CMPUT 690 Project Report: Hyperpartisan News Detection (SemEval-2019 Task 4)

Maisha Binte Moin
bintemoi@ualberta.ca

Abstract

Yellow journalism has increased the spread of hyperpartisan news on the internet. Hyperpartisan news exhibit blind, prejudiced allegiance to one party, faction, cause, or person. In this paper, we describe our system for SemEval-2019 Task 4: "Hyperpartisan News Detection". We attempted this classification task by using traditional machine learning approaches. Our best accuracy is 53% and is obtained by using support vector machine. We have utilized pre-trained word embeddings and some additional explainable features to train the models. Finally, we provide some explanation for the results we achieved and propose improvements that can be made to our work in the future.

1 Introduction

Social media usage has undeniably become a part of our daily routine. This has caused a shift in our manner of news consumption as people are being more exposed to news found on social media platforms than that of traditional news outlets. However, social media can be an outlet for misinformation. A subtle form of this misinformation is that of hyperpartisan news (Ross et al., 2021), which exhibits extreme bias towards a single side, particularly in a political spectrum.

Study by Vosoughi et al., 2018, show that fake news can spread at an alarming speed and the spread of hyperpartisan news is the most viral. The speed of spread for hyperpartisan news re-exhibit a common inclination towards pre-existing bias in us humans. The speed and extent of the spread can be tied to political and financial incentives, both online and offline (Torabi Asr and Taboada, 2019). As such, an automated way of detecting such form of news is necessary in order to implement countermeasures and to stop this misinformation spread. SemEval-2019 Task 4, imposes the goal of 'Hyperpartisan News Detection' (Kiesel et al., 2019). In this project, we have employed traditional machine learning approaches for solving the proposed task, but we plan to cover deep learning-based approaches in future. In summary, the contribution of this paper are as follows:

1. Study and characterization of useful features for model explainability.

2. The usage and comparison of several machine learning algorithms for the task.

3. Proposal of possible ways of extending this work by using deep learning based approaches.

2 Related Work

Fake news detection has been widely studied (Mahir et al., 2019, Cardoso Durier da Silva et al., 2019, Zhou et al., 2019, Ross et al., 2021), in comparison to studies related to hyperpartisan news detection. While a previous study (Potthast et al., 2017) have analyzed the writing style of hyperpartisan news to distinguish them from mainstream news. The dataset made available through SemEval-2019 Task 4, made it possible to further investigate the detection of such biased news.

In the current literature, we see several different groups of input features, machine learning and deep learning models utilized to tackle this task. Machine learning models used for the task include support vector machines (SVM), logistic regression (LR), random forest (RF), XGBOOST, naïve bayes (NB). Saleh et al., 2019, trained a logistic regression model with features ranging from simple bag-of-words to text readability features. Alabdulkarim and Alhindi, 2019, trained SVM model that uses TF-IDF of tokens, Language Inquiry and Word Count (LIWC) features and structural features from the data.

In deep learning based approaches, we see that BERT (Drissi et al., 2019, Naredla and Adedoyin, 2022), BiLSTM (Zhang et al., 2019), CNN (Jiang et al., 2019) have been utilized for this task.

The most popular type of neural networks among the participants were convolutional ones (CNNs) (Kiesel et al., 2019). The winning team 'Team Bertha von Suttner' (82.2% accuracy) used a convolutional neural network (CNN) model, which used ELMo based sentence embeddings to represent the article (Jiang et al., 2019).

3 Methodology

As stated earlier, the primary goal of this project is to investigate machine learning approaches and useful features for hyperpartisan news detection. For this, we have used several machine learning models to approach the task, as machine learning models are widely used to provide solutions in both classification and regression tasks. High-level overview of the pipeline is data preprocessing, using pre-trained word embeddings for training, fitting the models, then finally comparing the model performances.

For data preprocessing, we used a standard process of stemming, lemmatization, and removing stop words. Next, we used this preprocessed data to create word2vec word-embeddings, where each word is converted into vectors and all the word vectors associated with an individual article were summed up. The use of pre-trained word embedding features make the models perform better and they are less likely to be affected by the change of topic. Afterwards, we explore some additional information the model may leverage to make decisions. These additional features are based on the work by Palić et al., 2019, which include: publication date, website referencing, sentiment analysis and trigger words. The intuition behind the use of publication date is that hyperpartisan news tend to appear more during election times. Website referencing is also important, it had also been utilized to create the larger 'by-publisher' dataset, as there are known websites that display political bias.

Finally, the models that we used were: SVM, Logistic Regression, Random Forest, K-Nearest Neighbors, Decision Tree for the classifiaction. Figure. 1 graphically show the accuracy values for the different algorithms. The best model from our results is SVM through GridSearch.

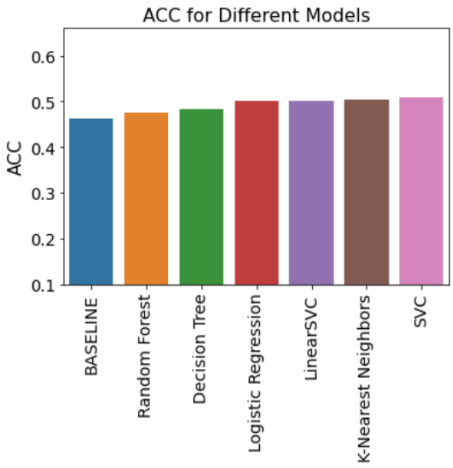

Figure 1: Accuracy comparison

4 Evaluation

This section describes the dataset used and the results obtained.

