Harvey Mudd College at SemEval-2019 Task 4: The D.X. Beaumont Hyperpartisan News Detector

Evan Amason
Harvey Mudd College
301 Platt Boulevard
Claremont, CA 91711
eamason@hmc.edu

Jake Palanker
Harvey Mudd College
301 Platt Boulevard
Claremont, CA 91711
jpalanker@hmc.edu

Mary Clare Shen
Harvey Mudd College
301 Platt Boulevard
Claremont, CA 91711
mshen@hmc.edu

Julie Medero
Harvey Mudd College
301 Platt Boulevard
Claremont, CA 91711
jmedero@hmc.edu

Abstract

We use the 600 hand-labelled articles from SemEval Task 4 (Kiesel et al., 2019) to hand-tune a classifier with 3000 features for the Hyperpartisan News Detection task. Our final system uses features based on bag-of-words (BoW), analysis of the article title, language complexity, and simple sentiment analysis in a naive Bayes classifier. We trained our final system on the 600,000 articles labelled by publisher. Our final system has an accuracy of 0.653 on the hand-labeled test set. The most effective features are the Automated Readability Index and the presence of certain words in the title. This suggests that hyperpartisan writing uses a distinct writing style, especially in the title.

1 Introduction

Hyperpartisan news is becoming more mainstream as online sources gain popularity. Hyperpartisan news is news written from an extremely partisan perspective, such that the goal is reinforcing existing belief structures in the party's ideology rather than conveying facts. Such hyperpartisan writing tends to amplify political divisions and increase animosity between opposing political ideologies. Hyperpartisan news sources also output fake news at startling rates (Silverman et al., 2016). Automatic detection of fake news is difficult, but detecting hyperpartisan news can help, and it can also expose biases in journalism. This task is challenging to automate because it is even difficult for humans: fake and biased news articles get shared on social media at high rates, and even labels that were hand-generated by professionals have errors (Silverman et al., 2016). We attempt to use various features of political news articles to train a multinomial Naive Bayes classifier to complete this task. We use a set of bag-of-words (BoW) features for words appearing in the title of each article, and for words appearing in the article text. With these features, we identified a set

of words that characterize hyperpartisan writing. We also considered complexity features such as type-to-token ratio and automated readability index. Based on the performance of these features we attempt to answer the question of whether hyperpartisan writing is more or less complex than non-hyperpartisan writing. A successful classifier could be very useful in today's society. For example, it could be used to create a browser plug-in to check online articles for political bias in real time as the user reads. People on social media could use it to verify the legitimacy of a political article before sharing it with their followers. Encouraging people to share factual news rather than inflammatory hyperpartisan articles would hopefully improve communication between opposing parties and create a more informed population.

The rest of this paper begins with a description of previous work on the related task of fake news detection in Section 2. We then describe our model and features in Section 3, and our results in Section 4. Section 5 discusses some lessons learned with respect to what features are most useful in identifying hyperpartisan news, and Section 6 closes with a brief description of our system's namesake, fictional magazine editor D.X. Beaumont.

2 Previous Work

Since the 2016 election, there has been a lot of interest in fake news, which is closely related to the hyperpartisan news we focus on. Our approach to the hyperpartisan news task leverages lessons learned in prior work on fake news detection, and explores the extent to which that work is successful in a different but related task. Fake news detection has been widely studied (e.g., the survey paper by Fuhr et al. (Fuhr et al., 2018)), and we base many of our classifier's features on previous studies of fake news.

The content of fake and real news articles differ

substantially. Fake news articles have been found to require a lower reading level than real news articles, to be less technical, and to use more personal pronouns. Further, their titles tend to be longer, use more proper nouns, and use more words that are all capitalized (Horne and Adali, 2017). Our work differs in that we were trying to determine whether an article is hyperpartisan, which is similar to but not the same as identifying fake news articles. In particular, a hyperpartisan news article may be factually correct (i.e., not contain any mistruths) but still be written with a hyperpartisan slant. We hypothesize, nonetheless, that the stylistic features that distinguish between real and fake news may be useful in identifying hyperpartisan news articles. Potthast, et. al., also showed that there are significant stylistic differences between hyperpartisan and mainstream news articles (Potthast et al., 2017). Consequently, we include reading level and features of each article's title as features in our model.

The success of these features on identifying fake news motivates our decision to focus on article titles as a differentiating feature, and to include reading level in the set of features available to our model.

Perez-Rosa et al. also examine fake news articles to create a classifier for them (Prez-Rosas et al., 2018). Their results identify additional features related to text readability, with fake news articles tending to be written at a lower reading level than real news articles. We incorporate features from their work, including *Average Word Length*, *Type-Token-Ratio*, and *SMOG Readability Formula* .

3   Methodology

Each article's content and title was tokenized using `spacy`'s default English model (AI, 2016–).

We use a multinomial naive Bayes classifier from `scikit-learn`, extracting a large number of features and then using feature selection to reduce the number of features available to our classifier.

3.1   Features

We make use of features related to the words in the article as a whole, the title of the article, sentiment, and text complexity.

Bag of Words Features: Using a vocabulary of 30,000 words, we count the number of times each vocabulary word occurs in the full article text. We then drop a fixed number stop words, selected automatically by frequency. We experimented with both 50 and 100 stop words, and the run of our system that was submitted to the SemEval task used 50 stop words.

Title Bag of Words: Next, using the same vocabulary but without excluding stop words, we add word counts for the title of the article. We also count the number of words in the title that are entirely capitalized, generally a feature of hyperpartisan titles (Horne and Adali, 2017).

