Ideology Detection of Personalized Political News Coverage: A New Dataset

Khudran Alzhrani
Department of Information Systems
Al-Qunfudhah Computing College, Umm Al-Qura University
Al-Qunfudhah, Mecca, Saudi Arabia
Email:kmzhrani@uqu.edu.sa

ABSTRACT

Words selection, writing style, stories cherry-picking, and many other factors play a role in framing news articles to fit the targeted audience or to align with the authors' beliefs. Hence, reporting facts alone is not evidence of bias-free journalism. Since the 2016 United States presidential elections, researchers focused on the media influence on the results of the elections. The news media attention has deviated from political parties to candidates. The news media shapes public perception of political candidates through news personalization. Despite its criticality, we are not aware of any studies which have examined news personalization from the machine learning or deep neural network perspective. In addition, some candidates accuse the media of favoritism which jeopardizes their chances of winning elections. Multiple methods were introduced to place news sources on one side of the political spectrum or the other, yet the mainstream media claims to be unbiased. Therefore, to avoid inaccurate assumptions, only news sources that have stated clearly their political affiliation are included in this research.

In this paper, we constructed two datasets out of news articles written about the last two U.S. presidents with respect to news websites' political affiliation. Multiple intelligent models were developed to automatically predict the political affiliation of the personalized unseen article. The main objective of these models is to detect the political ideology of personalized news articles. Although the newly constructed datasets are highly imbalanced, the performance of the intelligent models is reasonably good. The results of the intelligent models are reported with a comparative analysis.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Supervised learning by classification**.

KEYWORDS

Machine Learning, Text Classification, Neural Language Processing, News Bias, Political Science, News Personalization

ICCDA 2020, March 9–12, 2020, Silicon Valley, CA, USA

ACM ISBN 978-1-4503-7644-0/20/03…$15.00
https://doi.org/10.1145/3388142.3388149

1 INTRODUCTION AND RESEARCH APPLICATIONS

Each political party is founded on a set of shared beliefs about how society should be organized and governed [9]. Hence, the party members take political stances accordingly to appeal to the strong partisan voters. Politicians join parties to receive recognition and funding for campaigns and advertisements. However, news media coverage has a high impact on the voters' turnout and choices[12].

News websites and blogs are maintaining their role as an influential medium in society through the use of social media to reach a large audience. News media credibility is usually associated with factual reporting. Fact-checking is a convenient tool to hold news websites accountable, yet they are also subject to news personalization. Political scientist have not agreed on a clear definition for news personalization and privatization [14] [4] [2]. In this paper, we don't distinguish between all the different terminologies and assume that news personalization includes any news coverage about a particular political individual. Nevertheless, we expect that political news media are affected by their political leanings. For instance, the news media tends to present the opposition party candidates or elected politicians in a skeptic, and cynical way [7]. On the other hand, the same news media will focus on the positive political and non-political traits of other preferable politicians. Therefore, it's possible to influence inattentive readers and direct their thoughts toward a particular ideology belief. Based on these assumptions, we argue that it is possible to automatically detect the political ideology of the personalized news with no human intervention. In other words, an intelligent model trained on labeled news articles will be able to capture the hidden attributes of news personalization and predict news articles' political affiliation that lines up with the composed message.

The are multiple applications for political affiliation detection of personalized news articles. Last year, YouTube labeled videos funded by the state or public sector [1]. One way to reason with actions taken by YouTube lately is to provide awareness to its viewers of any political or non-political messages delivered in these videos that might profit the funders. Nonetheless, this approach is indirect and extending the scope to include ideology labels based on the contents could improve viewers' awareness and judgments on controversial issues. The labeling process requires prior knowledge and it is time-consuming, yet it's achievable with automated solutions. Another clear advantage of using machine learning to detect the political ideology of news articles, unlike inattentive readers who

make decisions based on most accessible information in their memory, the intelligent model will be able to identify characteristics of the framed messages. In addition, editors can use intelligent models to examine the unintentional framing of politicians and ensure that articles don't exhibit any kind of bias. The contributions of this research paper are summarized as follows:

- The construction of large experimental datasets with more than 178,000 articles that address the problem of personalized news automated ideology detection.
- The implementation of multiple intelligent models to detect the political affiliation of articles based on its content.

