Headline Evaluation and Style Independency in Journalism

Rafael Arbex-Murut, Yeshwanth Somu

Abstract

The written news media plays a significant role in society. However, it is generally accepted that many newspapers have distinct editorial styles, such as having a political or philosophical bias. Several researchers have used statistical methods and machine learning to quantify these biases. This project uses existing NLP software to quantify certain journalistic practices that have not yet been fully explored.

Our findings indicate that it is possible to measure stylistic idiosyncrasies between newspapers. We measured the propensity of news outlets to have descriptive headlines for their own articles, and on the same dataset, we also measured the uniqueness of their editorial styles. These findings are of interest to journalists, editors, and readers alike, as they provide quantified insights into the modern media environment.

Introduction

While the news media environment has evolved throughout the years with the introduction of technologies such as the radio, television, smartphones, etc., written news media has remained popular. As digitized mediums overtook print and radio, newspapers have adapted. Of the 86% of U.S. adults who get their news from digital devices, 67% get the news from news websites or apps, 49% from social media, and 71% from ‘Search’ [1]. Newspapers have a presence on all digital mediums, and paid digital-only subscribers total in the tens of millions for the top English-language papers [8].

However, distrust in the media has also risen in recent years. Two-thirds of Americans trust the media “Not very much” or “Not at all” [9]. A common explanation for this is that the media is biased [10]. Researchers have attempted to measure biases with varying levels of success, and we discuss this further in the next section.

Our motivation for this project was to provide insight into the media environment by applying existing models and NLP techniques. Our project borrows elements from bias detection, sentiment analysis, author identification, and other techniques to provide this insight.

We intended on quantifying the journalistic practices of certain news outlets that have not been fully explored, thus providing the reader with the proper tools to navigate the media landscape.

Background

Previous work has been done on analyzing journalistic datasets with natural language processing techniques. Our project is alike and different to these studies in key ways.

D'Alonzo and Tegmark [2] is a 2021 MIT study that presented an automated method for measuring the political bias of different news sources. This study analyzed phrases used by newspapers and used the data collected to place newspapers on a traditional left-right axis of bias, as well as an establishment bias (the amount of affinity that a newspaper has for the status quo). We share the assumption made in D'Alonzo and Tegmark that it is worthwhile to analyze news outlets as singular entities. Though, we opted to measure other journalistic practices that are not traditional political biases. In our case, we are measuring headline accuracy and style distinction. However, we accept that traditional political biases may be an influencing factor on our measurements.

Abbasi et al. [4] published a joint paper that uses ensemble learning to detect and classify the authors of texts. This project was conducted on the same dataset used for our project. The difference is that our project conducts authorship identification on the basis of the publications themselves, not the individual writers within them. Interestingly, Abassi et al. effectively do the same in some cases, as the 'author' values in this dataset are oftentimes the entire editorial board of the respective news outlet.

Liu et al. [3] published a joint paper on detecting fraudulent headlines. The researchers developed a deep learning model with attention to classify headlines as either 'True' or 'Fraudulent' and compared their performance against the real labels. Likewise, we investigate headlines under the premise that they are an important journalistic practice. Though we do not share the assumption that headlines are binarily 'True' of 'Fraudulent', and we take issue with how the researchers assigned these values in the first place. Rather, we work with the assumption that accuracy lies on a spectrum and that a relative measure between news sources is a better evaluation of headlines.

Bukhtiyarov and Gusev [5] is a unique study that uses fine-tuned models to generate abstractive headlines that are compared to the real ones by Rouge scores. We are also interested in generating headlines, but not for the purpose of achieving accuracy. Rather, we wish to analyze trends within and between publications. Furthermore, their models were trained on a Russian news corpus as opposed to an English corpus.

Data

Data set description

There are numerous publicly available datasets that aggregate written news media. This project required a non-topic-specific dataset (many existing public datasets focus on one topic, such as financial or environmental news). We also limited our project to articles written in English and published within the last decade.

Therefore, we opted to use the *All the News* [11] data set available on Kaggle. This dataset includes over 140,000 news articles from 15 different American publications. The dataset was built from RSS feeds, thus capturing the most prominent articles of each publication.

This dataset includes 11 attributes, including title, publication, author, date, year, month, url, and content. It is of high quality and has been used in several publications.

Data cleaning

Before commencing the project, we cleaned the database in two steps. First, we normalized the data in terms of article length and year of publication. We kept articles that were between 100 and 500 words in length to remove abnormally short articles (such as breaking news) and longer articles (often opinionated essays) and for ease of computation. We only kept articles published in 2015 or more recently.

Second, we removed self-references from headlines. In some cases, we removed prependings or appendings that might bias the classification model and distort headline evaluations. For example, all New York Times articles were prepended with the text " - The New York Times." In the remaining cases, we removed self-referencing rows entirely if they did not follow an identifiable pattern.

Methods (design and implementation)

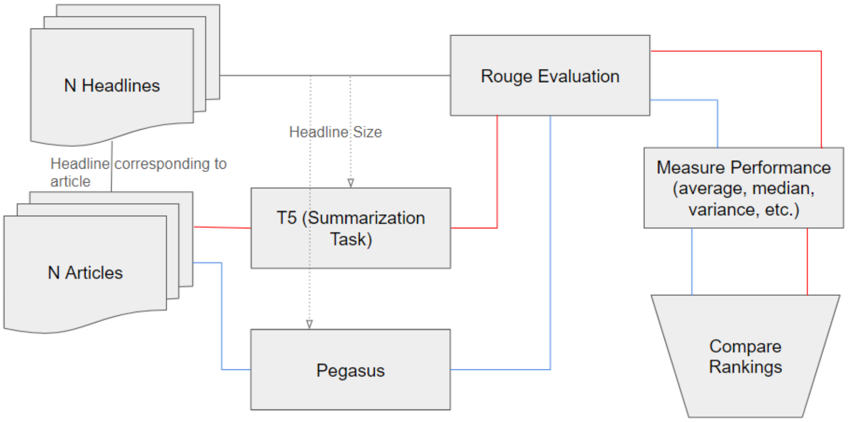
Headline Evaluation

The headline evaluation portion of this project is a new algorithm that applies existing NLP tools to measure how much a headline deviates from the contents of its own article.

This is done by producing a summary of the article using a summarizing language model (or a language model capable of summarizing) and subsequently evaluating the generated summary against the real headline, which is used as the reference. In essence, we are attempting to capture how well a headline describes its own article.

We analyzed how the news sources performed relative to each other. To standardize these rankings, we ran the same parameterized model on a subset of 100 articles from each publication and then collected evaluation performance. In an attempt to validate these relative rankings, we ‘triangulate’ the data by applying the same process a second time with a different model using the same parameters. The models used were T5-base and Pegasus.

A visual representation of this algorithm:



This raises the question of which parameters ought to be used. To determine this, we first bound the length of the produced summaries to three discrete intervals based on the token length of the respective headline, under the premise that headlines can be thought of as summaries of their articles.

$$token\_length(Real\ Headline) + x \geq token\_length(Generated\ Summary)$$
$$token\_length(Real\ Headline) - x \leq token\_length(Generated\ Summary)$$
$$x = \{0, 5, 10\}$$

The next step was to determine the number of beams (`num_beams`) and the number of no-repeat n-grams (`no_repeat_ngram_size`), two parameters accepted by T5 and Pegasus.

To accomplish this methodically, we sampled five random articles from each of the fifteen publications and ran the algorithm on every possible permutation of 3 ≤ `num_beams` ≤ 5 and 3 ≤ `no_repeat_ngram_size` ≤ 5.

We selected the combination with a high variance and high average/median score under the premise that some headlines should perform better than others. The parameter determination tables are included in the appendix.

Once the parameters were determined, the aforementioned algorithm was executed on the *All the News* dataset, and the results are discussed in the next section.

Style Analysis

Style analysis can be thought of as a type of author identification, with the key difference being that entire news outlets are seen as a single author. Abbasi et al. used an ensemble approach that combines a BERT model with XGBoost to analyze term frequency-inverse document frequencies that automatically learn features without human interference [4].

Our proposed approach was simpler in that we trained a BERT model to predict the news outlet of the articles. We then observed the confusion matrix, shown in the Results section, for misclassification of articles. Our hypothesis was that a high rate of accurate classification is indicative of a unique editorial style that distinguishes a publication from other outlets. Conversely, if an outlet's articles are highly misclassified, there is less 'uniqueness' in that outlet's [7] writing style. Under this hypothesis, high confusion between articles could also be indicative of style overlap.

Therefore, we were more concerned with the correct identification of writing styles than achieving a high degree of accuracy.