Sentiment Analyzer: We use two sentiment lexicons (Hu and Liu, 2004). The first contains 2000 words with positive sentiment, and the second contains 4000 words with negative sentiment. We count the occurrence of words from each list, hypothesizing that hyperpartisan articles will likely have many more words with polarized sentiment than non-hyperpartisan articles.

Complexity Features: Finally, we include features designed to capture the articles' complexity. This category includes features such as *Average Word Length*, *Type-Token-Ratio*, and *SMOG Readability Formula*. Each of these is designed to capture the complexity of a given text; *Average Word Length* gives us insights into the vocab choices and uses of "advanced" words, *Type-Token-Ratio* measures the amount of "novel" words in the text, the *SMOG Readability Formula* is based on the number of polysyllabic words per sentence (which is influenced both by vocabulary choice, and sentence length). Since prior work shows that hyper-partisan articles are often written at an easier reading level, with more repeating words, and simpler sentence structure, we expect that these complexity features will be useful in identifying hyperpartisan articles.

3.2   Feature Selection

The above feature space was very large compared to the number of available articles, so we implemented two different methods of feature selection: one using variance, and one using a χ^2 test. In each case, we perform statistics on the training set,

attempting to describe which features are the most *distinguishing*. Given these statistics, we score each feature, and select a subset of the total feature set using either a threshold score or a target feature count. By experimenting on the smaller hand-labeled data set, we found that reducing to the best 3000 features maximized our performance for 10-fold cross validation. This modification was made after the evaluation, however; our results on the SemEval task represent the performance of our task without feature selection.

4 Results

Our final system achieved an accuracy of 0.653, which ranked 28th out of 42 submissions on the test set hand-labeled by article.

4.1 Feature Selection

As part of additional analysis, we examined the effectiveness of feature selection on the validation set. Table 1 shows that reducing the number of features to 3000 had a negligible effect on both accuracy and f-measure. Since the validation set is qualitatively different from the hand-labeled test set used in the official competition, these results are not directly comparable to our final system performance. In particular, we note that our system performs slightly better on the validation set than on the test set regardless of the number of features used, which may indicate that our classifier learned some characteristics of the source-labeled validation set that distinguished it from the hand-labeled test set.

Feature Selection	Accuracy	f1-measure
all	0.611	0.675
3000	0.5983	0.667

Table 1: Validation set performance using `all` of our features or the `3000` most informative features.

5 Discussion

Hyperpartisan news has been a concern since the rise of social media, and that concern has only grown since the 2016 election. Giving consumers of social media the knowledge of whether or not what they are reading is hyperpartisan could help to reduce the number of people fooled by fake or misleading facts, and it could help to reduce the partisan divide within the United States.

Feature Title	Category	χ^2	p-value
"trump"	Polarity	416	1.77e-92
A.R.I.	Complexity	377	3.67e-84
"*"	Title	208	2.95e-47
"class"	Title	179	9.45e-41
"american"	Title	170	7.41e-39
"most"	Title	143	5.04e-33
"political"	Title	137	1.08e-31
"israel"	Title	133	1.16e-30
"like"	Polarity	128.7	7.85e-30
"these"	Title	126	2.92e-29

Table 2: Highest ranked features from our hand labeled data-set.

Using our χ^2 feature selection system, we found the top 10 features over the hand-labeled article set, shown in Table 2. The size of the hand-labeled set is rather small, so the extremely small p-values are likely inflated by this.

The Automated Readability Index feature (a complexity feature measuring word length and sentence length) is the second highest performing, indicating that this way of capturing complexity is worthy of further study.

A number of BoW features on the title are also important. The selected words included fall under a few categories such as controversial topic (*trump*, *Israel*), generalization (*most*, *these*), and political terms (*political*, *class*). Some, like the presence of "*" in title, seem like strange outliers that are likely a consequence of a combination of formatting artifacts and the small size of the hand-labeled dataset.

While an earlier, simpler version of our model achieved 10-fold cross-validation accuracy of .787 on the hand-labeled training set, the submission we submitted performed much more poorly on the final test set. We hypothesize that one source of this difference may have been in the tuning of our hyper-parameter related to feature selection. We tuned this parameter manually using results from 10-fold cross-validation on the hand labeled data-set. Because the hand labeled data was significantly smaller, it is possible that it took far fewer features to properly classify the space. Improved tuning of this parameter on a larger set could have given us better results. Nonetheless, our work demonstrates that BoW, complexity, and polarity features are all useful in identifying hyperpartisan news articles.

969

6 Namesake

Our system is named after D.X. Beaumont, a magazine editor and publisher on the short-lived TV Series *My Sister Eileen* that aired on CBS in 1960-61 (Wikipedia contributors, 2018). The series, based on autobiographical short stories published in The New Yorker by Ruth McKenney (Lippman, 2018). Ruth, who aspired to be a writer, worked for Beaumont (shown in Figure 1 as portrayed by Raymond Bailey). We imagine that the proliferation of hyperpartisan news in modern communication would have caused the orderly Ruth a great deal of frustration, and hope that our contribution to this task will benefit future writers and their publishers.

Figure 1: Darren McGavin, who portrayed D.X. Beaumont in the TV Series *My Sister Eileen*(NBC Television, 2017).

Acknowledgments

References

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

Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, YiquN Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. 2018. An information nutritional label for online documents. *SIGIR Forum*, 51(3):46–66.

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. In *The 2nd International Workshop on News and Public Opinion at ICWSM*. Cornell University.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD*.

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.

Laura Lippman. 2018. In praise of ruth mckenney. *The New York Times*. [Online; accessed 18-February-2019].

NBC Television. 2017. Accessed: 2019-02-20 (Public domain). [link].

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

Vernica Prez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. 2016. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate.

Wikipedia contributors. 2018. My sister eileen — Wikipedia, the free encyclopedia. [Online; accessed 18-February-2019].