The reminder of the paper is organized as follows: related work is presented in Sec. 2. Sec. 3 details our research methodology. Sec. 4 describes the news personalization and political affiliation dataset. The results of our experiments are presented in Sec. 5. Finally, the paper is concluded in Sec. 6.

2 RELATED WORK

The recent advancements in the field of machine learning and deep neural networks provided researchers with the opportunity to address unconventional problems. In this section, we list and briefly describe some of the research papers examined the problem of automated ideology detection. Similar to advertising, the electronic news-press delivers news in video, image, audio, text or a combination of these. The prominent news outlets reach out to the younger generation by establishing their presence on social networks, such as Snap, Instagram, and Twitter. Compared to other data types, texts are easier and faster to generate. Therefore, digital textual remains one of the most common medium for news media. Clickable news' titles attract the audience to read the article and usually give hints on the news' attitude towards the reported politician or issue. However, headlines are short by nature and don't provide sufficient information to train on. Commonly, research papers analyze articles' content alone or with their titles.

In the literature, the emphasis is on topical news articles' ideology detection. Meaning, the corpus consists of various named entities, such as locations, organizations, and persons. Thus, it is easier to distinguish between the political ideologies of topics or issues. We argue that linking political news topics to specific politicians provides a clear advantage over other approaches listed below.

Identifying political affiliation has been applied to a variety of textual information, user-generated content such as Twitter, news articles, and political speeches. Iyyer1 et al. [6], detected the ideology of US Congressional floor debate transcripts and ideology books using Recursive Neural Networks(RNN). Besides RNN, the authors experimented with multiple extracted features as an input to a Logistic Regression. The results on both datasets for Sentence-level bias detection accuracy ranged from 62.1% to 70%. Kulkarni et al. [11], trained a Deep Neural Network Model on the US News articles' title, content, and hyperlinks found in the article body to detect its political ideology. The annotation of the dataset used in this paper was done by readers blindly without knowledge of their source. We believe this approach is limited since it relies on human intervention that may lack the required knowledge or interest. Also, only one of the baseline classifiers was trained on the content of the articles. The f1-measure reported in this paper ranged from 59.12 to 79.67. Ideology detection is used on Twitter as well to profile users on Twitter [5], [13] or predict tweets ideology [10].

3 RESEARCH METHODOLOGY

To examine the applicability of personalized news ideology detection through intelligent systems such as machine learning. We followed the most known practices for extracting and processing texts. Also, we implemented several classifiers to study their performance on the newly constructed datasets.

3.1 Data Pre-processing

Text classification as a model consists of several stages see Fig 3. The following is a brief description of each one of them. The datasets are stored in a flat text file in raw format. Classifiers are unable to processes unstructured texts; therefore, the collection of training texts articles are converted to a matrix of count tokens. However, only informative tokens that could boost the classifier performance are desired. Hence, there are simple yet critical procedures that have to take place before producing the numerical tokens vector.

The tokenization process involves breaking down texts into unique letters separated by white space. In the Tokenization step, all the one and two- grams are extracted from the raw datasets as unique tokens. The definition of a gram in this experiment is any word with two or more letters. Accents, punctuation, symbols are cleared from the texts. Then Tokens characters are turned into lower cases to avoid feature duplication in later stages. A set of predefined English stop-words acquired from the Natural Language Toolkit (NLTK) were also removed token lists. The stop-words don't provide a meaningful context such as "the", "you", and "did". The remaining tokens are counted in each document to output a matrix of tokens counts. any token that appeared in all the articles or just one article were also removed. Each column in the tokens count matrix is unique -features- and the corresponding rows are documents. This will often result in a sparse matrix, where each document will consist of a small set of extracted features.

In the second stage, the document to features matrix of counts is transformed into a term frequency-inverse document frequency (TFIDF) with L2 normalized representation. The TFIDF was originally developed for information retrieval, but it's also effective for text classification. The advantage of TFIDF over document frequency is better features' weights and lessen the effect of high or low features occurrence in the documents.

Finally, the list of labels that correspond to the documents is encoded to numerical values to prepare it for classifiers learning and testing. Similar to the procedures that we followed in the prepossessing stages for the training set, the testing set is tokenized; yet all the tokens are removed except for the ones found in the training set. Then the counts and TFIDF of the same features are calculated in the testing set. Testing set labels are also encoded to match the numerical values in the encoded training labels.

