spotBias - Identifying Political Bias in News Articles

Ayush Kumar
MT21020
ayush21020@iiitd.ac.in

Ayush Singh Chilwal
MT21021
ayush21021@iiitd.ac.in

Manas N Kulkarni
MT21048
manas21048@iiitd.ac.in

Padmaa Jaulkar
MT21061
padmaa21061@iiitd.ac.in

Sakshi Kumari
MT21141
sakshi21141@iiitd.ac.in

ABSTRACT

Whenever a political incident is reported, written or published, it tends to reflect an underlying political bias of the reporter, narrator or media house. Our web based platform spotBias uses without feature selection method and applied Random Forest Classifier OneVsAll model to identify if there exists any bias towards any political party- BJP, Congress, AAP, None (NOT Biased).

1 INTRODUCTION

News and Media play an important role in forming public opinion. While reporting various incidents and happenings from around the world, these news outlets, bloggers, social media journalists, tend to spread their own underlying bias. There are various types of biases that compel them in deciding whether to report a certain news item or not, and/or how much time to be allotted for a certain news item which generally happens in case of TV news. The different types of media slant include Gate-keeping bias, which tells about the way an outlet selects a particular news article to publish. Then there is Coverage bias where the outlet decides how long or how much time it is going to spend on covering a particular news or event. Then we have framing bias, this can be defined as the way an author or journalist frames or presents a topic in his article. Which generally includes his/her underlying opinion on the subject. There is another bias similar to framing bias, which is ideological bias.

Our main motivation for selecting this topic includes identifying the political bias present in a news article. Certain media outlets tend to publish articles which strengthen a political party's view point or propaganda. Thus, they are approaching news items with a biased agenda. Since the general public most often then not, are not aware of the preferences of a Media Outlet, tend to believe them and the ideas/incidents reported in their articles. We believe it's highly important for the general public to know the truth (before forming an opinion), without falling prey to bias of the Media outlets or the bloggers' published articles.

2 PROBLEM STATEMENT

Developing an NLP/ML based web application that in real time reports whether the reporting done in news articles is politically biased or not.

3 DATASET

The dataset was built using BeautifulSoup on News Catcher API and Google News. It currently consists of roughly 900 articles and respective annotations. The annotations represent the bias in the news articles.The annotations are done by our team only. Two of the teams members have annotated the articles and rest three members have verified the annotations. Link to the dataset: Dataset.

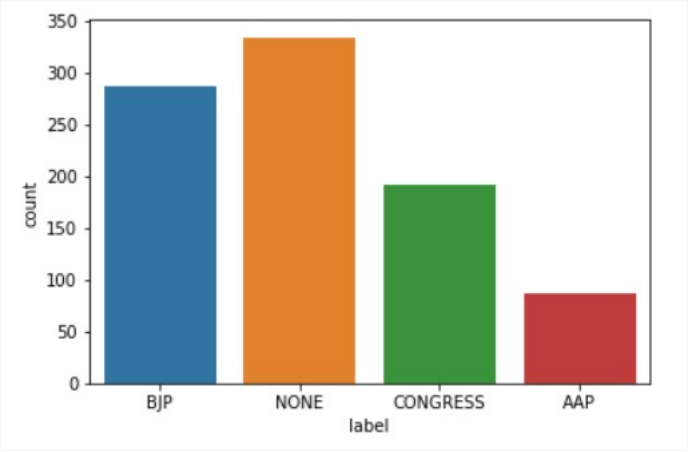

Figure 1: Dataset Distribution

4 METHODOLOGY

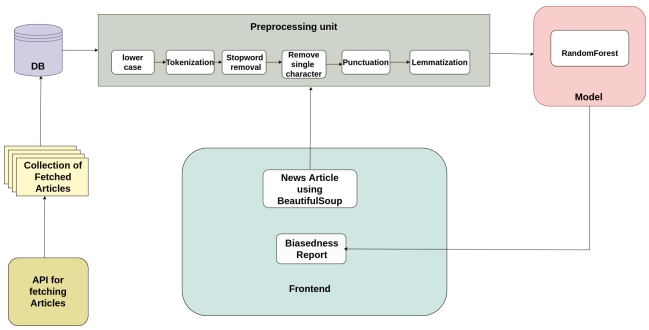

Figure 2: Flowchart

A new news article is fetched using Beautiful Soup on Google News. We preprocess it by changing into lowercase, tokenization, removing Stopwords, removing single characters, punctuation, and Lemmatization. The preprocessed article is then fed into model which consists of Random forest classifier. This outputs "None"

1.

in case the article is not biased towards any political party, else it outputs name of one of the political party BJP, Congress, AAP.

