Exploring the Relationship between News Source, Media Bias and Propaganda through Propaganda Detection

Ella Happel
University of Illinois Urbana-Champaign / 120 6th Avenue, La Grange IL
ehappel2@illinois.edu

Abstract

This paper attempts to create a propaganda detection classifier, trained on a sentence-level human-annotated dataset of news articles. It then explores the relationships between news outlets and predicted propaganda, and media bias and predicted propaganda. The two propaganda detection classifiers built are logistic regression and SVM models with TF-IDF. The logistic regression model performs slightly better, with an accuracy score of 0.75. After applying the logistic regression model to a new dataset of news articles, each news publication's proportion of articles predicted to contain propaganda is recorded, and there is no correlation observed between media bias and predicted propaganda. The logistic regression model is also tested on a separate dataset of annotated tweets, with an accuracy of 0.72.

1 Introduction

In an online world, propaganda has the potential to spread quickly, unchecked, infesting the minds of users through their phones and computers. It is vital to the preservation of democracy, prevention of misinformation, and fight against hateful movements that propaganda is able to be detected and prevented, especially within news sources. It is also of interest to identify news outlets that contain higher levels of detected propaganda within their articles, and to evaluate if there is any correlation between those with higher detected propaganda rates and media bias. This goal was inspired and motivated by the allsides media bias dataset, which reports media bias of news sources, as annotated by Americans. The hope is to create a propaganda chart by news source, and to explore if media bias has an impact on propaganda levels.

The training data was sourced from data collected for a shared task at the 2019 Workshop on NLP4IF, a conference focused on censorship, disinformation, and propaganda. We train and evaluate logistic regression and SVM classifiers with TF-IDF, finding that logistic regression slightly outperforms SVM, reaching an accuracy of 0.75.

Then, using the logistic regression classifier, we annotate a dataset of 3824 news articles collected from December 2016 to March 2017. We explore predicted propaganda proportions by news source, and if this relates to the sources' media bias.

2 Related Work

A comprehensive review of the conference shared task can be found in (Da San Martino et al., 2019).

3 Data Exploration

Our code and datasets are publicly available here.

The training corpus contains 500 articles from 48 news sources, manually annotated by 6 professional annotators. The annotators applied 18 propaganda techniques to the data:

- Loaded Language
- Flag-Waving
- Whataboutism
- Name Calling, Labeling
- Causal Oversimplification
- Reductio ad Hitlerum
- Flag-Waving
- Whataboutism
- Causal Oversimplification
- Reductio ad Hitlerum
- Slogans
- Red Herring
- Appeal to Authority
- Bandwagon
- Black-and-White Fallacy
- Straw Men
- Thought-terminating Cliches
- Obfuscation, Intentional Vagueness, Confusion

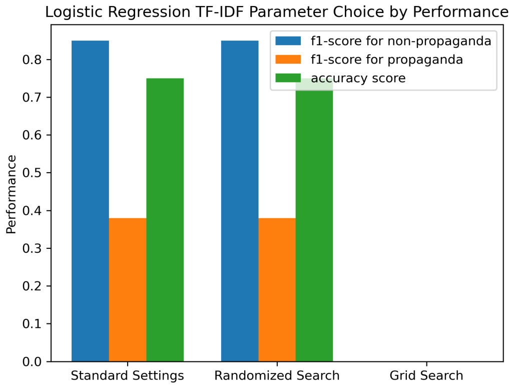

Figure 1: TF-IDF Parameter Performance for Logistic Regression

The annotators also conducted sentence level binary classification, annotating if each sentence contained propaganda. The data then presented two tasks: Fragment Level Classification, and Sentence Level Classification. For the remainder of this paper, we will focus on the ladder.

4 Classification

We applied both logistic regression and SVM to the annotated data, testing the classifier against a portion of its own dataset. These models were chosen due to their straightforward and traditional nature, as opposed to more complex choices, like BERT, as seen in many of the previously published papers on this shared task.

To clean the data, we first utilized two provided files from the dataset: createdataframeslc.py and createlabelsdataframeslc.py. These files loaded the annotated data into a pickle file, and we added an output of a csv file as well. We dropped na values, of which there were many due to empty lines in the original dataset. Then we replaced the values "propaganda" and "non-propaganda" with 1s and 0s.