3.2 Building the Models

After processing and preparing the data, the choice of the classifier is essential for the model performance. In this research, several text classifiers were developed to experiment on our corpus. As illustrated in Fig 3, first the classifiers are trained on the training

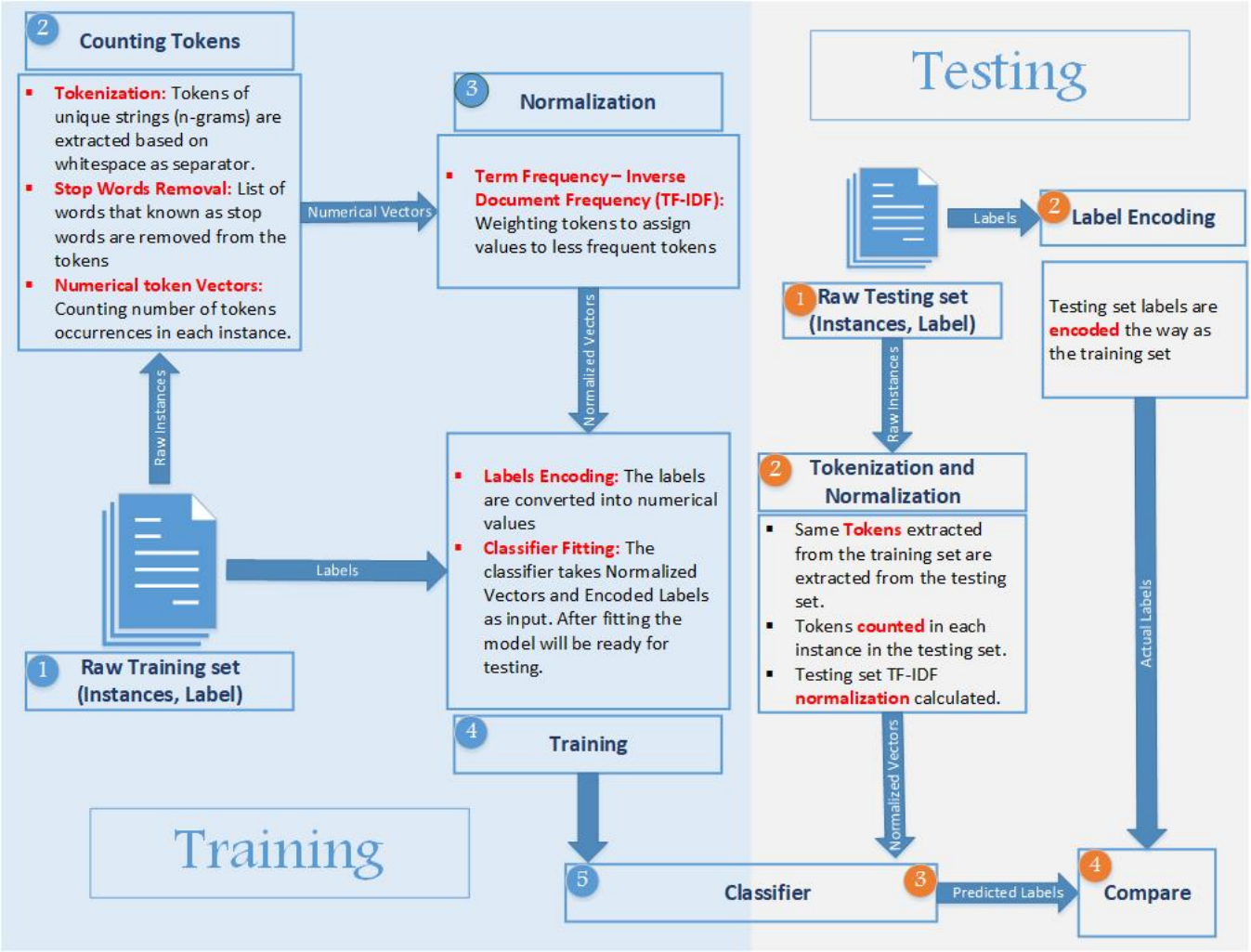


Figure 1: This figure illustrates the procedures followed in our methodology from extracting features extraction to training and testing. To replicate a real-world scenario, the training and testing sets are separate which means the new features in the testing set are not used.

set and then predict the class of the new unseen documents which are derived from the testing set. The classifiers used in this paper are briefly explained below.

