Harvey Mudd College at SemEval-2019 Task 4: The Carl Kolchak Hyperpartisan News Detector

Celena Chen cechen@hmc.edu
Celine Park cpark@hmc.edu
Jason Dwyer jdwyer@hmc.edu
Julie Medero jmedero@hmc.edu

Abstract

We use various natural processing and machine learning methods to perform the Hyperpartisan News Detection task. In particular, some of the features we look at are bag-of-words features , the title's length, number of capitalized words in the title, and the sentiment of the sentences and the title. By adding these features, we see improvements in our evaluation metrics compared to the baseline values. We find that sentiment analysis helps improve our evaluation metrics. We do not see a benefit from feature selection. Overall, our system achieves an accuracy of 0.739, finishing 18th out of 42 submissions to the task. From our work, it is evident that both title features and sentiment of articles are meaningful to the hyperpartisanship of news articles.

1 Introduction

In 2019, Task 4 of the SemEval Workshop asked participants to automatically identify hyperpartisan texts (Kiesel et al., 2019). Hyperpartisan news detection is the problem of building a classifier using natural language processing techniques in order to label news articles as either hyperpartisan or neutral in content bias. Hyperpartisan articles, in this case, can be defined as articles which are very polarized and extremely biased towards one political party. This task is quite relevant in today's political climate, with reports of "fake news" articles heavily influencing votes and people's support of some candidates running for government offices. The issue is especially egregious because these biased news articles are informing political opinions of people who believe them to be factual and impartial, with no easy way to remove or detect hyperpartisan news articles. This is also a non-trivial task, as hyperpartisan news is not extremely explicit in its bias, and even human readers do not always agree on which articles should be classified as hyperpartisan or what about those articles merits such a classification. There is no one unifying feature of hyperpartisan news, and even news publishers which do produce hyperpartisan news are not guaranteed to *only* publish hyperpartisan news. Therefore, each article must be evaluated for its degree of hyperpartisanship, along many different axes of measurements.

A functional and accurate hyperpartisan news detector would be useful for social media sites and other carriers of news to make sure that their news content is unbiased, and to be able to detect and perhaps remove or block sources of hyperpartisan news. Facebook, for example, has been under great public scrutiny due to the quantity and popularity of hyperpartisan news on its site. Hyperpartisan news detection could also be useful for researchers seeking to understand the scope and impact of hyperpartisan news on the 2016 presidential election and how it can continue to inform voters today and in future elections.

2 Previous Work

Hyperpartisan news detection has become a popular application of natural language processing due to its relevance in contemporary politics. Specifically, there has been research to hash out what features are prevalent in hyperpartisan articles. Buzzfeed conducted a manual analysis of nine different pages on Facebook: three that were mainstream news, three that were hyperpartisan left, and three hyperpartisan right (Silverman et al., 2018). They rated every post as mostly false, mixture of true and false, mostly true, or not factual, for posts like memes or jokes. They determined that hyperpartisan articles on both the left and the right side have more in common with each other than with articles in the mainstream, and detecting whether or not an article was hyperpartisan was easier than detecting the actual orientation of the bias (Potthast et al., 2017). Likewise, in our application, we build a hyperpartisan news detector which labels hyperpartisanship but not whether an

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 957–961
Minneapolis, Minnesota, USA, June 6–7, 2019.

article is left- or right-leaning.

Fake news, which hyperpartisan sites are more likely to produce, tends to have certain qualities about its titles that make them distinct (Horne and Adali, 2017). These qualities are longer titles, simpler, more readable vocabulary words, and multiple words in the title which are all capitals. We use these qualities to inform our feature extraction of the article title, extracting the length of titles, the average length of title words, and the number of words in all capitals and adding these features to our larger feature matrix.

Polarity indicates how positive or negative a text may be, or the direction of the bias, while subjectivity indicates how strongly the text represents an opinion versus an objective fact, or the magnitude of the bias (Liu, 2010). We hypothesize that hyperpartisan news articles will carry relatively strong observable opinionation in comparison to non-hyperpartisan articles, so we use the two metrics of subjectivity and polarity as an addition to the other features in our feature matrix.

