Automated political bias classification in news agencies: a word-based feature selection approach

Anonymous ACL submission

Abstract

This study offers a new solution to the problem of developing political bias classification models in news agencies. Our method uses search engine score functions to develop a measure of the relevance of each word in text scrapped from news websites. With these scores, we train models using existing feature selection methods and a custom feature selection algorithm that we developed. The resulting models are contrasted with each other and neural network-based counterparts. Models trained using our proposed method and custom algorithm outperformed others by achieving macro F1 scores of 0.81 and 0.78 on right-wing and left-wing bias detection respectively.

1 Introduction

The ever-increasing popularity of online news platforms and the ease of publishing news online, have created a demand for automated political bias detection. This study addresses this need by crafting eight data sets based on expert labeling of data that are free of manual trimmings of texts. Furthermore, it proposes quantitative measures for assessing the role of different words in the documents and introduces a variety of new model training approaches.

The data sets are built by compiling the first 12 articles appearing on the U.S politics page of each of the 78 news agencies along with the titles of the articles appearing on the same section of the site on a given day. The type of bias of each news agency is labeled using data from Ad Fontes Media[1], which uses a body of experts to label the political bias of news agencies. The three search engine ranking functions of Okapi BM25, tf–idf and a Divergence from randomness model (DFR in short) were used to develop a measure of the relevance of each word. Using this information, the study examines methods to detect bias inducing words through a mix of conventional feature selection methods such as f-regression and PMI, and also, a custom feature selection algorithm that makes use of the cross validation score in the training data. Finally, various models are trained, tested and contrasted with each other. The method proposed in the paper is also compared against deep learning methods by conducting a multi-label text classification experiment using the RoBERTa (Liu et al., 2019) network model.

This study offers three new contributions to the problem of automated political bias classification:

- Creation of expert-labeled data sets that make use of search engine score functions to quantify the relevance of words.
- A new word-based feature selection algorithm which chooses meaningful features by making use of the relevance scores of each word to calculate cross validation scores.
- An evaluation of bias classification, which empirically demonstrates that our feature selection outperforms existing counterparts.

2 Related Work

Linguistic patterns that introduce political bias have been examined including the use of individual words (Chen et al., 2020). There has also been research conducted to identify bias-inducing words (Spinde et al., 2021). However, much work remains to be done in order to systematically and quantitatively study the appearance of such words. We seek to fill this gap by offering a new approach.

There has been some progress towards identifying bias in individual news articles. One previous study offered such a system by examining potential word-choices for a single semantic concept within an article(Hamborg et al., 2019), other studies have automatically extracted information from variety of sources to detect political bias in individual news articles (Baly et al., 2020). A lot of attention has been

[1] https://adfontesmedia.com/

given to fact verification problems and challenges such as the Fake News Challenge[2] and workshops such as FEVER[3] seek solutions to such problems. However, little attention has been given to bias detection within news sources based on a collection of articles which is the main focus of this study.

3 Approach

Since different themes are discussed in right biased and left biased media (Carlisle, 2005), vocabulary that might be a good bias predictors for one category might not be a great choice for the other. Therefore, left-wing and right-wing bias classification are examined separately in a binary manner with a news media either belonging to one of these categories or not. In order to quantify the relevance of each word in the data collected from the news agencies, probabilistic and bag-of-words search engine scoring functions are used and contrasted to produce such measures. Finally, we introduce a new feature selection strategy to select meaningful features for bias classification. We train multiple models using both out feature selection strategy and a variety of conventional ones. The resulting models are then evaluated and contrasted with each other and against deep learning methods using the macro F1 score as an evaluation metric.

3.1 Data set creation and train/test split

Between the arbitrary dates of January and March 2022, jsoup[4] was used to extract information from news agency web sites. For each news agency, the politics page URL of the website was inputted to jsoup, creating a .txt file consisting of the headlines of the day. Furthermore, the URL of the first 12 articles appearing on the website were given as input to jsoup, creating 12 additional .txt files for each website. This process was repeated for 78 news agencies. These documents were labeled as right-wing biased, left-wing biased, and non-biased according to the labels from Ad Fontes Media.[1] The software package CoreNLP[5] was used for tokenizing and lemmatizing the words from the .txt files. Apache Lucene[6] with the scoring functions of Okapi BM25, tf–idf and a DFR model were used to measure the relevance of each word in our index. The index is a searchable compilation of first 12 news articles plus the headlines of the U.S politics page on a given day for 78 news agency; in other words it is a compilation of 936 .txt files containing news article content and 78 .txt files containing headlines. The relevance scores are then written into a .CSV file. The final result is six data sets for the total combinations of three scoring functions and the two categories of left-wing bias and right-wing bias. In each of these data sets the rows are the news agencies. For every word that has appeared in out index, there is a column representing relevance scores across news agencies for that word. The table below is an example of the structure of the data set that displays the relevance of the words "donate" and "us" in the appropriate cells for the news agency "nytimes".