Results and discussion

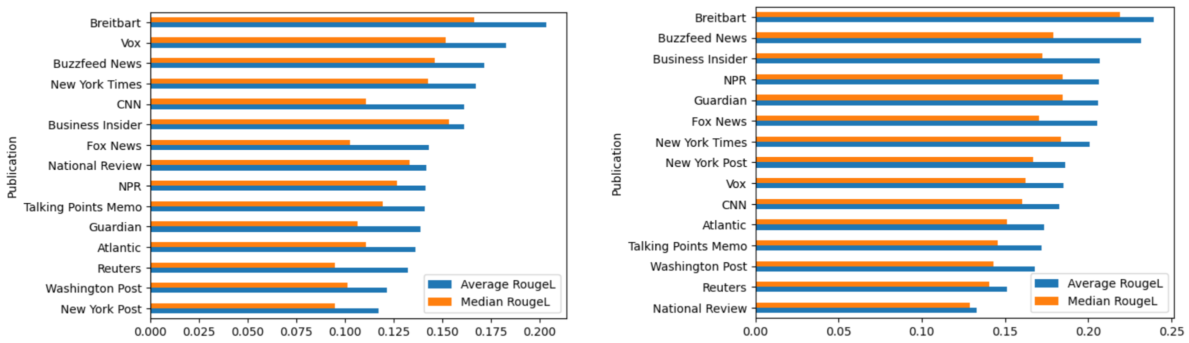
Headline Evaluation

The algorithm produced comparable results for both models. Pegasus resulted in average RougeL scores that were only about 0.04 higher on average than T5.

Breitbart, Buzzfeed News, and Business Insider ranked highly on both models, as measured by RougeL. This is indicative of their headlines being relatively more descriptive summaries of their articles.

Other news outlets, such as Reuters, the Washington Post, and the Atlantic, ranked low in both models. Inferentially, this is indicative of their headlines being relatively less descriptive summaries of their articles.

However, some publications, such as the Guardian and National Review, had noticeably different performances, varying over 5 spots in ranking over the two models, thus making it difficult to triangulate their performance.

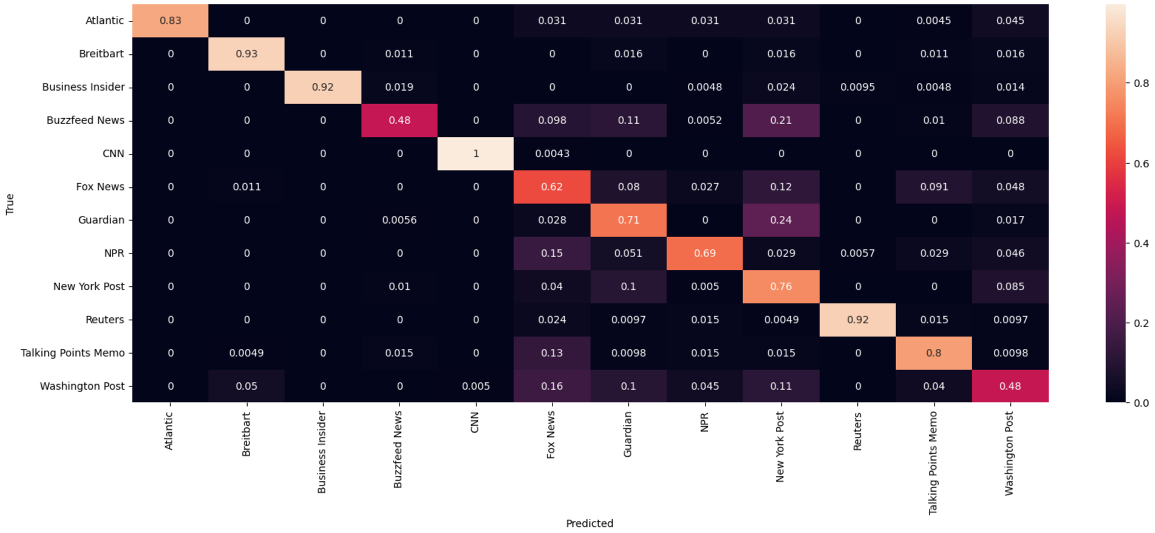


Full tabulated results are documented in the appendix, including scores for Rouge1, Rouge2, and variance scores.

Style Analysis

Text classification tasks have been conducted in numerous research papers. In this portion of the project, we apply a well-known model to a new dataset. To prevent majority class prediction, we limited this exercise to 12 publications that had at least 1000 articles.

After testing and fine-tuning, we ran a bert-base-cased model with a final softmax classification layer. Full details and parameters of the final model are included in the appendix.



	Atlantic	Breitbart	Business Insider	Buzzfeed News	CNN	Fox News	Guardian	NPR	New York Post	Reuters	Talking Points Memo	Washington Post
Atlantic	0.83	0	0	0	0	0.031	0.031	0.031	0.031	0	0.0045	0.045
Breitbart	0	0.93	0	0.011	0	0	0.016	0	0.016	0	0.011	0.016
Business Insider	0	0	0.92	0.019	0	0	0	0.0048	0.024	0.0095	0.0048	0.014
Buzzfeed News	0	0	0	0.48	0	0.098	0.11	0.0052	0.21	0	0.01	0.088
CNN	0	0	0	0	1	0.0043	0	0	0	0	0	0
Fox News	0	0.011	0	0	0	0.62	0.08	0.027	0.12	0	0.091	0.048
Guardian	0	0	0	0.0056	0	0.028	0.71	0	0.24	0	0	0.017
NPR	0	0	0	0	0	0.15	0.051	0.69	0.029	0.0057	0.029	0.046
New York Post	0	0	0	0.01	0	0.04	0.1	0.005	0.76	0	0	0.085
Reuters	0	0	0	0	0	0.024	0.0097	0.015	0.0049	0.92	0.015	0.0097
Talking Points Memo	0	0.0049	0	0.015	0	0.13	0.0098	0.015	0.015	0	0.8	0.0098
Washington Post	0	0.05	0	0	0.005	0.16	0.1	0.045	0.11	0	0.04	0.48

Model 1: 12 Outlet Classification with 1000 news articles per outlet

Buzzfeed and Washington post were predicted accurately less than 50% of the time. This is indicative of those news outlets having non-unique editorial styles given the high rate of confusion with other outlets. These findings reinforce existing knowledge about Buzzfeed and The Washington Post, platforms known for their diverse content spanning various topics and authored by individuals with diverse backgrounds.

Conversely, Breitbart, Business Insider, CNN, and Reuters seem to have strongly unique writing styles, with the model correctly classifying their articles with over 90% accuracy. A news outlet such as CNN, which reports on a wide variety of topics and has a unique writing style, was surprising to observe.

The remaining news outlets performed in the 62%–83% range, not quite high enough to display a strongly unique writing style.

We did a follow-up investigation by reducing the number of news outlets to test whether we see similar results among a subset of the outlets. We filtered to 6 outlets that had at least 1000 articles. We observed a high level of accuracy in the test set for all classes except New York Post which was slightly misclassified as Breitbart, as documented in the appendix. Nevertheless, this low incidence of confusion did not provide any insight to the ‘uniqueness’ of each publication. We conclude that reducing the number of outlets for prediction is not conducive for determining writing style.

Conclusion

Our intention behind this project was to use NLP algorithms in descriptive ways to find trends and patterns in journalism that are of practical and linguistic interest.

This project is unique in the sense that we do not use linguistic machine learning tools in the traditional sense of aiming for a high accuracy while appropriately fitting the data. Rather, we employed existing models and methods in an analytic manner to generate heuristic results that readers, editors, and other stakeholders can reference to gain insight into written journalism.

We were willing to accept a high variance in both steps of the project. Varied rouge scores may be inferred as some headlines outperforming others in the headline evaluation portion of this project. Varied classification accuracies might similarly be understood as some publications having more ‘identifiable’ editorial styles than others.

We intend for our findings to be used as a baseline for future research into media bias.

