Biased News Data Influence on Classifying Social Media Posts

Marija Stanojevic, Jumanah Alshehri, Eduard Dragut and Zoran Obradovic

Center for Data Analytics and Biomedical Informatics (DABI), Temple University, Philadelphia, Pennsylvania, USA

{marija.stanojevic, jumanah.alshehri, edragut, zoran.obradovic}@temple.edu

1. Abstract

A common task among social scientists is to mine and interpret public opinion using social media data. Scientists tend to employ off-the-shelf state-of-the-art short-text classification models. Those algorithms, however, require a large amount of labeled data. Recent efforts aim to decrease the compulsory number of labeled data via self-supervised learning and fine-tuning. In this work, we explore the use of news data on a specific topic in fine-tuning opinion mining models learned from social media data, such as Twitter. Particularly, we investigate the influence of biased news data on models trained on Twitter data by considering both the balanced and unbalanced cases. Results demonstrate that tuning with biased news data of different properties changes the classification accuracy up to 9.5%. The experimental studies reveal that the characteristics of the text of the tuning dataset, such as bias, vocabulary diversity and writing style, are essential for the final classification results, while the size of the data is less consequential. Moreover, a state-of-the-art algorithm is not robust on unbalanced twitter dataset, and it exaggerates when predicting the most frequent label.

2. Motivation

In recent years, social media platforms have become leading channels for the exchange of knowledge, debates, and product or opinion advertising. Thus, we need systems that classify data with limited human involvement.

Hypothesis: using recent language models (LM), which allow self-supervised learning and additional domain-specific data, can improve classification with small labeled datasets.

Hypothesis 2: type and bias of additional data can also hurt the performance.

Pre-trained Language Model → Fine-Tune on new dataset → Text Classifier

3. Experiments

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

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.

Six news outlets texts, discussing US election 2016 with different bias, are used (Table 1).

We used WTM103 pre-trained language model (103 million tokens from Wikipedia [1]).

For the fine tuning, ten combinations of twitter and news data are used.

In the final step, labeled twitter data are classified.

There are two twitter labeled data mixes: with labels ratios left:neutral:right (380:323:323) and (380:823:323), respectively.

Both test and cross-validation have 200 examples for each mix.

7. References

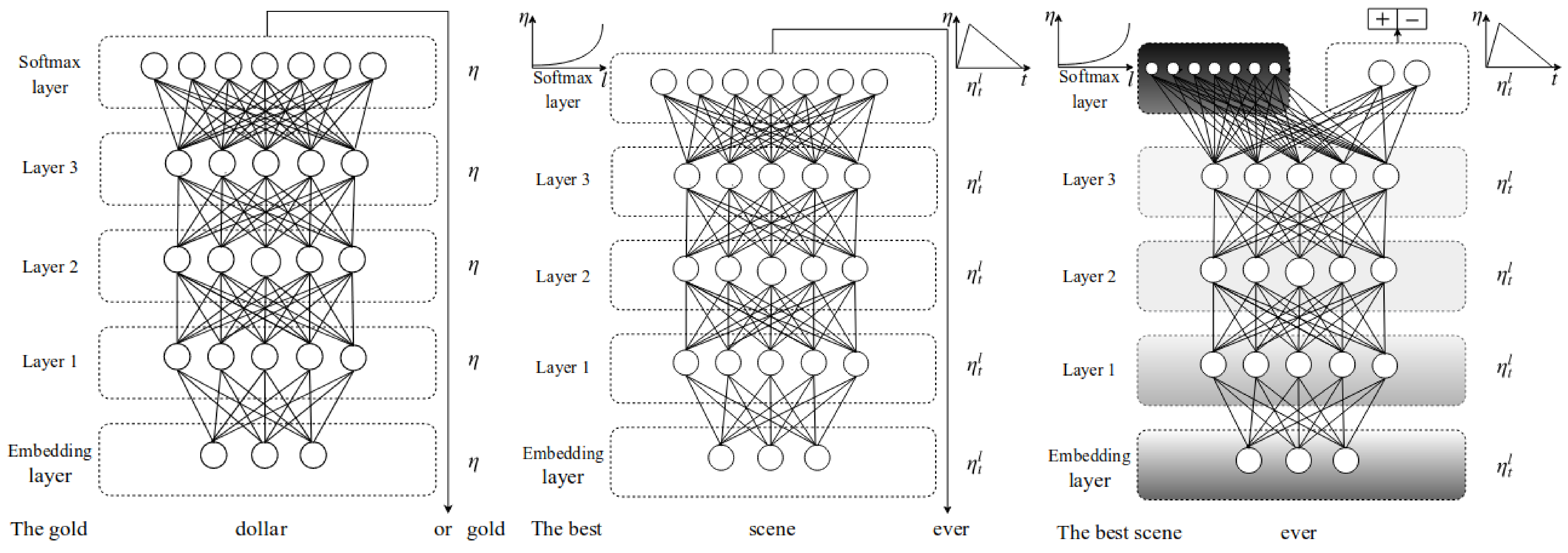
[1] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.