5 LITERATURE REVIEW

We have studies following research papers -

- **[1]Predicting Factuality of Reporting and Bias of News Media Sources** In this paper, they predict the factuality of reporting and biases of the news media. Then they have collected information from various sources such as Wikipedia page, Twitter account, the sample from the target website, and URL. Then they have rated the model of factuality on a scale of 3 points and bias on a scale of 7 points. After preparing dataset from different sources, they applied an SVM classifier and did hyper-parameter tuning to find the best parameters for their model separately for both bias and factuality. Then they concluded that articles from the target website, Wikipedia page, and Twitter are essential.
- **[2]Detecting Media Bias in News Articles using Gaussian Bias Distributions** They have pointed out that recent methods of bias detection such as neural text classification methods and feature based methods which are primarily dependent upon low level lexical analysis (neural nets, bagging of words). Hence these methods fail to show bias predictiveness when the words appear in new contexts. They have effectively used second order information to improve the detection accuracy. The second order information includes the probability distributions on positions, frequency and order of tokens and informational sentence-level bias in a Gaussian Mixture Model. They have compared their results with the standard model of sentence level bias detection and the comparison reveals that the models which use the second order information show better results.
- **[3]Detecting Political Bias in News Articles Using Headline Attention** This paper proposed a headline attention network. When this network is applied to the articles based on its headline, enables it to attend to more critical content to predict the bias on a dataset of 1329 news articles collected from various Telugu newspapers. This neural network-based method performed much better as compared to traditional methods.
- **[4]Identifying Political Bias in News Articles** Authors in this paper talked about different types of biases inherent in the News/Media industry. Their main goal includes to detect the newspaper's choices that reveal its underlying political inclination. They used 3 datasets namely Guardian, Telegraph's online available political articles from 1996 to 2015, 2000 to 2015 respectively. Third is UK Parliament speeches from 1934 until 2015. They used quotation extraction from news articles, the text search by Elastic search, and the named entity recognition by IBM Alchemy AP. They performed phrase queries with the politicians' first and last names and searched in the title, subtitle, body and image captions of the articles. Then plotted the graphs per article which shows per political party affiliations for them.
- **[5]Detection of hyperpartisan news articles using natural language processing Technique**This paper ripped through the issue of partisan selectivity in the era of prevalent false, misleading, and biased information. Hyperpartisan news is an extremely biased version of particular political news in terms of political misinformation. In order to mitigate hyper partisan news for the users there is a need for an automated model for its detection; hence different machine learning models are used in this research work to do so. This research makes use of the by-article dataset published at SEMEVAL-2019, consisting of 1273 news articles. The three models BERT, ELMo, and Word2vec work well in the identification of hyperpartisan news. This research to detect hyperpartisan news articles has surpassed all the machine learning and neural networks that are used to detect hyperpartisan news in SEMEVAL-2019. The accuracy of the hyperpartisan article detection model developed by Huang and Lee (2019) using BERT was 0.68, whereas the accuracy obtained in this research by using BERT is 0.83 with the same dataset, which exceeds by 20 percent. This research, evidenced from the described machine learning models, would integrate governments, news' readers, and other political stakeholders to detect any hyperpartisan news, and also help to regulate misinformation about the political parties and their leaders.
- **[6]Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking**In this paper they have done analysis of news media language in the context of political fact-checking and the detection of false news.To categories news articles they tokenize the text using NLTK library and each lexicon's per-document counted, as well as report averages per article of each kind.We investigate the viability of categorizing news articles into four groups: trustworthy, satire, hoax, or propaganda.Then they check the truthfulness of news article. Then they train the LSTM model over it and predict the Politifact rating.Then they look at honesty and the language features that contribute to it in a variety of contexts, such as online news sources and public utterances.
- **[7]Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias student name** This paper has emphasized how biases on the news article can be best visualized and communicated to the audience. For that, they created three manually annotated datasets. They presented the results of a prototypical user study in which they tested the effectiveness of communicating bias-related news using different visualization types and components. They include the factors of how users perceive media bias in the information. They used various articles: Plain Visually highlighted phrases representing the facts Visually highlighted framing effects Visually highlighted annotation of biased or unbalanced language They concluded that perceived journalized bias was directly and significantly related to the political extremeness.
- **[8]Automated identification of bias inducing words in news articles using linguistic and context-oriented features**As news consumers are bombarded with a constant stream of fake news, propaganda, hoaxes, rumors, satire, and advertising — that often masquerade as credible journalism