CITATIONS

[1] Shearer, E. (2021, January 12). 86% of Americans get news online from smartphone, computer or tablet. Pew Research Center. Retrieved December 8, 2023, from https://www.pewresearch.org/short-reads/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

[2] D'Alonzo, S., & Tegmark, M. (2021). *Machine-Learning media bias*. Retrieved from http://arxiv.org/abs/2109.00024

[3] Deep learning model for detecting if an article is fraudulent- Liu, H., He, D., & Chan, S. (2021). *Fraudulent News Headline Detection with Attention Mechanism*. Computational intelligence and neuroscience, 2021, 6679661. https://doi.org/10.1155/2021/6679661

[4] Classification of news articles' authors- Abbasi, A., Javed, A.R., Iqbal, F. et al. *Authorship identification using ensemble learning*. Sci Rep 12, 9537 (2022). https://doi.org/10.1038/s41598-022-13690-4

[5] Bukhtiyarov, A., & Gusev, I. (2020). *Advances of Transformer-based models for news headline generation*. Retrieved from https://arxiv.org/pdf/2007.05044.pdf

[6] Qian, C., He, T., & Zhang, R. (2017). *Deep Learning based Authorship Identification*.https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf

[7] Khudhair Hassoon, Fatimah. (2018). *A Stylistic Analysis of Selected Newspaper's Stories*. https://www.researchgate.net/publication/324455115_A_Stylistic_Analysis_of_Selected_Newspaper's_Stories

[8] Watson, Amy. "Digital News Subscriptions Worldwide 2023." Statista, 23 Aug. 2023, www.statista.com/statistics/785919/worldwide-number-of-digital-newspaper-subscribers/

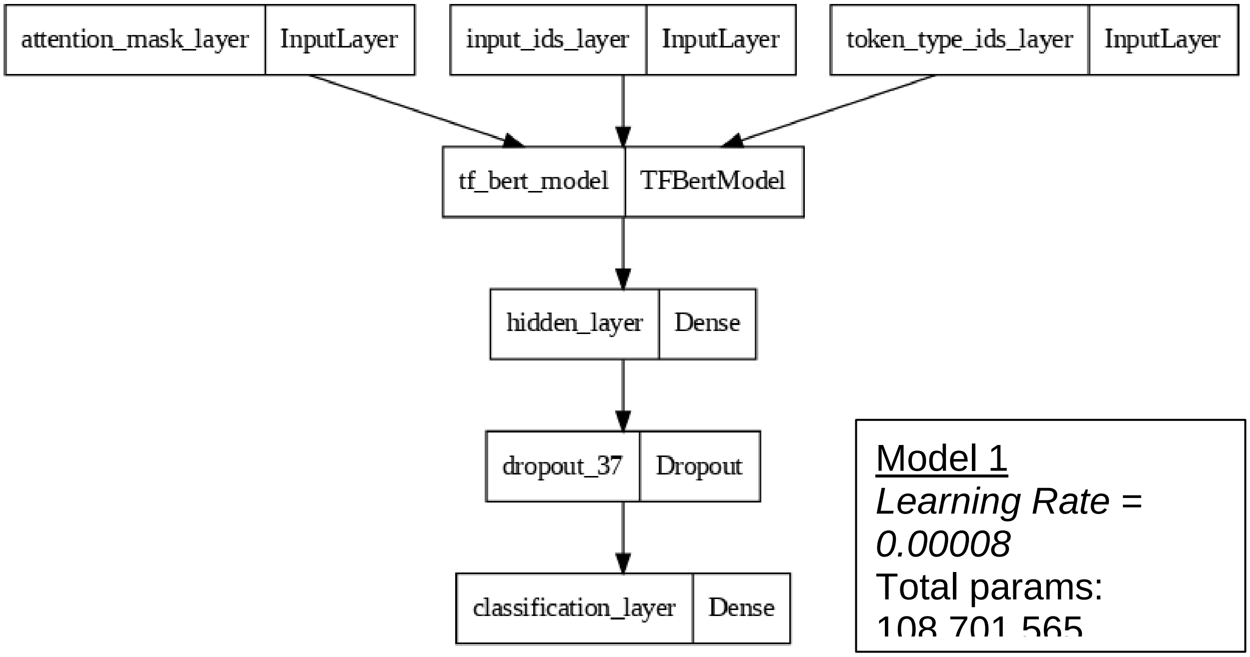
[9] Brenan, Megan. "Americans' Trust in Media Remains near Record Low." Gallup.Com, Gallup, 5 June 2023, https://news.gallup.com/poll/403166/americans-trust-media-remains-near-record-low.aspx#:~:text=The%20percentage%20of%20Americans%20with%20a%20great%20deal%20or%20fair,has%20averaged%2042%25%20since%202004.

[10] Lewis, S. C. (2019). *Lack of trust in the news media, institutional weakness, and relational journalism as a potential way forward*. Journalism, 20(1), 44-47. https://doi.org/10.1177/1464884918808134

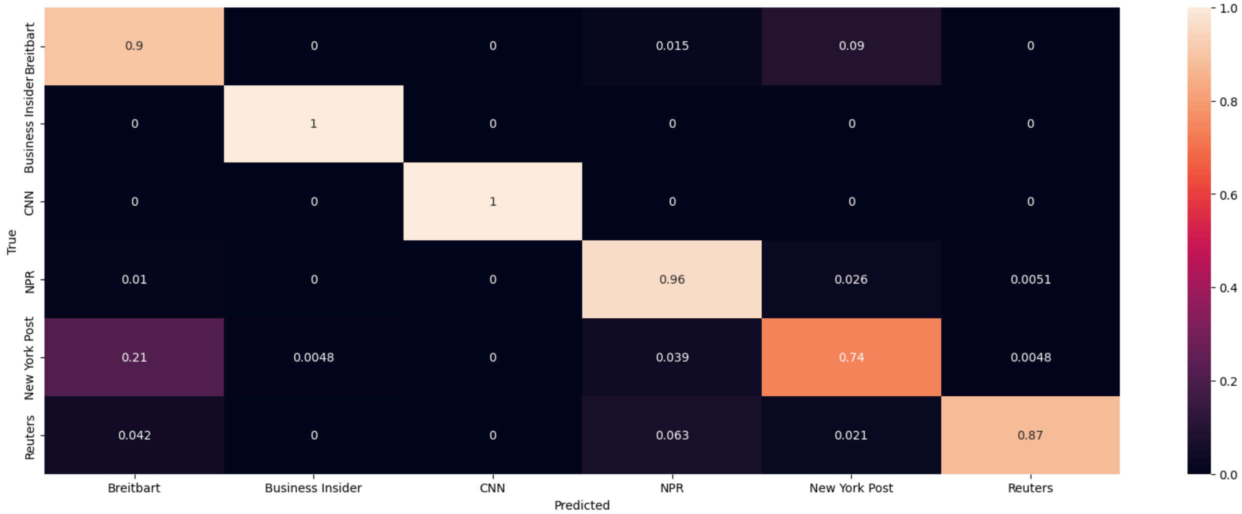
[11] All the news. (n.d.). Kaggle, from https://www.kaggle.com/datasets/snapcrack/all-the-news

Appendix

News Outlet BERT model architecture

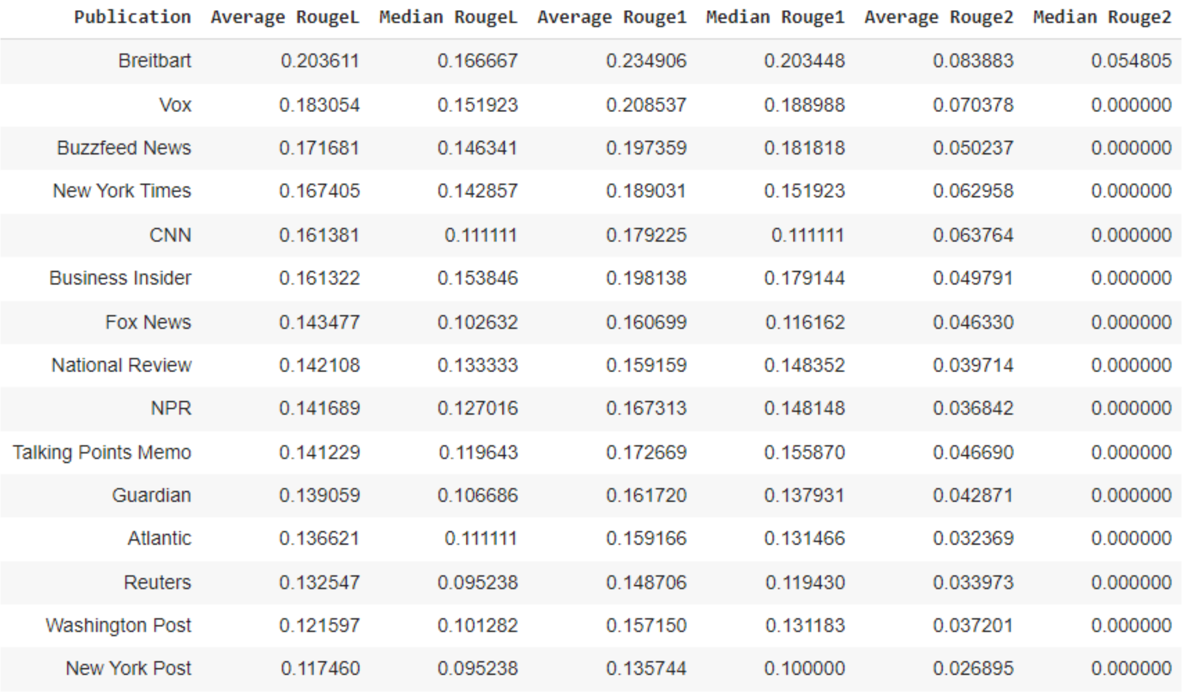


Model 1: (12 news outlets with 1000 articles per outlet) - Architecture and Parameters