3.2.1 Ridge Classifier . Ridge Classifier is based on a ridge regression linear model that penalizes the coefficients' size to minimizes the residual sum of squares between the observed variables in the dataset and the independent variables predicted by the linear approximation. In our case, the independent variables are qualitative predictors. The mathematical formulation of the Ridge Regression model is as follows :

$$J(w) = \lambda \|w\|^2 + \sum_i (w^T d_i - y_i)^2. \quad (1)$$

Where $w = (w_1, ..., w_n)$ are the coefficients. Ridge regression controls the shrinkage degree to get approximately close to the best possible population parameters by the λ in the left part of the equation. While the right part of this equation works the same way as in the Ordinary Least Squares algorithm.

3.2.2 Support Vector Machine with SGD. Support Vector Machine(SVM) classifiers have proven to be effective for a variety of text classification problems. Since texts have a high number of feature dimensions, SVM as a linear classifier is well suited for relatively small textual datasets [8]. In theory, the performance of SVM is greatly degraded when applied to imbalanced datasets [3]. We use SVM with Stochastic Gradient Descent (SGD) learning, which is considered preferable when the dataset is large.

3.2.3 Nearest Centroid. The Nearest Centroid also is known as the Rocchio classifier computes the centroid of each class instance. The Euclidean distance algorithm calculates the distances between class's instances vectors and other points to determine classes' boundaries. The boundaries between different classes are set by

finding points that have equal distances to all classes' centroids. This will create a region for each class, the new unseen documents will be classified based on their distance from each class's centroid. Assuming the training documents' vectors represented in the dataset as $\vec{d}$ and the set of labels $y_i \in \mathbf{Y}$, then the set of documents vectors and their labels are $\{(\vec{d}_1, y_1), \ldots, (\vec{d}_n, y_n)\}$. In the training phase, the nearest centriod for each class is given by

$$\vec{\mu_a} = \frac{1}{|J_a|} \sum_{i \in J_a} \vec{d}_i \tag{2}$$

Where J_a is the set of indices that points to the class a documents. In the testing phase, the distance between the new document's vector $\vec{d}$ and each class' centroid is computed, then the document will be assigned to the class with the nearest distance by

$$\hat{y} = \arg\min_{l \in \mathbf{Y}} \|\vec{\mu}_a - \vec{d}\|. \tag{3}$$

3.2.4 Naive Bayes. In addition to the two linear classifiers and nearest centroid, we are experimenting with two variations of Naive Bayes namely Bernoulli and Multinomial. Naive Bayes classifiers are based on Bayes' theorem, hence the name. All Naive Bayes classifiers' variations are probabilistic and assume features independency.

Bernoulli Naive Bayes is a simple classifier that predicts the document's class C by computing the probability p of a term t occurrence or absence in a document d.

$$p(\mathbf{d} \mid C_a) = \prod_{t=1}^{n} p_{at}^{d_t} (1 - p_{at})^{(1-d_t)} \tag{4}$$

Unlike Bernoulli, Multinomial Naive Bayes takes into consideration the frequency of a term in documents to determine the probability p of a term t occurrence in a document that belongs to the observed class C.

$$p(\mathbf{d} \mid C_a) = \frac{(\sum_t d_t)!}{\prod_t d_t!} \prod_t {p_{at}}^{d_t} \tag{5}$$

4 DATASET

In this section, we will go over the datasets' details and their construction process.

4.1 News Sources

As we have stated repeatedly in this paper, the literature lacks the data needed to study the problem of personalized political news articles. Therefore, we collected and reconstructed two datasets that should be suitable for our research experiments. While most news sources claim to be bias-free, there are ways to categorize them based on their political ideology. For instance, the audience could be polled with several questions to determine their political ideology and favorite news source. Our dataset was built out well-known news websites, but we did not include any content from the mainstream news media. However, most of the mainstream news media state that they are all about professional journalism and are not biased to any particular political side. Therefore, all the news websites that were included in this research are self-identified as either to the far left or right. The news sources are listed in Table 1.