[2] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

4. UMLFiT Model

The ULMFiT model [2] consists of three training components:

- a) LM pre-training which learns vectors for word embedding using general text (e.g. Wikipedia, books);
- b) LM fine-tuning model which modifies embedding based on contexts in domain-specific text;
- c) classifier which learns how to classify texts based on the small labeled dataset and word embeeding.



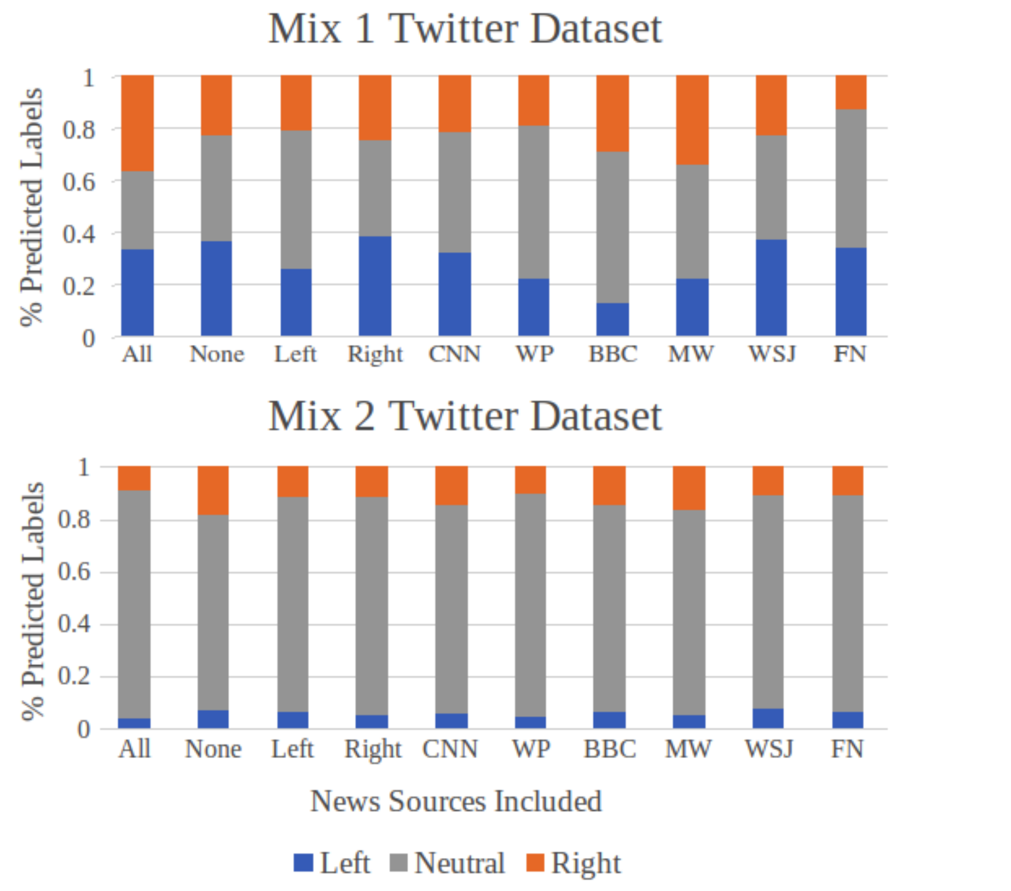
(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

5. Results

- All experiments are repeated four times. Average and stdev are reported in Table 2.
- 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.
- As evident from figures below, 80 − 90% of predicted labels for Mix 2 are neutral.
- The classification accuracy difference between Mix 1 and 2 is the largest (11.9%) when "left-biased news" is used for fine-tuning.
- Mix 1 figure reveals that 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 created for each experiment reveal that model recognizes the right label easier than the left label in Mix 2.

Table 2: Classification results

News sources included	Mix 1	Mix 2
(Left : Neutral : Right)	(380 : 323 : 323)	(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)	$53.2 \pm 2.9\%$	$61.9 \pm 3.3\%$



6. Conclusions

1. High stdev (1.2 − 5.3%) indicates the model's sensitivity to the number of labeled examples.
2. Results provide evidence that the model is not robust to unbalanced datasets.
3. 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).
4. 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.
5. Fine-tuning with biased news datasets can influence accuracy in contrasting ways.
6. The size of the fine-tuning data does not influence the results.