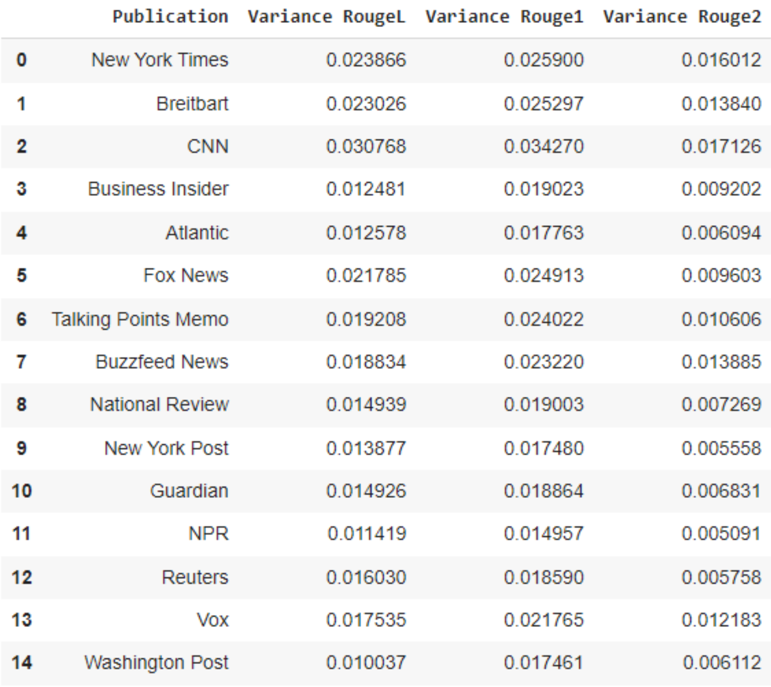


Model 2: 6 News Outlets Classification with 1000 news articles per outlet (identical model architecture to Model 1)

T5 Full Results

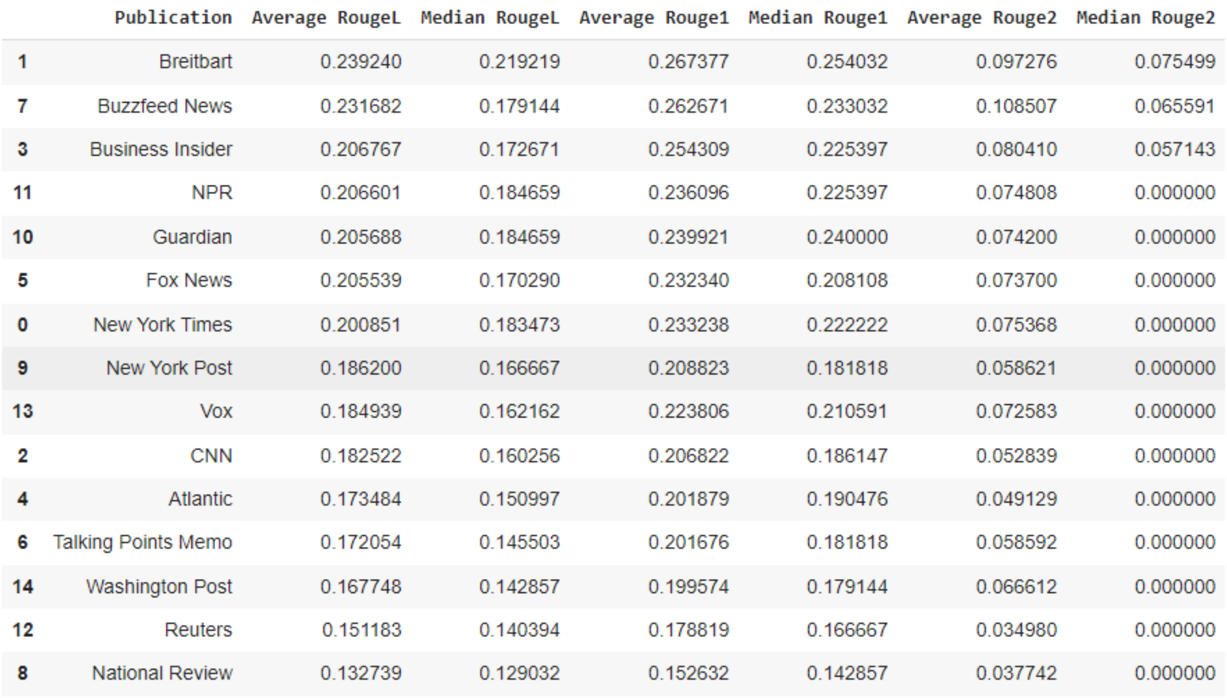


Publication	Average RougeL	Median RougeL	Average Rouge1	Median Rouge1	Average Rouge2	Median Rouge2
Breitbart	0.203611	0.166667	0.234906	0.203448	0.083883	0.054805
Vox	0.183054	0.151923	0.208537	0.188988	0.070378	0.000000
Buzzfeed News	0.171681	0.146341	0.197359	0.181818	0.050237	0.000000
New York Times	0.167405	0.142857	0.189031	0.151923	0.062958	0.000000
CNN	0.161381	0.111111	0.179225	0.111111	0.063764	0.000000
Business Insider	0.161322	0.153846	0.198138	0.179144	0.049791	0.000000
Fox News	0.143477	0.102632	0.160699	0.116162	0.046330	0.000000
National Review	0.142108	0.133333	0.159159	0.148352	0.039714	0.000000
NPR	0.141689	0.127016	0.167313	0.148148	0.036842	0.000000
Talking Points Memo	0.141229	0.119643	0.172669	0.155870	0.046690	0.000000
Guardian	0.139059	0.106686	0.161720	0.137931	0.042871	0.000000
Atlantic	0.136621	0.111111	0.159166	0.131466	0.032369	0.000000
Reuters	0.132547	0.095238	0.148706	0.119430	0.033973	0.000000
Washington Post	0.121597	0.101282	0.157150	0.131183	0.037201	0.000000
New York Post	0.117460	0.095238	0.135744	0.100000	0.026895	0.000000

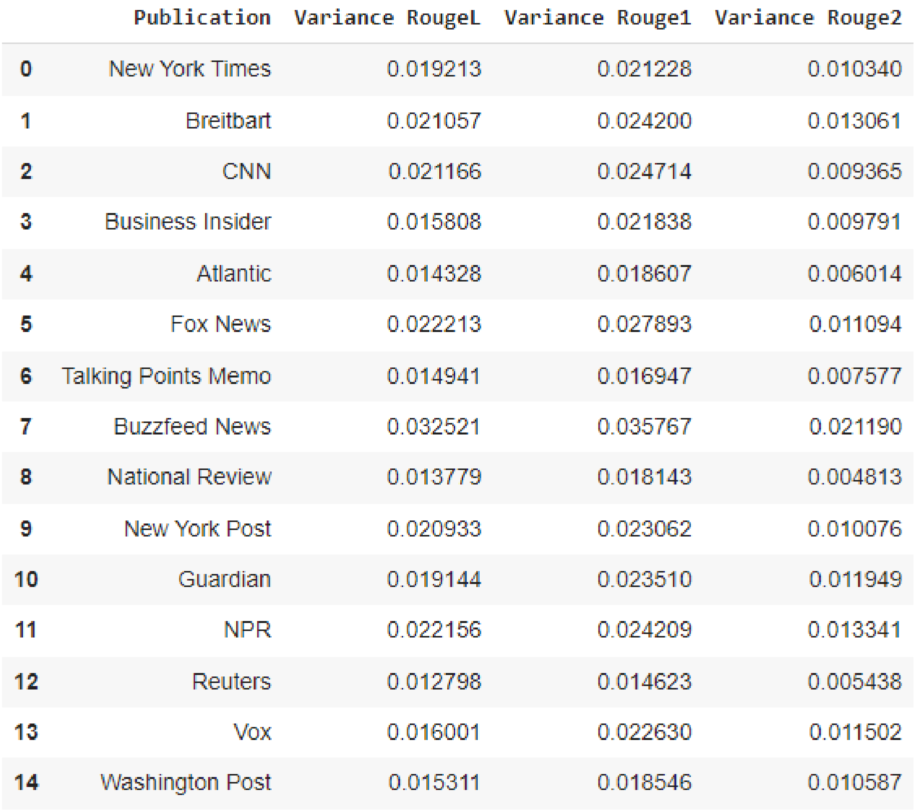


	Publication	Variance RougeL	Variance Rouge1	Variance Rouge2
0	New York Times	0.023866	0.025900	0.016012
1	Breitbart	0.023026	0.025297	0.013840
2	CNN	0.030768	0.034270	0.017126
3	Business Insider	0.012481	0.019023	0.009202
4	Atlantic	0.012578	0.017763	0.006094
5	Fox News	0.021785	0.024913	0.009603
6	Talking Points Memo	0.019208	0.024022	0.010606
7	Buzzfeed News	0.018834	0.023220	0.013885
8	National Review	0.014939	0.019003	0.007269
9	New York Post	0.013877	0.017480	0.005558
10	Guardian	0.014926	0.018864	0.006831
11	NPR	0.011419	0.014957	0.005091
12	Reuters	0.016030	0.018590	0.005758
13	Vox	0.017535	0.021765	0.012183
14	Washington Post	0.010037	0.017461	0.006112