Table 1 shows that we collected news regarding two political leaders, President Trump, which is considered a conservative president, and his predecessor President Obama, liberal. The same table shows we collected articles from seven news websites, five of them are conservative, and the other two are liberal. The conservative news websites are DailyWire, I Love My Freedom, National Review, TheBlaze, and NewsBusters. On the other hand, liberal news websites are DailyKos and World Socialist. We found that liberal news websites produce more content than conservative ones. The number of collected articles from the DailyKos is far more than all the other conservative websites combined. Since President Obama has been in the spotlight longer than President Trump, it's understandable that he received more coverage.

Table 1: This table display news sources, ideology alignments, number of articles for Trump and Obama datasets. Although we collected articles from five conservative news sources and just two liberals, the news sources aligned with liberal ideology produces more articles than their conservative counterparts. The number of articles published by the DailyKos for Obama dataset is greater than all other news sources combined.

News Website	Ideology Alignment	Trump Articles	Obama Articles
DailyWire	Conservative	11031	2489
I Love My Freedom	Conservative	6351	1180
DailyKOS	Liberal	39868	81737
National Review	Conservative	9982	9945
TheBlaze	Conservative	614	485
World Socialist	Liberal	2954	4054
NewsBusters	Conservative	4669	3300
		75497	103190

4.2 Datasets Construction

News articles covering two political leaders, Obama and Trump, are collected from the seven news sources. Each article is labeled with either Liberal or Conservative based on the news source affiliation. Although both of the datasets are collected from the same news sources, they are separate from one another. The datasets' construction process goes as follows. First, we crawled news websites and extracted any contents that tagged one of the political leaders or appeared on the websites' search engine. Normally, due to the differences in the structuration of news website sources, multiple crawls programs are needed to extract the required information. In the extraction process, we made sure to retain articles publishing date and time, title, article body, and author. An ID and labels are assigned for each article at the extraction time.

Once the crawling data was done, we have removed all the articles with fewer than 10 words. Then the datasets are shuffled and split into two parts training and testing. The statistics of the two datasets are in Table 2. As shown in the table, news articles from the news sources are split between the training and testing datasets. However, in the training procedure, the source of the news article, time and data, titles are disregarded and only articles contents and its labels are used.

Two classes are presented in our datasets, they vary in size see Table 3. Both datasets are skewed towards the liberal class, but the

Table 2: News Sources to the number of training and testing articles for each dataset. Although the datasets were shuffled before splitting it into training and testing set, percentage-wise, the number of articles from each source are proportionally similar.

	Trump Dataset		Obama Dataset	
News Source	Train	Test	Train	Test
DailyWire	7797	3234	1758	731
I Love My Freedom	4489	1862	826	354
DailyKOS	27853	12015	57341	24396
National Review	6929	3053	6880	3065
TheBlaze	432	182	351	134
World Socialist	2055	899	2827	1227
NewsBusters	3287	1382	2288	1012
Total	52842	22627	72271	30919

Obama dataset is highly imbalanced with over 83% of the dataset belongs to the liberal class.

Table 3: This table shows the total number of training and testing articles for each class in Trump and Obama datasets. It's quite clear that the number of liberal articles outnumbers the conservative ones in both datasets. Moreover, the Obama dataset suffers more from dataset imbalance, the liberal articles populate around 83% of Obama dataset, training and testing set.

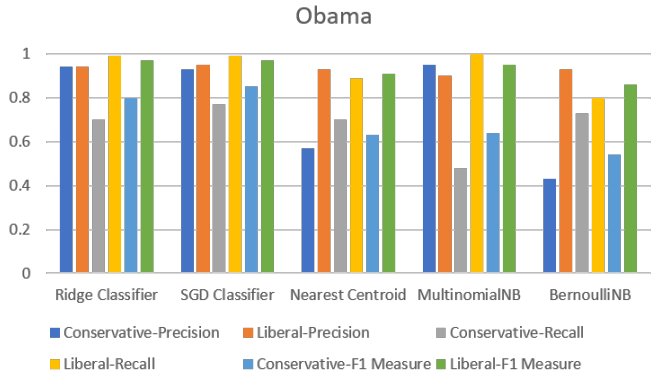
	Trump Dataset		Obama Dataset	
Class	Train	Test	Train	Test
Conservative	22934	9713	12103	5296
Liberal	29908	12914	60168	25623