4.1 Dataset

Two types of dataset have been made available for the task. The 'by-article' dataset consists of 645 manually annotated articles and the 'by-publisher' dataset consists of 750,000 articles that are automatically annotated publisher-wise (Cruz et al., 2019), there is a half and half split in the 'by-publisher' dataset for news that is hyperpartisan and news that is not. Among the 645 manually annotated data 238 (37%) are hyperpartisan and 407 (63%) are not (Kiesel et al., 2019). The test set is balanced and contains 628 articles. The systems are evaluated based on their accuracy score (ACC) on the test set and the baseline accuracy is 46%, based on the competition statistics (Kiesel et al., 2019). In this project, we have used the 'by-article' dataset, as previous studies (Jiang et al., 2019, Palić et al., 2019), showed that the larger 'by-publisher' dataset contain a lot of noise and can be potentially harmful for the classifiers, for example, weather report is tagged as hyperpartisan based on the source of the news.

4.2 Results

We have compared the performance of several well known machine learning classifiers for this task. In Table 1, we see that there is no drastic difference in the accuracy values produced by different classifiers. SVC with GridSearch obtained the maximum accuracy on the dataset.

Table 1: Classifier Results

No.	Model	ACC
1	Decision Tree	0.476
2	Random Forest	0.479
3	Logistic Regression	0.500
4	Linear SVC	0.501
5	K-Nearest Neighbors	0.504
6	SVC	0.530

4.3 Discussion

In our experiments, the use of website referencing improved the performance of our SVM model by 2%. To exemplify how it was useful, in Figure 2, we show the correlation of hyperpartisan in the dataset based on publisher bias.

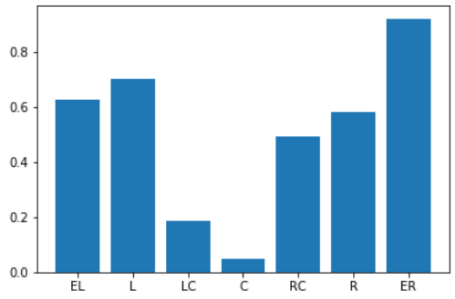

Figure 2: Percentage of hyperpartisan articles by publisher bias

Furthermore, while the use of pre-trained word2vec word embeddings makes models more robust, one limitation is that they generate the same embedding for the same word in different contexts. Because of this, the use of contextualized embeddings can further improve this work.

5 Conclusion

To conclude, this task was primarily challenging due to the complexity in labeling such articles, and differences in writing styles across domains, publishers and individuals (Alabdulkarim and Alhindi, 2019). We used several machine learning classifiers along with explainable features to classify hyperpartisan argumentation. This set of explainable features can be extended in the future. As especially in the current political climate, it is necessary that hyperpartisanship detectors not only reach a high accuracy, but also reveal their reasoning (Kiesel et al., 2019).

For future work, we shall investigate deep learning approaches as they have achieved state-of-the-art results in various other NLP tasks. While various deep learning based approaches have already been applied, we propose experimenting with more recent Longformer (Beltagy et al., 2020) and Big Bird (Zaheer et al., 2020) Transformer architectures, as they are able to handle longer sequence of text compared to BERT models and have the potential to achieve higher accuracy in this task considering the article lengths.

References

Amal Alabdulkarim and Tariq Alhindi. 2019. Spider-Jerusalem at SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 985–989.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Fernando Cardoso Durier da Silva, Rafael Vieira, and Ana Cristina Garcia. 2019. Can machines learn to detect fake news? A survey focused on social media. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

André Cruz, Gil Rocha, Rui Sousa Silva, and Henrique Lopes Cardoso. 2019. Team Fernando-Pessa at SemEval-2019 Task 4: Back to basics in hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Mehdi Drissi, Pedro Sandoval, Vivaswat Ojha, and Julie Medero. 2019. Harvey MUDD college at SemEval-2019 Task 4: The Clint Buchanan hyperpartisan news detector. *arXiv preprint arXiv:1905.01962*.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Team Bertha von Suttner at Semeval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq, et al. 2019. Detecting fake news using machine learning and deep learning algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE.

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):100064.

Niko Palić, Juraj Vladika, Dominik Čubelić, Ivan Lovrenčić, Maja Buljan, and Jan Šnajder. 2019. TakeLab at SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 995–998.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Robert M Ross, David G Rand, and Gordon Pennycook. 2021. Beyond" fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. *Judgment & Decision Making*, 16(2).

Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeno, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. 2019. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection. *arXiv preprint arXiv:1904.03513*.

Fatemeh Torabi Asr and Maite Taboada. 2019. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):2053951719843310.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Chiyu Zhang, Arun Rajendran, and Muhammad Abdul-Mageed. 2019. UBC-NLP at SemEval-2019 Task 4: Hyperpartisan News Detection With Attention-Based Bi-LSTMs. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1072–1077.

Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019. Fake news detection via NLP is vulnerable to adversarial attacks. *arXiv preprint arXiv:1901.09657*.