Pegasus Full Results



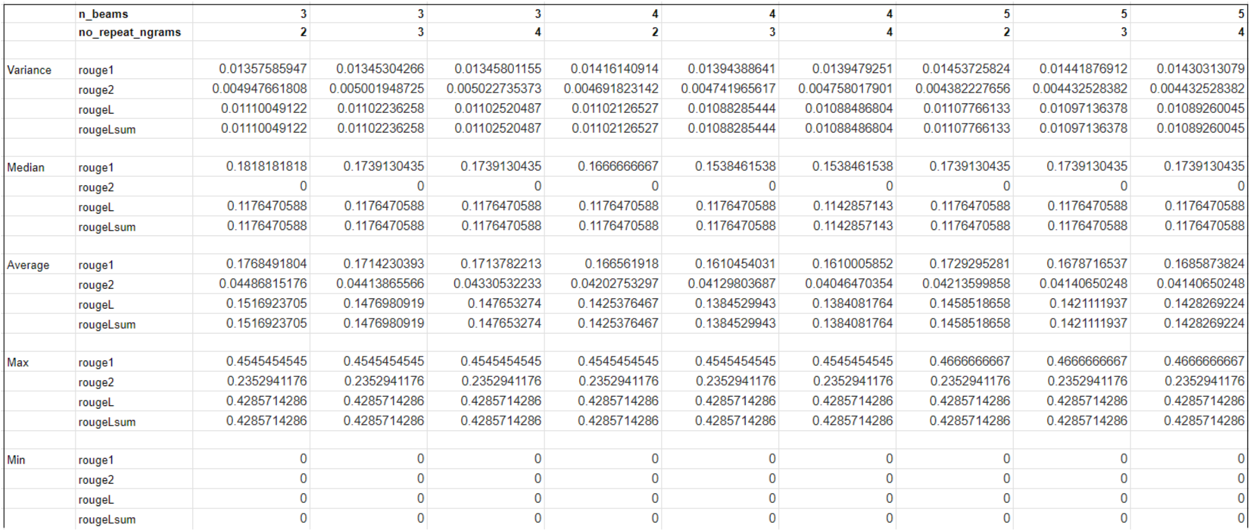
	Publication	Average RougeL	Median RougeL	Average Rouge1	Median Rouge1	Average Rouge2	Median Rouge2
1	Breitbart	0.239240	0.219219	0.267377	0.254032	0.097276	0.075499
7	Buzzfeed News	0.231682	0.179144	0.262671	0.233032	0.108507	0.065591
3	Business Insider	0.206767	0.172671	0.254309	0.225397	0.080410	0.057143
11	NPR	0.206601	0.184659	0.236096	0.225397	0.074808	0.000000
10	Guardian	0.205688	0.184659	0.239921	0.240000	0.074200	0.000000
5	Fox News	0.205539	0.170290	0.232340	0.208108	0.073700	0.000000
0	New York Times	0.200851	0.183473	0.233238	0.222222	0.075368	0.000000
9	New York Post	0.186200	0.166667	0.208823	0.181818	0.058621	0.000000
13	Vox	0.184939	0.162162	0.223806	0.210591	0.072583	0.000000
2	CNN	0.182522	0.160256	0.206822	0.186147	0.052839	0.000000
4	Atlantic	0.173484	0.150997	0.201879	0.190476	0.049129	0.000000
6	Talking Points Memo	0.172054	0.145503	0.201676	0.181818	0.058592	0.000000
14	Washington Post	0.167748	0.142857	0.199574	0.179144	0.066612	0.000000
12	Reuters	0.151183	0.140394	0.178819	0.166667	0.034980	0.000000
8	National Review	0.132739	0.129032	0.152632	0.142857	0.037742	0.000000



	Publication	Variance RougeL	Variance Rouge1	Variance Rouge2
0	New York Times	0.019213	0.021228	0.010340
1	Breitbart	0.021057	0.024200	0.013061
2	CNN	0.021166	0.024714	0.009365
3	Business Insider	0.015808	0.021838	0.009791
4	Atlantic	0.014328	0.018607	0.006014
5	Fox News	0.022213	0.027893	0.011094
6	Talking Points Memo	0.014941	0.016947	0.007577
7	Buzzfeed News	0.032521	0.035767	0.021190
8	National Review	0.013779	0.018143	0.004813
9	New York Post	0.020933	0.023062	0.010076
10	Guardian	0.019144	0.023510	0.011949
11	NPR	0.022156	0.024209	0.013341
12	Reuters	0.012798	0.014623	0.005438
13	Vox	0.016001	0.022630	0.011502
14	Washington Post	0.015311	0.018546	0.010587

