Detecting Bias in News Articles using NLP Models

Muhammad Umar Nadeem and Sarah Raza
[munadeem, sraza007]@sanford.edu

Problem/Background

This project analyses various natural language processing (NLP) algorithms in order to build a deeper understanding of the machine learning techniques required to detect biased political leanings in news sources. Since news is the first and most direct source from which people learn about current unfolding events, the introduction of unjust subjectivity can be very harmful. Thus, the ability to detect bias in news sources is key to maintaining truth when disseminating information. The first classification model we implement is a Tensorflow deep neural network (DNN) using bag of words (BOW) to represent the input sentences. Then, we build on this deep learning algorithm by adding term frequency–inverse document frequency (TD-IDF) as a weighting factor to the DNN input data. In an attempt to further improve results, we shift towards an unsupervised K-Means clustering algorithm to analyze patterns discovered amongst articles from the various news source. Finally, we implement SimCSE, a contrastive learning framework for sentence embeddings. We find that contrastive learning is the most accurate NLP model of those tested for detecting the nuances of political bias in news article sentences.

Methods

Data

The datasets we are using for training and testing our algorithm are Wei's NewB news source sentences about former United States President Donald Trump [3]. For training data, Wei has compiled approximately 250,000 sentences from 11 news sources, five liberal sources (Newsday, New York Times, CNN, LA Times, Washington Post), one neutral source (Politico), and five conservative sources (Wall Street Journal, New York Post, Daily Press, Daily Herald, Chicago Tribune) into a .txt file that labels each sentence with the corresponding source. We preprocessed this training data using pandas dataframes into a .csv file, and then extracted 1000 sentences from each source. We stored these 11,000 training sentences as eleven lists of strings that we inputed as JSON data to an nltk tokenizer. Similarly, for the testing data, Wei has compiled approximately 11,000 sentences from the same 11 news sources. We preprocessed this testing data using pandas dataframes into a .csv file, and then extracted 100 sentences from each source. We stored these 1100 training sentences as eleven single lists of strings that we inputted as labeled dictionaries to our evaluation function.

Approach

1. Bag of Words & TF-IDF - Tensorflow DNN with a Softmax activation function and an Adams optimizer. Our batch size is 8 and there are 4 layers in our DNN model. To then implement TF-IDF, we maintained the same model and model configurations, just replacing the input with TF-IDF weighted, vectorized news source sentences.

2. K – Means - K-Means model trains on TF-IDF weighted, vectorized news source sentences. We trained the model first with 11 clusters in hopes of identifying 11 unique patters for 11 unique sources. We then used the Elbow Method to find the optimal number of clusters, namely 2

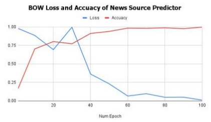
3. Supervised SimCSE - We use an Adam optimizer with warmup steps and a linear learning rate scheduler. The parameters for our model are batch size is 256, number of epochs is 3, learning rate for training is .0001, and learning rate for testing is .00005.

Experiments

Bag Of Words (Baseline)

Our baseline model correctly classifies news source sentences to their respective new sources with an average accuracy of 15.6%

News Source	Baseline (%)
Newsday	19.2
NYT	22.6
CNN	15.6
LA Times	12.2
WashPost	12.8
Politico	13.8
WSJ	14.2
NYPost	22.6
DailyPress	10.6
DailyHerald	21.6
ChicagoTribune	9

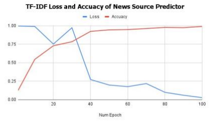

BOW Loss and Accuacy of News Source Predictor

TF-IDF

Our DNN with TF-IDF as a weighting factor for the input vectors correctly classifies news source sentences to their respective new sources with an average accuracy of 15.1%

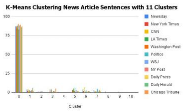
Baseline vs TF-IDF Prediction Accuracy

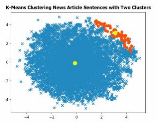
News Source	Baseline (%)	TF-IDF (%)
Newsday	19.2	18.6
NYT	22.6	23
CNN	15.6	11.6
LA Times	12.2	9.6
WashPost	12.8	11.2
Politico	13.8	12.6
WSJ	14.2	16.2
NYPost	22.6	19.2
DailyPress	10.6	11.6
DailyHerald	21.6	21.4
ChicagoTribune	9	10.6


TF-IDF Loss and Accuacy of News Source Predictor

K-Means

Our K-Means algorithm was not effective in finding patterns amongst article sentences from the same source or even amongst article sentences across different sources


K-Means Clustering News Article Sentences with 11 Clusters

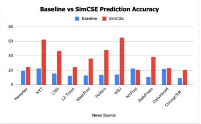

K-Means Clustering News Article Sentences with Two Clusters

SimCSE

Our SimCSE model classifies news source sentences to their respective new sources with an average accuracy of 37.2%

Baseline vs SimCSE Prediction Accuracy

News Source	Baseline (%)	SimCSE (%)
Newsday	19.2	24.3
NYT	22.6	62.5
CNN	15.6	46.8
LA Times	12.2	24.1
WashPost	12.8	36.3
Politico	13.8	48.3
WSJ	14.2	64.9
NYPost	22.6	20.5
DailyPress	10.6	38.7
DailyHerald	21.6	22.8
ChicagoTribune	9	20.2


Baseline vs SimCSE Prediction Accuracy

Analysis

Our analysis of four different NLP algorithms for bias detection in news source highlights a couple key insights regarding the machine learning techniques required to detect biased political leanings in news sources. First, we see that TF-IDF as a feature scaling method does not improve or even really change the classifier's accuracy above our BOW baseline DNN. In this case uniform column scaling was not helpful in any way due to the fact that there were no "informative words" or "common words" that stood out amongst the different new article sentences. All the sentences, in training and testing data and from all sources, are all versions of someone talking about Donald Trump. Therefore, the vocabulary is very very similar across all sentences, and word vectorization is unable to pick up the nuances of political bias in how those words were ordered, phrased, or emphasized. This same reasoning regarding word vectorization can extend to K-means and its inability to find patters amongst sentences from the same or varying sources. Second, we see that SimCSE outperforms all the other models. This tells us that contrastive based learning is the most successful for detecting the nuances between different news sources. This likely because SimCSE is the only model taking into account factors such as framing rather than just word choice because it considers the whole sentence.

Conclusions

It is clear that sentence embeddings and overall greater contextual understanding are necessary to detect the nuances of political bias and leaning in published news articles. Simple word embedding models such as BOW with TF-IDF are not successful in capturing varying framings of similar topics because the overall vocabulary of each news source does not have significant variance. Similarly, an unsupervised model such as K-Means with word embedding input vectors is not able to find patterns across articles from a specific news source or across news sources because of a limited variance in vocabulary. The limitations of our work are the specific news sources from which we obtained our training and testing data, and well as the computing power required to train and test on entire articles rather than sentences. As follows, future avenues of work include expanding this research for a wider range of online news sources, and scaling the models to larger input embeddings.

References

[1] Premanand Ghadekar, Mohit Tilokchandani, Anuj Jevrani, Sanjana Dumpala, Sanchit Dass, and Nikhil Shinde. Prediction and classification of biased and fake news using nlp and machine learning models. In Debabala Swain, Prasant Kumar Pattnaik, and Tushar Athawale, editors, Machine Learning and Information Processing. Springer Singapore, 2021.
[2] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." arXiv preprint arXiv:2104.08821 (2021).
[3] Wei, J (2020) NewB, Github Repository. Software available from github.com/JerryWei03/NewB.