For the logistic regression classifier, we split the dataset into training and test data, in an 80% training, 20% test split. Then we vectorized with TF-IDF, and trained it on the sentences from our dataset. Next, we predicted the test sets' annotations, and printed a classification report.

We first spent time working to find the best tf-idf parameters for the logistic regression model. This proved unfruitful, as our randomized search performed exactly the same as the standard parameters for the key values of: f1-score for sentences with no propaganda, f1-score for sentences with propaganda, and accuracy. Our grid search was unsuccessful due to the amount of time it took to run on our machine. Given these results, our final optimized model used the standard tf-idf parameters for logistic regression, as they are quickest and equally accurate to the randomized search.

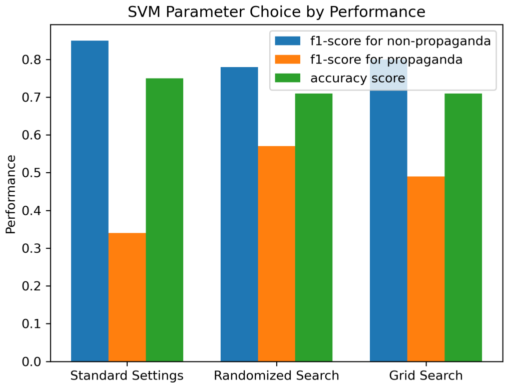

Figure 2: SVM Parameter Performance

	Logistic Regression	SVM
Standard Parameters	0.75	0.75
Randomized Grid Search	0.75	0.71
Grid Search	N/A	0.71

Table 1: Accuracy Levels of Parameter Search Methods for Different Models

For SVM, we again split the data along 80% training and 20% test. Due to the previous unhelpful results of optimizing TF-IDF features, we optimized for SVM features instead. However, this too was not helpful. Our randomized search and paired-down grid search significantly increased the propaganda f1-scores, but decreased the f1-scores of non-propaganda sentences. This may be due to the grid search's non-comprehensive nature, due to limited processing power. Overall, this actually meant a decrease in accuracy. This meant again, our standard parameters were the highest performing.

It is of note, that for both models the classifier was much better at classifying non-propaganda than propaganda. If the goal of the classifier was to optimize the f1-score for articles with propaganda, a model like SVM with randomized search parameters would be more ideal.

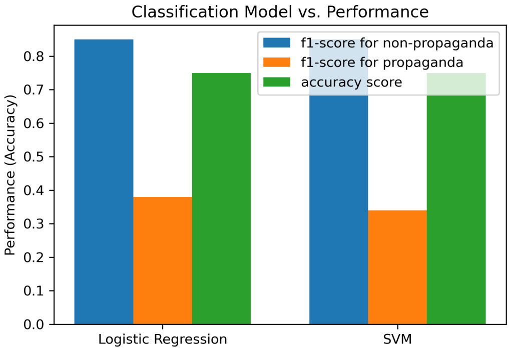

Figure 3: Optimized Classification Model vs. Performance

5 Application

The next dataset was found on Harvard dataverse. It is a dataset of 3824 news articles, with variables of publish date, title, subtitle and text. These articles were collected within the span of three months, from December 2016 to March 2017. They contain articles from ABC News, CNN news, The Huffington Post, BBC News, DW News, TASS News, Al Jazeera News, China Daily and RTE News.

Here we motivate future researchers; it was difficult to find a recent and reliable dataset of news articles. This is likely due to copyright issues. We encourage publication of more and updated news article databases for general use and research.

To clean the dataset, we discarded extra empty columns. Then we added a column indicating the source of the article, based on the url. We used spacy to split by sentence, and built a new dataframe holding each articles' source, full text, array of sentences, array of sentence labels (0 for non-propaganda, 1 for propaganda), and document label (also 0 for non-propaganda, 1 for propaganda). A document is labeled propaganda if any sentence within it is detected as containing propaganda.

Our first task of evaluation was to find the propaganda level by source. We were able to create a visualization of each news source vs. the proportion of articles in the dataset labeled as "propaganda." Here, we see that Al Jazeera and CNN lead in proportion, with around 61% of their articles being labeled as containing propaganda. Tass has the lowest proportion, at around 0.2. Overall, we find proportions ranging throughout 0.2 to 0.6, with no large outliers.