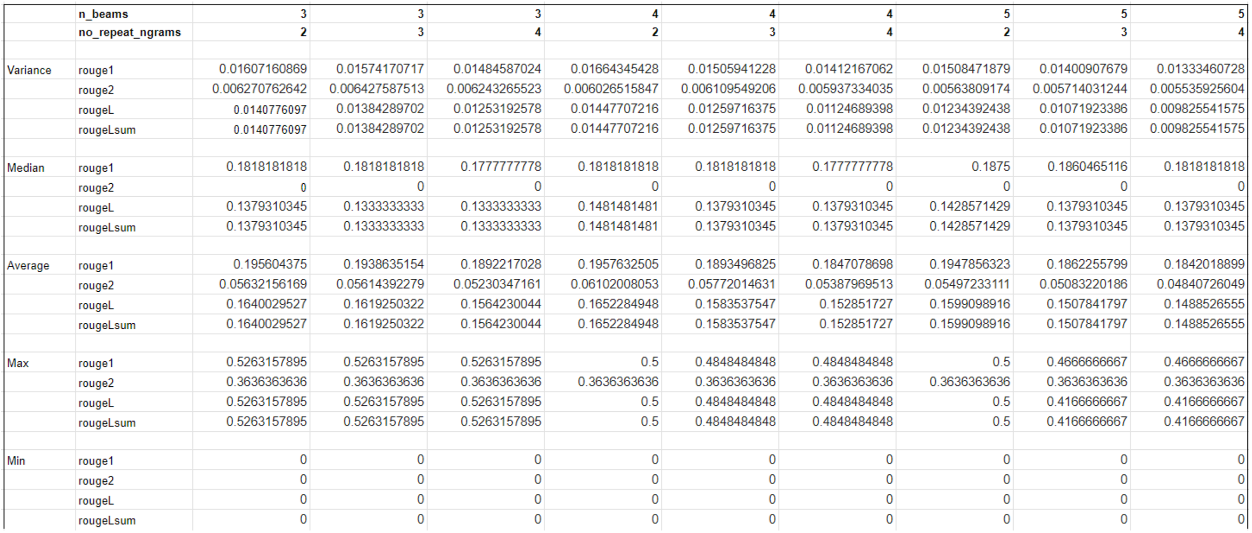
Parameter Determination Tables

T5 Token Range 0



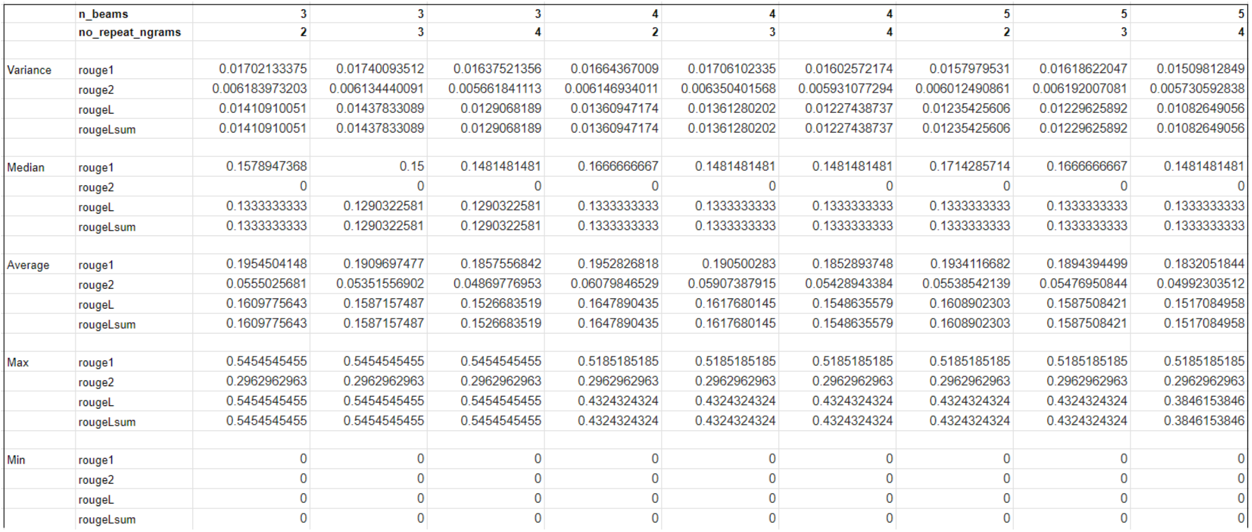
	n_beams	3	3	3	4	4	4	5	5	5
	no_repeat_ngrams	2	3	4	2	3	4	2	3	4
Variance	rouge1	0.01357585947	0.01345304266	0.01345801155	0.01416140914	0.01394388641	0.0139479251	0.01453725824	0.01441876912	0.01430313079
	rouge2	0.004947661808	0.005001948725	0.005022735373	0.004691823142	0.004741965617	0.004758017901	0.004382227656	0.004432528382	0.004432528382
	rougeL	0.01110049122	0.01102236258	0.01102520487	0.01102126527	0.01088285444	0.01088486804	0.01107766133	0.01097136378	0.01089260045
	rougeLsum	0.01110049122	0.01102236258	0.01102520487	0.01102126527	0.01088285444	0.01088486804	0.01107766133	0.01097136378	0.01089260045
Median	rouge1	0.1818181818	0.1739130435	0.1739130435	0.1666666667	0.1538461538	0.1538461538	0.1739130435	0.1739130435	0.1739130435
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0.1176470588	0.1176470588	0.1176470588	0.1176470588	0.1176470588	0.1142857143	0.1176470588	0.1176470588	0.1176470588
	rougeLsum	0.1176470588	0.1176470588	0.1176470588	0.1176470588	0.1176470588	0.1142857143	0.1176470588	0.1176470588	0.1176470588
Average	rouge1	0.1768491804	0.1714230393	0.1713782213	0.166561918	0.1610454031	0.1610005852	0.1729295281	0.1678716537	0.1685873824
	rouge2	0.04486815176	0.04413865566	0.04330532233	0.04202753297	0.04129803687	0.04046470354	0.04213599858	0.04140650248	0.04140650248
	rougeL	0.1516923705	0.1476980919	0.147653274	0.1425376467	0.1384529943	0.1384081764	0.1458518658	0.1421111937	0.1428269224
	rougeLsum	0.1516923705	0.1476980919	0.147653274	0.1425376467	0.1384529943	0.1384081764	0.1458518658	0.1421111937	0.1428269224
Max	rouge1	0.4545454545	0.4545454545	0.4545454545	0.4545454545	0.4545454545	0.4545454545	0.4666666667	0.4666666667	0.4666666667
	rouge2	0.2352941176	0.2352941176	0.2352941176	0.2352941176	0.2352941176	0.2352941176	0.2352941176	0.2352941176	0.2352941176
	rougeL	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286
	rougeLsum	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286	0.4285714286
Min	rouge1	0	0	0	0	0	0	0	0	0
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0	0	0	0	0	0	0	0	0
	rougeLsum	0	0	0	0	0	0	0	0	0

T5 Token Range 5



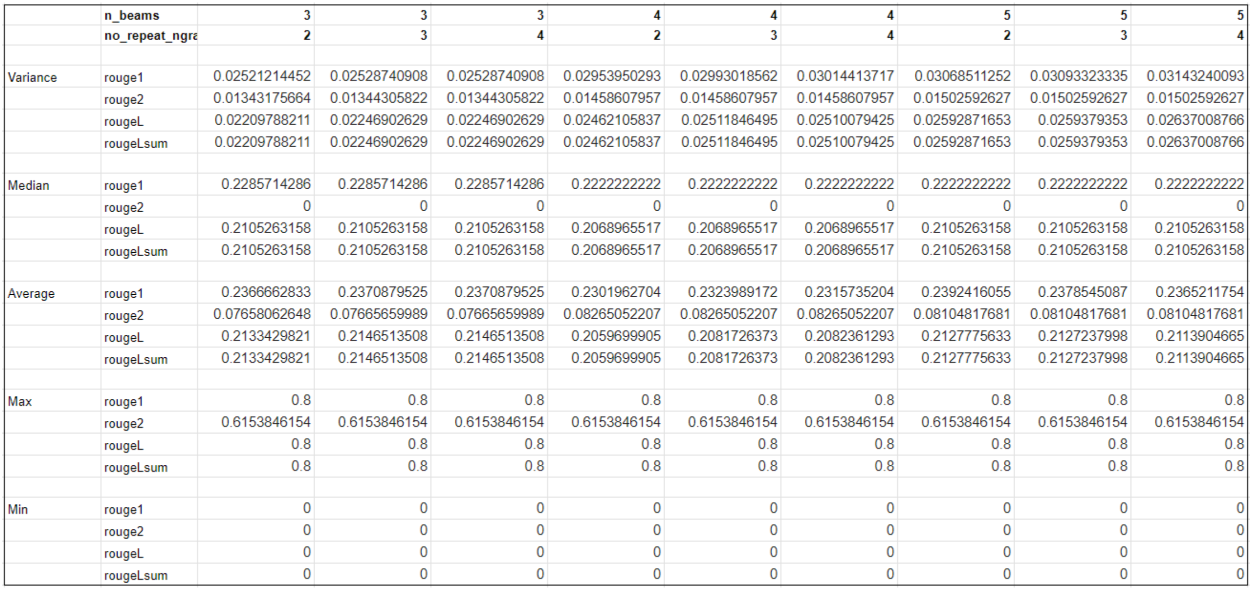
	n_beams	3	3	3	4	4	4	5	5	5
	no_repeat_ngrams	2	3	4	2	3	4	2	3	4
Variance	rouge1	0.01607160869	0.01574170717	0.01484587024	0.01664345428	0.01505941228	0.01412167062	0.01508471879	0.01400907679	0.01333460728
	rouge2	0.006270762642	0.006427587513	0.006243265523	0.006026515847	0.006109549206	0.005937334035	0.00563809174	0.005714031244	0.005535925604
	rougeL	0.0140776097	0.01384289702	0.01253192578	0.01447707216	0.01259716375	0.01124689398	0.01234392438	0.01071923386	0.009825541575
	rougeLsum	0.0140776097	0.01384289702	0.01253192578	0.01447707216	0.01259716375	0.01124689398	0.01234392438	0.01071923386	0.009825541575
Median	rouge1	0.1818181818	0.1818181818	0.1777777778	0.1818181818	0.1818181818	0.1777777778	0.1875	0.1860465116	0.1818181818
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0.1379310345	0.1333333333	0.1333333333	0.1481481481	0.1379310345	0.1379310345	0.1428571429	0.1379310345	0.1379310345
	rougeLsum	0.1379310345	0.1333333333	0.1333333333	0.1481481481	0.1379310345	0.1379310345	0.1428571429	0.1379310345	0.1379310345
Average	rouge1	0.195604375	0.1938635154	0.1892217028	0.1957632505	0.1893496825	0.1847078698	0.1947856323	0.1862255799	0.1842018899
	rouge2	0.05632156169	0.05614392279	0.05230347161	0.06102008053	0.05772014631	0.05387969513	0.05497233111	0.05083220186	0.04840726049
	rougeL	0.1640029527	0.1619250322	0.1564230044	0.1652284948	0.1583537547	0.152851727	0.1599098916	0.1507841797	0.1488526555
	rougeLsum	0.1640029527	0.1619250322	0.1564230044	0.1652284948	0.1583537547	0.152851727	0.1599098916	0.1507841797	0.1488526555
Max	rouge1	0.5263157895	0.5263157895	0.5263157895	0.5	0.4848484848	0.4848484848	0.5	0.4666666667	0.4666666667
	rouge2	0.3636363636	0.3636363636	0.3636363636	0.3636363636	0.3636363636	0.3636363636	0.3636363636	0.3636363636	0.3636363636
	rougeL	0.5263157895	0.5263157895	0.5263157895	0.5	0.4848484848	0.4848484848	0.5	0.4166666667	0.4166666667
	rougeLsum	0.5263157895	0.5263157895	0.5263157895	0.5	0.4848484848	0.4848484848	0.5	0.4166666667	0.4166666667
Min	rouge1	0	0	0	0	0	0	0	0	0
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0	0	0	0	0	0	0	0	0
	rougeLsum	0	0	0	0	0	0	0	0	0