news agency	donate	us
nytimes	0	0.20077

In order to contrast the methods proposed in this paper with deep learning methods, two data sets for training multi-label classification networks were created in addition to the six data sets mentioned above in each category of right-wing bias and left-wing bias. For each of these data sets, the 12 .txt files containing news articles and the one .txt file containing the headlines for each news agency are annexed and presented in the row belonging to each news agency, with each row representing a distinct news agency. The structure of this data set for the news agencies called "nytimes" can be seen in the table below. The Column with the label "Text" represents the compiled texts for that news agency.

news agency	Text
nytimes	...would supplant the rule of law with...

All data sets have the same set of test and train news agencies for coherency. The splitting process was done using the split function available in the scikit-learn library[7] for all data sets with a split of 30 percent. For the three data sets that make use of the scoring program, the indexes were created after the train/test split and the relevance scores were calculated and input-ed separately for the train and test data sets. This eliminates possible corruptions introduced by the relevance scoring program. The train and test data are balanced in terms of the distribution of non-biased, left-wing and right-wing media with both having roughly 0.33 left wing biased media, and 0.24 right-wing biased media.

[2] http://www.fakenewschallenge.org/
[3] https://fever.ai/
[4] https://jsoup.org/
[5] https://stanfordnlp.github.io/CoreNLP/
[6] https://lucene.apache.org/
[7] https://scikit-learn.org/stable/index.html

3.2 Models and Feature selection

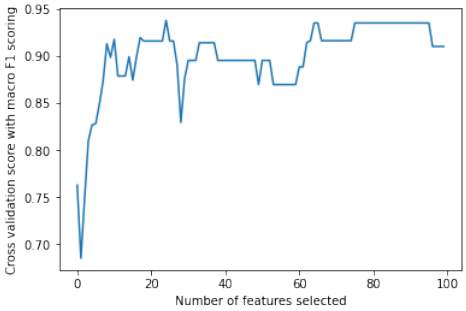
Logistic regression, random forests, gradient boosting classifiers and MLP neural networks were trained for the models making use of the relevance scores on all six data sets of such kind. A RoBERTa network model was trained on the two data sets containing compiled document texts. The conventional feature selection methods PMI and f-regression were used to train relevance-score based models. A custom feature selection algorithm was also developed for the same purpose. This algorithm was unable to extract features in the data sets making use of the tf-idf scoring function due to the values being too small, however, produced interesting lists of words in the other four data sets. This algorithm uses a logistic regression model to calculate the cross-validation score of a classifier using every word appearing in the columns of each data set as potential features. Each word that results in an F1 score of above 0.4 on the class of news agencies that have bias is stored in an array. The array is later sorted decreasingly based on the F1 scores. The pseudo-code for this algorithm is offered below with cvv function referring to the cross validation score calculation.

Algorithm 1: Custom algorithm

Input : A list $[a_i], i = 1, 2, \cdots, n$, where each element is a word from the data set.
Output : A sorted collection of words.

$features = [\,], scores = [\,]$
for $i \leftarrow 0$ **to** $n - 1$ **do**
  LRR()// new logistic regression model
  if cvv$(a_i, LRR) > 0.4$ **then**
    $features.append(a_i)$
    $scores.append($cvv$(a_i))$
  end
end
for $i \leftarrow 0$ **to** LEN$(features)$ **do**
  for $j \leftarrow 0$ **to** i **do**
    if $scores[j] > scores[i]$ **then**
      swap $scores[i]$ with $scores[j]$
      swap $features[i]$ with $features[j]$
    end
  end
end
return $features$

To estimate the number of features suitable for each combination of a model and feature selection algorithm, the cross validation scores with a macro F1 evaluation metric were calculated on a given training data set and plotted on a range of 1 to 100 of the top features extracted. For example, the figure below shows the plot for a logistic regression model trained on the top features extracted by the custom algorithm on the training left-wing bias data set that uses Okapi BM25.