5 EXPERIMENTAL EVALUATION

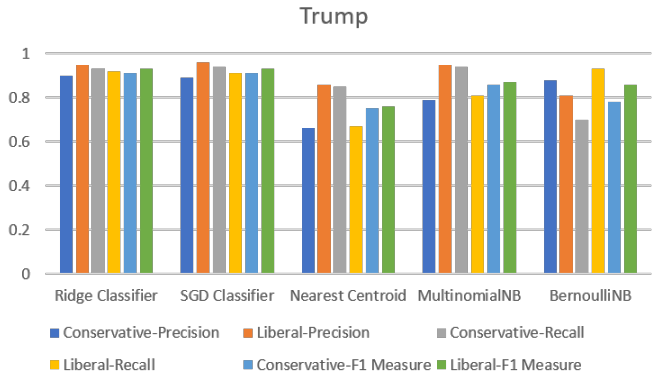
5.1 Setup

The operating system and software platforms used to conduct the experiments in this paper are Microsoft Windows 10 Home, Python 3.5.2 for reading and processing data and we took advantage of exiting tools in the well-known machine-learning library namely Scikit-learn to pre-process texts and built models. The experiments were performed on x64-based processor Intel(R) Core(TM) i7-4720HQ CPU @ 2.6GHz with 16GB RAM.

One set of parameters was used for all the classifiers for both datasets. As illustrated in 2, we experimented with 5 different classifiers Ridge Classifier, SVM-SGD, Multinomial and Bernoulli Naive Bayes and Nearest Centroid. Rather than Accuracy which is a common metric for evaluating text classifiers, Recall, Precision and F1-Measure are used. These measurements can capture models' performance on imbalance datasets and identify weaknesses in each model.



(a) This figure illustrates the performance of the 5 models on the Obama dataset. Based on the Recall, Precision, and F1-score of results, the SGD classifier reported better results compared to the other 4 classifiers. Around 83% of the Obama training and testing set consist of liberal articles, yet the SGD classifier was able to successfully recall 0.77 and 0.99 of the conservative, and liberal instances respectively. The Multinomial Naive Bayes suffered the most from the problem of the imbalanced dataset and only was able to achieve 0.48 on the recall metric.



(b) This figure illustrates the performance of the 5 models on Trump Dataset. Both Ridge and SGD classifiers are linear and achieved better results than the other three models. The Trump dataset is an imbalanced dataset, however; The linear classifiers were able to get over 0.90 on all metrics for both conservative and liberal class. Multinomial Naive Bayes is second to the linear classifiers, though the same classifier is ranked last on the Obama dataset. This could be attributed to the fact that the Obama dataset is highly skewed toward the liberal class.

Figure 2: Two figures that illustrate the results of 5 classifiers namely, Ridge, SGD, Nearest Centroid, Multinomial Naive Bayes, and Bernoulli Naive Bayes on Obama and Trump datasets. The metrics are on a scale from 0 to 1 for Recall, Precision, F1-Measure on two classes conservative and liberal.

5.2 Experimental Results

The linear classifiers are known well suited for sparse datasets, therefore on both experiments, linear classifiers achieved better results see Fig 2. However, relatively to the models' performance on

the Trump dataset, the models were not able to repeat the same feat on the conservative class of the Obama dataset. Since the Obama dataset is more skewed towards the liberal class than Trump Dataset, the influence of the dataset imbalance on the models' performance is clearer. All models achieved between 0.8-0.99 recall, 0.9-0.95 precision, and 0.86-0.97 f1-measure for the Obama dataset's liberal class. On the other hand, the same models scored between 0.48-0.77 recall, 0.43-0.95 precision, and 0.54-0.85 on f1-measure for the Obama dataset's conservative class. Although Multinomial Naive Bayes classifier reported the best precision on the Obama dataset's conservative class, its recall was the worst among the other classifiers. The results detailed earlier indicate that all models performed quite well in liberal classes. However, because only a small portion of the Obama dataset consists of conservative class, the precision of the liberal class didn't capture the magnitude of misclassified conservative articles. With 0.85 f1-measure on the Obama dataset's conservative class, the SGD classifier outperformed all other models. Similarly, the SGD classifier ranked first along with Ridge classifier on all three metrics for Obama's liberal class. The results prove that the SGD classifiers are highly suitable for imbalanced spares datasets.