T5 Token Range 10



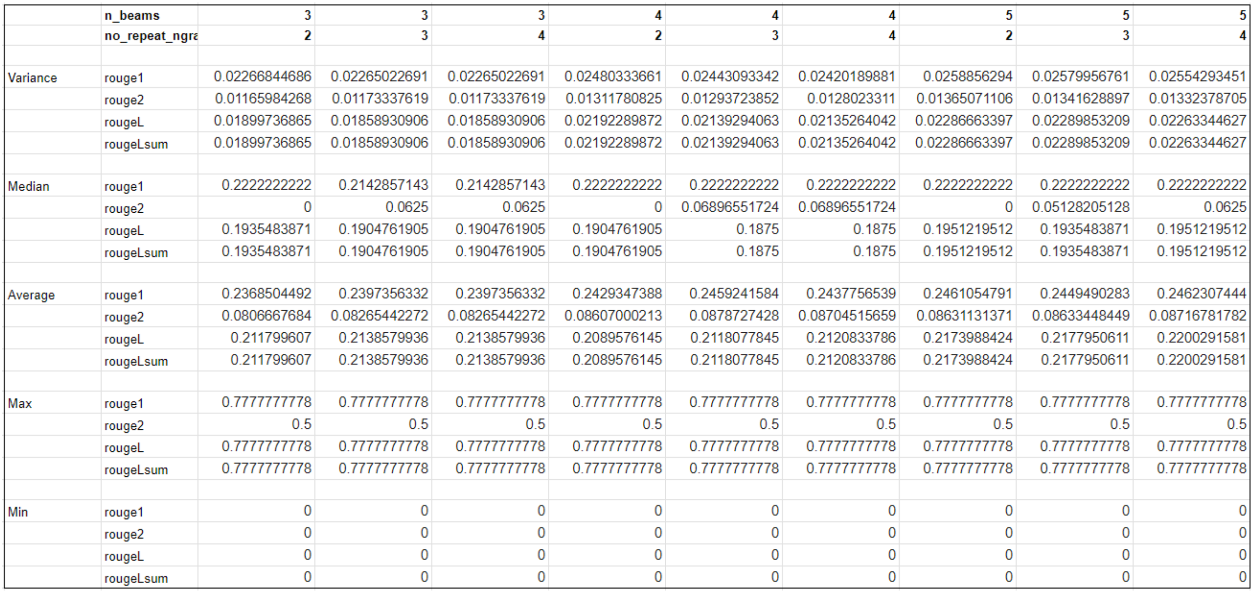
	n_beams	3	3	3	4	4	4	5	5	5
	no_repeat_ngrams	2	3	4	2	3	4	2	3	4
Variance	rouge1	0.01702133375	0.01740093512	0.01637521356	0.01664367009	0.01706102335	0.01602572174	0.0157979531	0.01618622047	0.01509812849
	rouge2	0.006183973203	0.006134440091	0.005661841113	0.006146934011	0.006350401568	0.005931077294	0.006012490861	0.006192007081	0.005730592838
	rougeL	0.01410910051	0.01437833089	0.0129068189	0.01360947174	0.01361280202	0.01227438737	0.01235425606	0.01229625892	0.01082649056
	rougeLsum	0.01410910051	0.01437833089	0.0129068189	0.01360947174	0.01361280202	0.01227438737	0.01235425606	0.01229625892	0.01082649056
Median	rouge1	0.1578947368	0.15	0.1481481481	0.1666666667	0.1481481481	0.1481481481	0.1714285714	0.1666666667	0.1481481481
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0.1333333333	0.1290322581	0.1290322581	0.1333333333	0.1333333333	0.1333333333	0.1333333333	0.1333333333	0.1333333333
	rougeLsum	0.1333333333	0.1290322581	0.1290322581	0.1333333333	0.1333333333	0.1333333333	0.1333333333	0.1333333333	0.1333333333
Average	rouge1	0.1954504148	0.1909697477	0.1857556842	0.1952826818	0.190500283	0.1852893748	0.1934116682	0.1894394499	0.1832051844
	rouge2	0.0555025681	0.05351556902	0.04869776953	0.06079846529	0.05907387915	0.05428943384	0.05538542139	0.05476950844	0.04992303512
	rougeL	0.1609775643	0.1587157487	0.1526683519	0.1647890435	0.1617680145	0.1548635579	0.1608902303	0.1587508421	0.1517084958
	rougeLsum	0.1609775643	0.1587157487	0.1526683519	0.1647890435	0.1617680145	0.1548635579	0.1608902303	0.1587508421	0.1517084958
Max	rouge1	0.5454545455	0.5454545455	0.5454545455	0.5185185185	0.5185185185	0.5185185185	0.5185185185	0.5185185185	0.5185185185
	rouge2	0.2962962963	0.2962962963	0.2962962963	0.2962962963	0.2962962963	0.2962962963	0.2962962963	0.2962962963	0.2962962963
	rougeL	0.5454545455	0.5454545455	0.5454545455	0.4324324324	0.4324324324	0.4324324324	0.4324324324	0.4324324324	0.3846153846
	rougeLsum	0.5454545455	0.5454545455	0.5454545455	0.4324324324	0.4324324324	0.4324324324	0.4324324324	0.4324324324	0.3846153846
Min	rouge1	0	0	0	0	0	0	0	0	0
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0	0	0	0	0	0	0	0	0
	rougeLsum	0	0	0	0	0	0	0	0	0

Pegasus Token Range 0



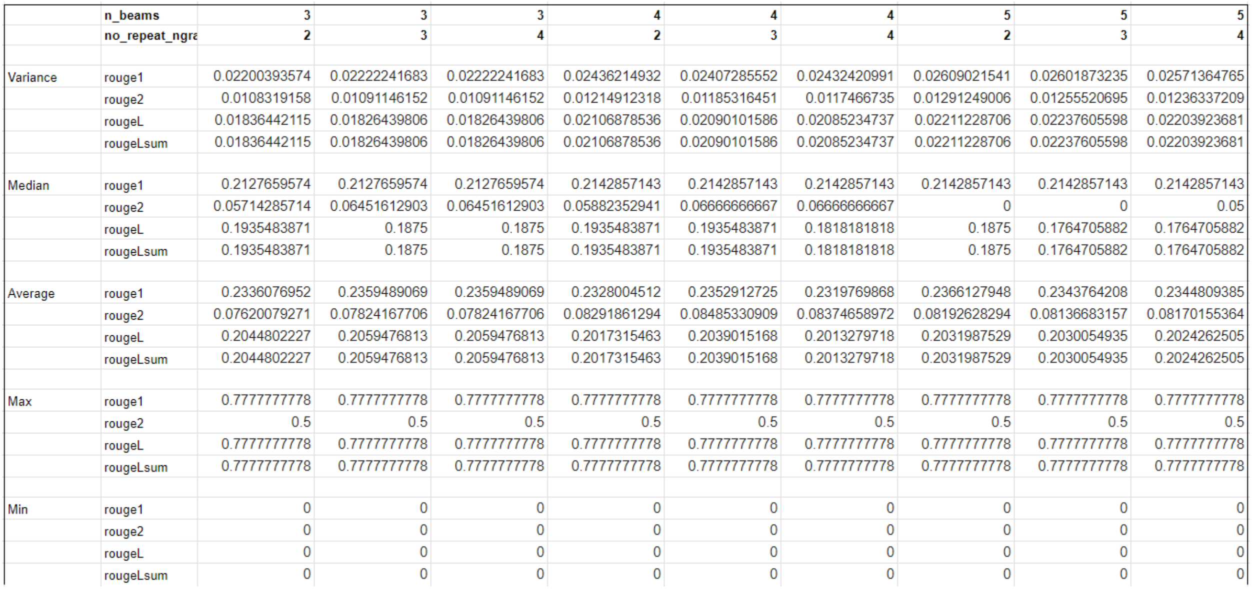
	n_beams	3	3	3	4	4	4	5	5	5
	no_repeat_ngra	2	3	4	2	3	4	2	3	4
Variance	rouge1	0.02521214452	0.02528740908	0.02528740908	0.02953950293	0.02993018562	0.03014413717	0.03068511252	0.03093323335	0.03143240093
	rouge2	0.01343175664	0.01344305822	0.01344305822	0.01458607957	0.01458607957	0.01458607957	0.01502592627	0.01502592627	0.01502592627
	rougeL	0.02209788211	0.02246902629	0.02246902629	0.02462105837	0.02511846495	0.02510079425	0.02592871653	0.0259379353	0.02637008766
	rougeLsum	0.02209788211	0.02246902629	0.02246902629	0.02462105837	0.02511846495	0.02510079425	0.02592871653	0.0259379353	0.02637008766
Median	rouge1	0.2285714286	0.2285714286	0.2285714286	0.2222222222	0.2222222222	0.2222222222	0.2222222222	0.2222222222	0.2222222222
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0.2105263158	0.2105263158	0.2105263158	0.2068965517	0.2068965517	0.2068965517	0.2105263158	0.2105263158	0.2105263158
	rougeLsum	0.2105263158	0.2105263158	0.2105263158	0.2068965517	0.2068965517	0.2068965517	0.2105263158	0.2105263158	0.2105263158
Average	rouge1	0.2366662833	0.2370879525	0.2370879525	0.2301962704	0.2323989172	0.2315735204	0.2392416055	0.2378545087	0.2365211754
	rouge2	0.07658062648	0.07665659989	0.07665659989	0.08265052207	0.08265052207	0.08265052207	0.08104817681	0.08104817681	0.08104817681
	rougeL	0.2133429821	0.2146513508	0.2146513508	0.2059699905	0.2081726373	0.2082361293	0.2127775633	0.2127237998	0.2113904665
	rougeLsum	0.2133429821	0.2146513508	0.2146513508	0.2059699905	0.2081726373	0.2082361293	0.2127775633	0.2127237998	0.2113904665
Max	rouge1	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	rouge2	0.6153846154	0.6153846154	0.6153846154	0.6153846154	0.6153846154	0.6153846154	0.6153846154	0.6153846154	0.6153846154
	rougeL	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	rougeLsum	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Min	rouge1	0	0	0	0	0	0	0	0	0
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0	0	0	0	0	0	0	0	0
	rougeLsum	0	0	0	0	0	0	0	0	0