— it is becoming more and more difficult to distinguish fact from fiction. This work includes the mold system for detection of bias induced words in news articles. The research focused on text content creators showcase concepts differently by word choice. As of now no such tool or technology is developed for the automated identification of media bias, due to lack of annotated data sets and highly content dependency. The workflow includes collecting sentences via crowdsourcing process backed up by biased words lexicon, explicitly for this domain. By defining bias inducing words as binary classifiers, the combination of different linguistic features is possible. Overall, we believe that the feature based approach is especially valuable, relating bias to specific features, which is difficult with automated feature extraction. The prototypical system achieves appreciable results as F1-score of 0.43, precision of 0.29, recall of 0.77 and ROC AUC of 0.79 as compared to other researchers.

- **[9]A New Benchmark Dataset for Fake News Detection**Detection of fake news is difficult.In this paper, they are using the LIAR dataset, which is publicly available. Then they applied a novel hybrid convolutional neural network to map metadata with text.This method improves a text-only deep learning model. Then, using the extracted feature, they compared various models applied over the dataset and found CNN was giving the best result on the test set to verify the result of CNN via a two-tailed paired t-test.

6 EVALUATION

We have split our dataset into a ratio of 75:25. Training dataset includes 674 articles and testing dataset includes 225 articles. We have trained various classification models and following are the baseline results:

Model	F1 Score	Accuracy
Naive Bayes without feature selection	0.586	64.44
Naive Bayes with feature selection	0.790	78.66
SVM without feature selection: For Unigram	0.758	76.44
SVM without feature selection: For Bigram	0.408	49.333
SVM with feature selection	0.862	86.22
Random Forest without feature selection:For Unigram	87.02	87.11
Random Forest with feature selection	90.29	90.22

Figure 3: Results

We tested two other models. From that the first one is Random Forest without feature selection whose accuracy is 87.11%. Its precision and recall scores are 87.16% and 87.11%. The other model consists of Random Forest with feature selection. It is 90.22% accurate. Its precision and recall scores are 90.52% and 90.22% respectively.

7 CONCLUSION

The highest accuracy obtained from Baseline model was 86.22%. Our current model is Random Forest without feature selection is 87.11% accurate and Random Forest with feature selection is 90.22%.

REFERENCES

[1] R. Baly, Georgi Karadzhov, Dimitar Alexandrov, James R. Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *EMNLP*.

[2] Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting Media Bias in News Articles using Gaussian Bias Distributions. *ArXiv* abs/2010.10649 (2020).

[3] Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting Political Bias in News Articles Using Headline Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy, 77–84. https://doi.org/10.18653/v1/W19-4809

[4] Konstantina Lazaridou and Ralf Krestel. 2016. Identifying Political Bias in News Articles. *Bull. IEEE Tech. Comm. Digit. Libr.* 12 (2016).

[5] Navakanth Reddy Naredla and Festus Fatai Adedoyin. 2022. Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights* 2, 1 (2022), 100064. https://doi.org/10.1016/j.jjimei.2022.100064

[6] Hannah Rashkin, Eunsol Choi, Jin Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. 2931–2937. https://doi.org/10.18653/v1/D17-1317

[7] Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. 2020. *Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias.* Association for Computing Machinery, New York, NY, USA, 389–392. https://doi.org/10.1145/3383583.3398619

[8] 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 Management* 58 (05 2021), 102505. https://doi.org/10.1016/j.ipm.2021.102505

[9] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *ACL*.