A reasonable estimate of the number of features, should appear in a stable range of the plot and produce a high cross validation score. The plots were analyzed by selecting the lengthiest range among the top ranges producing the highest cross validation scores as a starting point for the number features we should use. The higher end of this range is selected as a potential list of features and the words containing names of companies and individuals not holding positions in government were dropped to avoid producing models that are dependant upon short-term political trends.

Using Hugging Face,[8] RoBERTa network models were trained on each of the two right-wing and left-wing bias data sets created for the purpose of training network models. The roberta-base model from Hugging Face was used as the tokenizer. The models were trained with 10 epochs, batch size of 64, learning rate of 2e-5 with AdamW as the optimizer.[9]

4 Experiments

The purpose of the study is to evaluate the efficacy of the classification methods that make use of relevance scores for training models, and to determine the best performing model training approach. Therefore, two type of comparisons need to be made: comparing the various relevance score dependant model training approaches with each other and contrasting the network models with relevance-score based models. In order to compare the relevance score-based models, experiments using all words in the data set columns as features were used as a baseline to evaluate the efficacy of the different feature selection approaches. The best performing models from this category are then contrasted with

[8] https://huggingface.co/

[9] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

the network models. The models trained for left-wing bias classification and right wing-bias classification are only compared with other models trained for the same purpose. Given that each experiment is either a binary classification of left-wing vs not left-wing or right-wing vs not right-wing, the macro F1 score on the test data is used as an evaluation metric of a model. The micro F1 score is not informative due to the binary nature of the classifications and the experiments containing about 0.7 news sources of just one label.

5 Results

The RoBERTa network models achieved a macro F1 score of 0.38 and 0.40 on left-wing and right-wing bias data sets respectively. The tables below show the results of our experiments for the models trained on relevance scores. All scores in the tables are macro F1 scores. The "LR" column represents the logistic regression model, similarly, "MLP" represents MLP model, "RF" represents the Random forest model, and "GB" represents the Gradient boosting model.

feature selection	relevance scoring	**LR**	MLP	RF	GB
all words	Okapi BM25	0.40	0.48	0.38	0.62
f-regression	Okapi BM25	0.60	0.68	0.70	0.53
PMI	Okapi BM25	0.59	0.47	0.53	0.57
custom	Okapi BM25	0.68	0.77	0.59	0.74
all words	DFR	0.37	0.40	0.66	0.59
f-regression	DFR	0.66	0.56	0.73	0.53
PMI	DFR	0.44	0.48	0.53	0.38
custom	**DFR**	**0.78**	0.62	0.63	0.78
all words	tf-idf	0.40	0.62	0.38	0.64
f-regression	tf-idf	0.70	0.70	0.67	0.53
PMI	tf-idf	0.40	0.68	0.67	0.56

Table 1: Left-wing bias detection.

feature selection	relevance scoring	LR	**MLP**	RF	GB
all words	Okapi BM25	0.45	0.45	0.45	0.45
f-regression	Okapi BM25	0.60	0.68	0.70	0.53
PMI	Okapi BM25	0.59	0.47	0.53	0.57
custom	Okapi BM25	0.68	0.77	0.59	0.74
all words	DFR	0.67	0.20	0.45	0.43
f-regression	DFR	0.44	0.62	0.45	0.45
PMI	DFR	0.47	0.39	0.58	0.42
custom	**DFR**	0.75	**0.81**	0.45	0.45
all words	tf-idf	0.45	0.45	0.45	0.55
f-regression	tf-idf	0.60	0.67	0.70	0.53
PMI	tf-idf	0.59	0.50	0.58	0.49

Table 2: Right-wing bias detection.

6 Discussion

Right-wing and left-wing classification models showed similar patterns; All models trained on relevance-based data sets outperformed their network based counterparts. Among the relevance-based models, it can be seen that models which used all words stored in the data set columns as features had poor performances as they mostly classified all news agencies as being non-biased. Models using the PMI method also displayed a poor performance for the same reason. The type of the scoring function used made no difference in the performance of these models. The low scores can be explained by the absence of any specific patterns among the words selected as features for these models. In the absence of the custom feature selection algorithm, the f-regression models produced the best results among models built using the tf-idf scoring functions. They also had the second highest F1 scores when the Okabi BM25 and DFR search functions were used. The custom feature selection algorithm produced the best overall results among the models using the DFR and Okabi BM25 scoring functions, and also, the best overall results in both right-wing and left-wing bias classification. The DFR based models had a slightly better performance compared to their Okabi BM25 based models. The small difference does not seem particularly meaningful. The successes of these methods can be attributed to the clear patterns seen among words selected as features. In the left-wing category the custom feature selection algorithms produced a list of 87 words capturing the themes of social justice, empathy, minority rights and anti violence often seen in U.S left (Carlisle, 2005). The words "plea", "workers", "nonprofit" , "guns", "muslim", "activist" are some examples of such words. In the right-wing category the custom feature selection algorithms produced a list of 86 words capturing the themes of abortion, liberty, personal responsibility, religious views on sexual matters and conspiracy theories seen in the U.S right wing politics (Carlisle, 2005). The words "virtues", "lgbt", "pathetic", "freedoms", "liberty", "mothers", "churches" and "hoax" were seen among the features.

