Detecting Bias in News Article Content with Machine Learning

Nathan Martindale

Introduction

Concern has grown over biased and unreliable news in recent decades. The internet allows for the rapid spread of misinformation, which is potentially damaging to society. A tool that can automatically detect and notify users of potentially biased content would be useful to help combat this spread.

Our work explores this area by applying natural language processing and machine learning to label bias and reliability of news articles based on their content, using weakly supervised learning.

Weak Supervision

- Supervised learning refers to data with associated labels and training an algorithm to predict those labels.
- In fully supervised learning, labels are all correct. However, various problems can exist with the labels in which it becomes a weakly supervised approach:
  - Incomplete - not all data is labeled.
  - Inaccurate - some labels are incorrect.
  - Inexact - only coarse labels exist for a fine-grained labeling problem.
- We deal with labels about news sources (coarse-grained) and try to predict labels for individual news articles (fine-grained), and so this is an inexact weak supervision problem.

Data

- Data is difficult, expensive, and time consuming to obtain in this domain due to the subjective aspects of bias and overwhelming amount of content. There are very few datasets of individually labeled articles.
- Primary dataset: NELA (NEws LAndscape). This dataset contains over 700K scraped articles from 180 different news sources.
- NELA also contains bias and reliability labels about news sources from eight different assessment sites.
- Assessment sites include sites like AllSides, NewsGuard, Media Bias/Fact Check, and more.
- Additional dataset used for testing from Media Bias Chart. This set includes 1600 articles each individually labeled with a bias and reliability score.

Source	AllSides	MB/FC	Media Bias Monitor	Combined
CNN	Left-center	Left	Center	Left
Reuters	Center	Center	Left-center	Center
Drudge Report	Right-center	Right	Right	Right
....	...	...	...	...

Table 1: An example of bias labels from different assessment sites, including the combined set created via voting mechanism.

Dataset Preparation

- Used NELA assessment site labels to assign proxy labels to each article. For example, if a particular source is labeled as left-biased, every article from that source is also labeled left-biased.
- ~20,000 articles split into 10 folds for 10-fold CV. Every article from any given source in a single fold. Validation results are thus on articles from unseen sources.
- Resulting model tested on individually labeled articles.
- Different "selection sets" were created by varying which set of labels were used – e.g. using AllSides labels for CNN instead of Media Bias Monitor's. (See Table 1.)
- A combined selection set was created by voting between the assessment sites.
- Word embeddings were created for each article, in two different formats: sequence and aggregate. With sequence data, every article was represented as a series of 300 dimensional word vectors. With aggregate data, these vectors were averaged to create a single 300 dimensional vector.

Classification Problems

- Reliability (reliable or unreliable)
- Bias (biased or unbiased)
- Bias direction (left, center, right)
- For each problem above, multiple approaches were tested, including different word embeddings (Word2Vec, GloVe, FastText) and different machine learning models (support vector machine's, neural networks, and LSTMs.)

Using Validation Set Results

- Initial results on validation set data yielded dramatically varying per-source accuracies. On some sources, the model only correctly predicted 5%, on others up to 97%.
- On problems where the target is a binary label, this indicates that on 5% accuracy results the model is confident the articles from that source should be labeled differently.
- We tested flipping the label on sources with under 25% accuracy in an attempt to increase internal consistency.
- This does not pollute the results as the validation and testing sets are unrelated.
- As shown in Table 2, the combined selection set with the flipped labels does perform better than the combined set by itself. Note that the displayed accuracies are on the testing set data, rather than the validation sets.
- This shows that an incorrectly labeled source can damage accuracy, and that using validation set results to create better internal consistency may create models that generalize better.

Selection set	Accuracy
AllSides	67.5%
MBM	68.8%
Combined	67.5%
Combined/Flipped	69.9%

