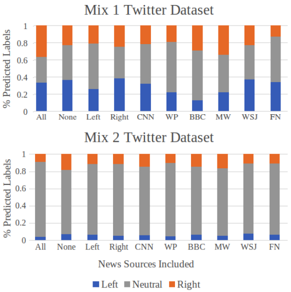
Experimental Settings

- General corpus: 103 million tokens from Wikipedia (WTM103) for LM pre-training.
- Discriminative corpus: 10 combinations of news data (0.5 - 16 millions of words) with different biases used together with 244,320 twitts (~ 4 millions of words) for LM fine-tuning.
- Classifier fine-tuning corpus: Mix1 or Mix2 data of 1,026 and 1526 labeled tweets.
- Validation and test data: Another 200+200 labeled tweets
- All experiments are repeated four times and average and standard deviation (stdev) are reported.

Results (1)

Table 2: Classification results

News sources included (Left : Neutral : Right)	Mix 1 (380 : 323 : 323)	Mix 2 (380 : 823 : 323)
All news	$53.2 \pm 3\%$	$59.4 \pm 3.7\%$
No news	$56 \pm 5.3\%$	$\mathbf{66.6 \pm 2.5\%}$
Left-biased (CNN+WP+BBC)	$49.2 \pm 2.9\%$	$61.1 \pm 3.3\%$
Right-biased (MW+WSJ+FN)	$51.7 \pm 3.8\%$	$63.0 \pm 3.2\%$
CNN	$\mathbf{58.7 \pm 1.2\%}$	$62.7 \pm 3.0\%$
Washington Post (WP)	$55.6 \pm 3.0\%$	$60.7 \pm 1.4\%$
BBC	$55.1 \pm 3.1\%$	$64.1 \pm 2.7\%$
MarketWatch (MW)	$56.5 \pm 2.6\%$	$64.2 \pm 1.8\%$
Wall Street Journal (WSJ)	$57.7 \pm 3.7\%$	$60.0 \pm 4.3\%$
FoxNews (FN)	$53.2 \pm 2.9\%$	$61.9 \pm 3.3\%$



- When Mix 1 and Mix 2 results are compared, the model always achieved better results for Mix 2 (Table 2) which has 54% of neutral labels as compared to 31.5% of neutral labels in Mix 1.
- $80 - 90\%$ of predicted labels for Mix 2 are neutral.

TEMPLE UNIVERSITY

Results (2)

- The classification accuracy difference between Mix 1 and 2 is the largest (11.9%) when "left-biased news" is used.
- Using "all news" data for fine-tuning achieves the best balance among predicted labels for Mix 1. However, almost half of predicted labels are wrong, so accuracy is low.
- The confusion matrices of experiments reveal that model recognizes the right label easier than the left label in Mix 2.
- Different influence of biased news is notable. In Mix1 between the best and the worst accuracy for different fine-tuning settings is 9.5%. In Mix 2 this difference is 7.2%.
- Influence of the bias is not uniform and it depends on other text properties (structure, jargon use, bias sensitivity).

TEMPLE UNIVERSITY®

Conclusions

- High stdev $(1.2 - 5.3\%)$ indicates the model's sensitivity to the number of labeled examples.
- Model is not robust to unbalanced datasets.
- Better results for Mix 2 are achieved because the algorithm exaggerates the most frequent (neutral) label in the imbalanced dataset (which contains 54% of examples of that class).
- Labeled Twitter data demonstrate diversity among posts with label "left". They often talk about one particular issue and have fewer hashtags to support the left political spectrum.
- Fine-tuning with biased news influences accuracy in both ways.
- The size of the fine-tuning data does not influence the results.

Summary

- In some cases, UMLFiT barely learns anything indicating a need for more labeled data and better fine-tuning dataset.
- All outlets try to appear neutral. Outlet bias labels come from experts following the outlet through time. Bias of individual articles can vary enormously.
- Using raw domain-data for fine-tuning can influence results in unpredictable ways. Domain-data has to be carefully selected and accommodated to the task.
- In the extension of this work, we want to understand how performance depends on the size of labeled data and what are the properties of good fine-tuning dataset for twitter classification.