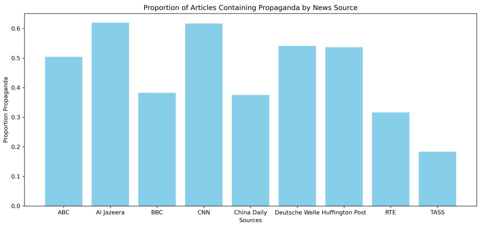

Figure 4: Proportion of Propaganda Articles by Source

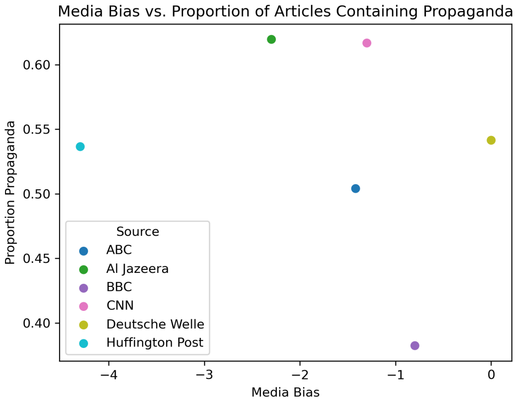

Figure 5: Media Bias vs. Proportion of Propaganda Articles

Then we applied the "allsides" media bias ratings[1] to the dataset, labeling each news source with their bias. Negative values indicate left-leaning and positive right-leaning. Here, we had to omit China Daily, RTE News, and TASS, as there was no available information on their media bias. We plotted the Media Bias against proportion propaganda. There seemed to be no apparent correlation between these values, as the data points were spread throughout the cooridinate field. It is of note that with more sources, and particularly right-leaning sources, it is possible we may find a correlation.

References

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China. Association for Computational Linguistics.

Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. 2024. HQP: A human-annotated dataset for detecting online propaganda. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6064–6089,

[1] https://www.allsides.com/media-bias

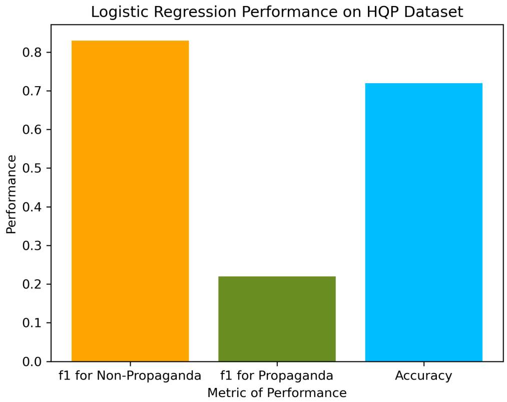

Figure 6: Performance of Logistic Regression Classifier on HQP Dataset by Metric

Bangkok, Thailand. Association for Computational Linguistics.

Limitations

There are many limitations to this analysis. Both the annotated and non-annotated news datasets are older, created in 2019 and 2016/2017. The model may be less effective on current data. The annotated data was also annotated by “professional annotators.” It is unclear what qualifies these annotators. It is also important to note that these professionals may not share the opinions or views of the general public. Whether or not a sentence contains propaganda can be viewed as subjective, and so can be vastly different between people, even if there is a scholarly objective consensus.

We also achieved an accuracy score of 0.75 for our logistic regression and SVM models. This means there is significant room for error in our annotations.

A Additional Testing on Tweets

We also tested our logistic regression model on a new, relevant register of tweets. Tweets, and social media at large, are also large sources of information for people, alongside traditional news articles. Propaganda detection in these contexts is now in high demand, and would be highly impactful. For this testing, we used HQP, a human-annotated corpus of 30,000 tweets, annotated by non-professionals. The creation of this dataset is documented extensively in (Maarouf et al., 2024).

To clean the dataset, we translated the propaganda category variable to a binary: True / contains propaganda or False / does not contain propaganda. We again used spacy to split the text into sentences, and annotated each sentence with our logistic regression classification model. Then, if any sentences in the tweet contained propaganda, we labeled the tweet as predicted propaganda.

When comparing our predicted labeling to the original annotated tweet labeling, we found an f1-score of 0.83 for articles with no propaganda, an f1-score of 0.22 for articles with propaganda, and an overall accuracy of 0.72. Once again, but to a larger extreme than on the original testing data, the f1-score of articles with propaganda is much lower than the f1-score of those with no propaganda. It would be interesting to explore further why this occurs.