In this vein, one past study used sentiment analysis on the comment sections of articles about the Trayvon Martin case. It determined that more well known commentators tended to have stronger sentiment in their comments (Ignatow et al., 2016), implying that sentiment is a useful metric for analyzing opinions on the World Wide Web. Another study that used sentiment analysis on social media data showed that sentiment analysis was a key technique for extracting features of an opinion, allowing the researchers to propose models for simulating and forecasting online opinions (Kaschesky et al., 2011). This research in particular was interesting, because it covers a similar area of interest as hyperpartisan news detection. That is, it examines the far-reaching effects of political opinions and their proliferation on the World Wide Web, and also uses similar natural language processing techniques to extract information about these opinions. This tells us that sentiment analysis is an important tool for computational analysis of political opinions.

3 Methodology

Our model was trained and tested on the pre-labeled dataset provided by the SemEval group and the basis of our approach was a bag-of-words model. In order to improve upon the bag-of-words model and integrate some known salient features of hyperpartisan news, we also included headline features as well as sentiment analysis scores of the articles.

3.1 Data Set

To train our model, we use training data provided by the SemEval 2019 Hyperpartisan News Detection task organizers (Kiesel et al., 2019). These data come in the form of news articles given in XML format. Each article was given with title and article body text, and had labels provided in a separate file to indicate whether they had been flagged as hyperpartisan.

This data came in two distinct training sets as provided by the task organizers. The larger of the two sets, with about 800,000 labeled articles, was labeled by publisher; that is, publishers were grouped by whether they were known to be hyperpartisan in general, and the corresponding label was applied to all articles by a given publisher. A smaller set, comprising around 650 articles, was entirely hand-labeled; that is, human readers determined on an article-by-article basis whether a given article should be labeled as hyperpartisan.

It should be noted here that the smaller set labels are more true to what is expected of this task. Specifically, it is more of interest to us whether we can detect hyperpartisanship of articles based on how humans would judge it. Though the labeling by publisher is useful for obtaining a larger data set, it introduces some error due to the possibility that hyperpartisan sources may sometimes publish non-hyperpartisan articles and vice versa. Despite the advantages of the hand-labeled data set, its small size makes it much less feasible as a training set, so we also made substantial use of the larger set as we were tuning our model.

The content of each article was pre-processed with the Python library `spaCy` to tokenize and sentence-segment the text (AI, 2016–).

3.2 Feature Extraction

We used a bag-of-words approach to use for our main set of features, using a vocabulary of common English words. Unknown words were ignored. We filtered out 100 stop words, and used a vocabulary of 30,000 words.

We also included features from the titles, as certain qualities about the titles in hyperpartisan articles may be different than mainstream articles. We included the number of words in the title, hypothesizing that long titles can often indicate the arti-

cle is misleading, or very biased. We included the number of fully capitalized words, which can often indicate that the article is not mainstream. Finally, we added a feature for the average length of the words in the title, as some research shows that hyperpartisan or biased articles tend to use more short, easily understood, words to appeal to the average reader (Horne and Adali, 2017).

We used the Python library TextBlob to do sentiment analysis on the articles and their titles (Loria, 2018). TextBlob has a sentiment analysis tool that provides both the subjectivity and polarity of a given sentence. We found the average subjectivity and average polarity of all the sentences in the article and used it as a feature. We also used the sentiment subjectivity and polarity of the article's title as a feature.

As our vocabulary used for extracting bag-of-words features was quite large, we used the built-in `SelectKBest` feature selection class provided by scikit-learn (Pedregosa et al., 2011) to narrow down the set of features we were using. However, testing with adjusting the parameters for feature selection did not seem to yield better results than simply using all possible features, so our final system made used of all of the available features.

3.3 Classifier

We feed our features to a multinomial naive Bayes (NB) classifier in scikit-learn (Pedregosa et al., 2011). For comparison to a baseline, we also use a majority-class dummy classifier.

4 Results

Table 1 shows the results of training with 10-fold cross-validation for each of our classifiers. As expected, the dummy classifier did not perform well. It represents a majority class baseline, though, and is useful for comparison purposes.

Our evaluation metrics improved with adding sentiment analysis and features of the title. Our final model does better than all of our previous models for every metric, including the dummy classifier, multinomial naive Bayes on BoW features, and adding title features. With an F-measure of 0.800, our precision and recall are well-balanced, and both are around 80%.

On the hand labeled SemEval test set, we achieved an accuracy of 0.739 and an F1 score of 0.745. Overall, our system ranked 18th out of 42 by accuracy, and 11th by F1 measure.