Trump dataset is smaller in size with a lower imbalance margin compared to the Obama dataset. This lessened the impact of skewed class on the models' performance, hence the better results. The models' results on scale 0 to 1 are 0.67-0.93 recall, 0.81-0.96 precision, and 0.76-0.93 f1-measure for Trump datasets' liberal class. As for the Trump datasets' conservative class, the reported results from the same models are 0.7-0.94 recall, 0.66-0.9 precision, and 0.75-0.91 f1-measure. Compared to the performance of the classifiers on the Obama dataset, the results on Trump dataset show a lower consistency on all three metrics for the liberal class. Meaning that the difference between the lowest recall, precision and f-measure and the highest is larger on Trump dataset. However, the discrepancy between the performances of the classifiers on the two classes on Trump dataset is lower in comparison to the Obama dataset, which indicates a better trad-off. One interesting point is that all models except for the BernoulliNB classifier got higher recalls for the Trump dataset. This could be contributed to the articles' length, the number of features, topics overlapping, and more importantly personalization. Nonetheless, the same classifiers obtained lower precision for Trump dataset's conservative class. Similar to the Obama dataset, the linear classifiers outperformed all other models with 0.91 and 0.93 f1-measure for conservative and liberal classes respectively. Unlike the results obtained from the Obama dataset experiment, the worst performing model is the nearest centroid classifier with 0.75 and 0.76 f1-measure for the conservative and liberal classes respectively.

6 CONCLUSION

As far as we are concerned, this is the first paper that studies the problem of automated ideology detection of personalized news articles. Even though the datasets are heavily imbalanced, the models performed well. The Conservative class in the Trump dataset achieved better f1-measure despite having a fewer number of articles compared to the Liberal class. This might indicate that the Trump personalization in the Conservative news coverage is stronger than its liberal counterpart. There is much future work to be done to improve upon the current results. Utilizing text augmentation methods might be able to combat the data imbalance problem, hence, the models' performance would improve. Deep Neural Networks have proven effective on text classification problems, but it requires a large number of data. Another approach is to incorporate more data in the training process such as time, news source or even authors' name. In conclusion, this paper proved that the ideology of personalized news articles is possible.

REFERENCES

[1] 2018. YouTube to label government and public-funded clips. https://www.bbc.com/news/technology-46139189

[2] Silke Adam and Michaela Maier. 2010. Personalization of politics–Towards a future research agenda. *A critical review of the empirical and normative state of the art. En Charles T. Salmon (Ed.), Communication Yearbook* 34 (2010).

[3] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. 2004. Applying support vector machines to imbalanced datasets. In *European conference on machine learning*. Springer, 39–50.

[4] W Lance Bennett. 2016. *News: The politics of illusion*. University of Chicago Press.

[5] Yupeng Gu, Ting Chen, Yizhou Sun, and Bingyu Wang. 2016. Ideology detection for twitter users with heterogeneous types of links. *arXiv preprint arXiv:1612.08207* (2016).

[6] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1113–1122.

[7] Nael Jebril, Erik Albaek, and Claes H De Vreese. 2013. Infotainment, cynicism and democracy: The effects of privatization vs personalization in the news. *European Journal of Communication* 28, 2 (2013), 105–121.

[8] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, 137–142.

[9] John T Jost, Christopher M Federico, and Jaime L Napier. 2009. Political ideology: Its structure, functions, and elective affinities. *Annual review of psychology* 60 (2009), 307–337.

[10] Sandeepa Kannangara. 2018. Mining twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 751–752.

[11] Vivek Kulkarni, Junting Ye, Steven Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. *arXiv preprint arXiv:1809.03485* (2018).

[12] Diana Owen. 2017. New media and political campaigns. In *The Oxford handbook of political communication*.

[13] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 729–740.

[14] Peter Van Aelst, Tamir Sheafer, and James Stanyer. 2012. The personalization of mediated political communication: A review of concepts, operationalizations and key findings. *Journalism* 13, 2 (2012), 203–220.