7 Conclusion

This work offered new systematic approaches for right-wing and left-wing political bias classification in U.S news agencies. The best results were achieved when the models were trained on the relevance scores of words using our custom feature selection algorithm. There seems to be no major difference among the Okabi BM25 and DFR scoring functions for this task. The data and code produced in this work is publicly available at *URL hidden for the blind review*.

8 Limitations

As mentioned in the paper, the names of individuals not holding political office positions and companies were manually dropped among the features for training models to avoid the creation of models that depend on data from a specific interval of time. Collecting data over longer periods of time can automatically eliminate the effect of such features on our feature selection algorithms. Despite the paper producing new effective models with only 936 articles extracted from the websites and 78 lists of article headlines, increasing the number of articles per news agency and examining its affect on models was not thoroughly examined in the paper.

9 Ethics Statement

Only news agencies allowing web scrapping in their user-terms were used for this study.

References

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Rodney P. Carlisle. 2005. Encyclopedia of politics : the left and the right.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.

Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019. Automated identification of media bias by word choice and labeling in news articles. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 196–205.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing I& Management*, 58(3):102505.

A Lists of words selected as features for the best performing models

The best performing model in both categories made use of the custom feature selection algorithm and the DFR scoring function.

A.1 List of right-wing features

'liberty', 'paid', 'woke', 'reagan', 'experts', 'freedoms', 'data', 'food', 'agency', 'pays', 'taxpayer', 'cbs', 'sanction', 'eastern', 'site', 'curriculum', 'jinping', 'xi', 'girls', 'remain', 'courage', 'sign', 'college', 'quo', 'dem', 'adverse', 'debunked', 'implementing', 'virtues', 'trucker', 'family', 'crises', 'company', 'sunshine', 'civilization', 'series', 'hoax', 'collusion', 'mothers', 'inspired', 'supports', 'healthy', 'keystone', 'excuses', 'lgbt', 'emergencies', 'heroes', 'repeatedly', 'confronted', 'protecting', 'build', 'mass', 'aoc', 'crushing', 'socialist', 'cases', 'conservatism', 'parenthood', 'effectiveness', 'ineffective', 'hearing', 'phone', 'dishonest', 'gettr', 'europeans', 'blasted', 'prison', 'swamp', 'beltway', 'grossly', 'huge', 'spirit', 'elites', 'babies', 'harvesting', 'damn', 'protected', 'panic', 'accomplished', 'pathetic', 'admin', 'denying', 'churches', 'moms', 'iran', 'conscience'

A.2 List of Left-wing features

'promise', 'actual', 'conspiracy', 'box', 'china', 'profile', 'knowing', 'plea', 'document', 'film', 'somehow', 'bernie', 'voice', 'named', 'theories', 'sole', 'crimes', 'misleading', 'guns', 'connection', 'fair', 'suggest', 'strange', 'emerged', 'letting', 'constitution', 'representation', 'organizing', 'remaining', 'jail', 'serves', 'nonprofit', 'doj', 'folks', 'violence', 'poverty', 'prize', 'solidarity', 'convince', 'muslim', 'primarily', 'sad', 'harassment', 'express', 'mail', 'formerly', 'permission', 'mention', 'incredible', 'complicated', 'prison', 'sources', 'worker', 'marshall', 'activism', 'broke', 'distributed', 'knowledge', 'adult', 'propaganda', 'appear', 'financially', 'profits', 'signing', 'walls', 'detailed', 'fundamentally', 'actors', 'bloody', 'detail', 'uncertainty', 'age', 'legitimate', 'investigate', 'donated', 'light', 'user', 'spoken', 'voter', 'uprising', 'gaining', 'entity', 'supremacy", 'notorious', 'existing', 'intercept', 'wildly'