5 Discussion

We can infer from the results of our system that the features we extracted from the text, such as title features and sentiment, were significant and correlated to the hyperpartisanship of articles. This was expected, as the design of our classifier was based on previous work which determined that such features were useful for detecting bias in text. Since we combined different aspects of other studies, we were able to build upon previous findings and gain a more holistic view of hyperpartisan news articles Since this is the first offering of the SemEval Hyperpartisan News Detection task, we see our work as providing a foundation for future groups to build on as they attempt to fine-tune and improve a classifier for this important task.

There are still many questions regarding hyperpartisan news identification that remain unanswered. For example, it would be interesting to incorporate bigrams or trigrams of words instead of just using the bag-of-words approach. We also hypothesize that noun phrase chunks would be indicative of hyperpartisanship due to the ubiquity of certain controversial noun phrases in current media. The temporal nature of these features presents a unique challenge, though,

We could also use sentiment analysis in other ways. For instance, instead of just taking the average subjectivity and average polarity, it might also be interesting to find the percentage of sentences with a absolute value of polarity above a certain threshold. This could indicate an article is hyperpartisan if there are a lot of sentences that are above some threshold for polarity, that is, very opinionated sentences either strongly positive or strongly negative.

It would also be interesting to be able to look into the comment sections of the articles and determine if the sentiment of the comments can indicate hyperpartisanship. It seems probable that hyperpartisan articles would tend to attract more hyperpartisan viewers than mainstream articles would, and these people would have similarly strong opinions and be willing to voice them. The alignment of the comments may not even be aligned with the article, as the article may attract people from the other side, looking to critique or complain about the article. This data was not available for the SemEval task, but polarity and subjectivity also seem

Classifier	Accuracy	Precision	Recall	F-measure
DC: most frequent	0.553	0.553	1.0	0.712
NB (BoW only)	0.552	0.61	0.297	0.401
+title features	0.556	0.615	0.315	0.417
+sentiment features	0.793	0.781	0.820	0.800

Table 1: Cross-validation results for dummy classifier and a Naive Bayes classifier using bag of words features with and without additional features related to article title and sentiment.

like they would be useful metrics to extract from article comments.

In addition to the comments, it would be interesting to analyze how often the articles were shared, viewed, commented on, or in other ways interacted with. As the Buzzfeed study showed, hyperpartisan articles and articles that may not be entirely true tended to get more shares than non-partisan, as these are more interesting and inflammatory, and so this may be another feature that would have helped determine the hyperpartisanship of the article (Silverman et al., 2018).

Trying different classifiers would also be an appropriate next step. We focused on feature selection above experimenting with different classifiers because we believed that feature selection would give more meaningful insights into the nature of hyperpartisan articles than merely optimizing a classifier, but both are likely necessary to successfully identifying hyperpartisan articles.

6 Namesake

Our system is named after Carl Kolchak, the main character from the television series *Kolchak: The Night Stalker*, which aired in 1974-75. On the show, Kolchak investigated mysterious cases that had been abandoned by the police. We believe the unlikely and often unbelievable scenarios encountered by Kolchak would have been likely fodder for fake and hyperpartisan news during its time, and hope that our system will contribute to a community effort to automatically separate truth from fiction (Wikipedia contributors, 2019).

Figure 1: Darren McGavin, who portrayed Carl Kolchak in the TV Series *Kolchak: The Night Stalker* (ClassicBecky, 2011).

References

Explosion AI. 2016–. spacy: Industrial-strength natural language processing.

ClassicBecky. 2011. Kolchak: The night stalker ... "the ripper". Accessed: 2019-02-20.

Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *CoRR*, abs/1703.09398.

Gabe Ignatow, Nicholas Evangelopoulos, and Konstantinos Zougris. 2016. *Sentiment Analysis of Polarizing Topics in Social Media: News Site Readers Comments on the Trayvon Martin Controversy*, chapter 10. Emerald Group Publishing Limited.

Michael Kaschesky, Pawel Sobkowicz, and Guillaume Bouchard. 2011. Opinion mining in social media: Modeling, simulating, and visualizing political opinion formation in the web. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, dg.o '11, pages 317–326, New York, NY, USA. ACM.

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 (SemEval 2019)*. Association for Computational Linguistics.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*.

Steven Loria. 2018. Textblob: Simplified text processsing.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *CoRR*, abs/1702.05638.

C. Silverman, L. Strapagiel, Shaban H., E. Hall, and J. Singer-Vine. 2018. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis. Accessed: 2018-12-22.

Wikipedia contributors. 2019. Darren mcgavin — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Darren_McGavin&oldid=883055125. [Online; accessed 19-February-2019].