Table 2: Bias prediction accuracies (using an SVM) with different selection sets

Results

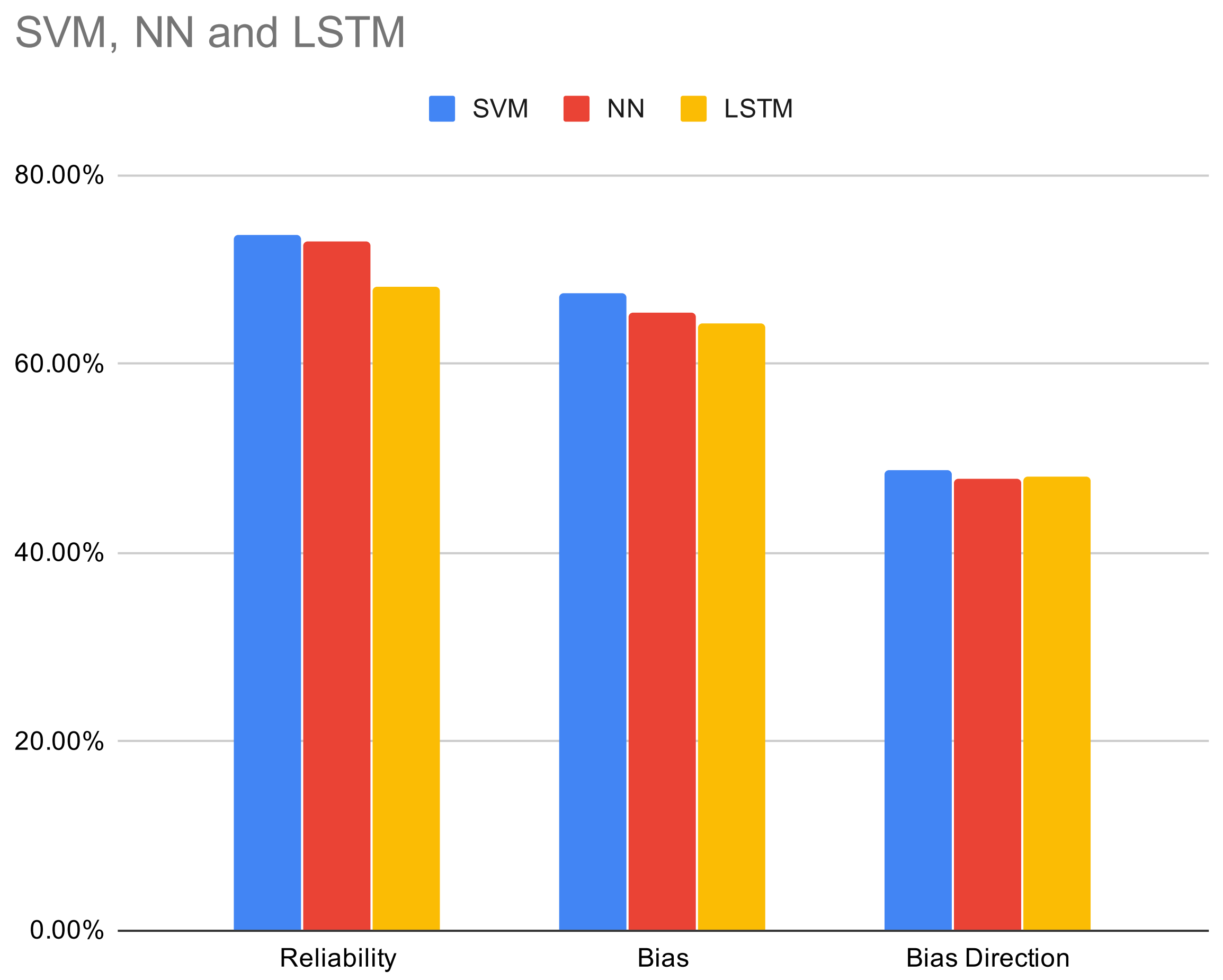

Figure 1: Performance differences between model types

- As shown in the algorithm comparisons in Figure 1, better than random results were achieved on each problem.
- In general, the aggregate data (SVM, NN) performed better than using sequence data.

	Trained on source-level labels	Trained on article-level labels
Reliability acc	73.6%	78.8%
Bias acc	69.9%	72.1%
Bias direction acc	53.5%	64.4%

Table 3: Training with weak supervision versus training with full supervision.

- We tested this weak supervision approach against simply training and testing on the 1600 individually labeled articles, or fully supervised learning. As shown in Table 3, all fully supervised learning approaches perform at least 5% better.
- While full supervision produces higher accuracies, the dataset for it is much harder to acquire, and similarly difficult to update over time. In contrast, using proxy labels, any new article published by a labeled news source can immediately be used as new labeled data.

Conclusion

- Weak supervision is a potentially viable approach to predicting bias of news articles.
- More work needs to be done to achieve higher accuracies than using full supervision.
- Future work could look into strategies for using individually labeled articles to help correct proxy labels.