Pegasus Token Range 5



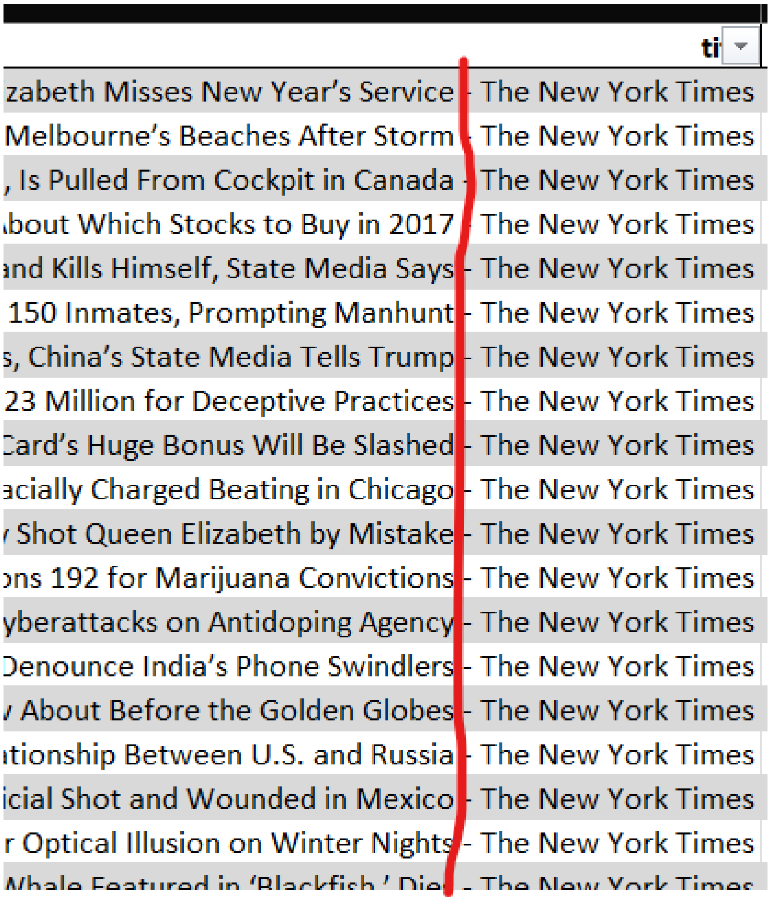
	n_beams	3	3	3	4	4	4	5	5	5
	no_repeat_ngra	2	3	4	2	3	4	2	3	4
Variance	rouge1	0.02266844686	0.02265022691	0.02265022691	0.02480333661	0.02443093342	0.02420189881	0.0258856294	0.02579956761	0.02554293451
	rouge2	0.01165984268	0.01173337619	0.01173337619	0.01311780825	0.01293723852	0.0128023311	0.01365071106	0.01341628897	0.01332378705
	rougeL	0.01899736865	0.01858930906	0.01858930906	0.02192289872	0.02139294063	0.02135264042	0.02286663397	0.02289853209	0.02263344627
	rougeLsum	0.01899736865	0.01858930906	0.01858930906	0.02192289872	0.02139294063	0.02135264042	0.02286663397	0.02289853209	0.02263344627
Median	rouge1	0.2222222222	0.2142857143	0.2142857143	0.2222222222	0.2222222222	0.2222222222	0.2222222222	0.2222222222	0.2222222222
	rouge2	0	0.0625	0.0625	0	0.06896551724	0.06896551724	0	0.05128205128	0.0625
	rougeL	0.1935483871	0.1904761905	0.1904761905	0.1904761905	0.1875	0.1875	0.1951219512	0.1935483871	0.1951219512
	rougeLsum	0.1935483871	0.1904761905	0.1904761905	0.1904761905	0.1875	0.1875	0.1951219512	0.1935483871	0.1951219512
Average	rouge1	0.2368504492	0.2397356332	0.2397356332	0.2429347388	0.2459241584	0.2437756539	0.2461054791	0.2449490283	0.2462307444
	rouge2	0.0806667684	0.08265442272	0.08265442272	0.08607000213	0.0878727428	0.08704515659	0.08631131371	0.08633448449	0.08716781782
	rougeL	0.211799607	0.2138579936	0.2138579936	0.2089576145	0.2118077845	0.2120833786	0.2173988424	0.2177950611	0.2200291581
	rougeLsum	0.211799607	0.2138579936	0.2138579936	0.2089576145	0.2118077845	0.2120833786	0.2173988424	0.2177950611	0.2200291581
Max	rouge1	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778
	rouge2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	rougeL	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778
	rougeLsum	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778
Min	rouge1	0	0	0	0	0	0	0	0	0
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0	0	0	0	0	0	0	0	0
	rougeLsum	0	0	0	0	0	0	0	0	0

Pegasus Token Range 10



	n_beams	3	3	3	4	4	4	5	5	5
	no_repeat_ngra	2	3	4	2	3	4	2	3	4
Variance	rouge1	0.02200393574	0.02222241683	0.02222241683	0.02436214932	0.02407285552	0.02432420991	0.02609021541	0.02601873235	0.02571364765
	rouge2	0.0108319158	0.01091146152	0.01091146152	0.01214912318	0.01185316451	0.0117466735	0.01291249006	0.01255520695	0.01236337209
	rougeL	0.01836442115	0.01826439806	0.01826439806	0.02106878536	0.02090101586	0.02085234737	0.02211228706	0.02237605598	0.02203923681
	rougeLsum	0.01836442115	0.01826439806	0.01826439806	0.02106878536	0.02090101586	0.02085234737	0.02211228706	0.02237605598	0.02203923681
Median	rouge1	0.2127659574	0.2127659574	0.2127659574	0.2142857143	0.2142857143	0.2142857143	0.2142857143	0.2142857143	0.2142857143
	rouge2	0.05714285714	0.06451612903	0.06451612903	0.05882352941	0.06666666667	0.06666666667	0	0	0.05
	rougeL	0.1935483871	0.1875	0.1875	0.1935483871	0.1935483871	0.1818181818	0.1875	0.1764705882	0.1764705882
	rougeLsum	0.1935483871	0.1875	0.1875	0.1935483871	0.1935483871	0.1818181818	0.1875	0.1764705882	0.1764705882
Average	rouge1	0.2336076952	0.2359489069	0.2359489069	0.2328004512	0.2352912725	0.2319769868	0.2366127948	0.2343764208	0.2344809385
	rouge2	0.07620079271	0.07824167706	0.07824167706	0.08291861294	0.08485330909	0.08374658972	0.08192628294	0.08136683157	0.08170155364
	rougeL	0.2044802227	0.2059476813	0.2059476813	0.2017315463	0.2039015168	0.2013279718	0.2031987529	0.2030054935	0.2024262505
	rougeLsum	0.2044802227	0.2059476813	0.2059476813	0.2017315463	0.2039015168	0.2013279718	0.2031987529	0.2030054935	0.2024262505
Max	rouge1	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778
	rouge2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	rougeL	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778
	rougeLsum	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778	0.7777777778
Min	rouge1	0	0	0	0	0	0	0	0	0
	rouge2	0	0	0	0	0	0	0	0	0
	rougeL	0	0	0	0	0	0	0	0	0
	rougeLsum	0	0	0	0	0	0	0	0	0

Data Cleaning New York Times Example



ti ▾
zabeth Misses New Year's Service - The New York Times
Melbourne's Beaches After Storm - The New York Times
, Is Pulled From Cockpit in Canada - The New York Times
ıbout Which Stocks to Buy in 2017 - The New York Times
ınd Kills Himself, State Media Says - The New York Times
150 Inmates, Prompting Manhunt - The New York Times
s, China's State Media Tells Trump - The New York Times
23 Million for Deceptive Practices - The New York Times
Card's Huge Bonus Will Be Slashed - The New York Times
acially Charged Beating in Chicago - The New York Times
/ Shot Queen Elizabeth by Mistake - The New York Times
ons 192 for Marijuana Convictions - The New York Times
yberattacks on Antidoping Agency - The New York Times
Denounce India's Phone Swindlers - The New York Times
/ About Before the Golden Globes - The New York Times
ıtionship Between U.S. and Russia - The New York Times
icial Shot and Wounded in Mexico - The New York Times
r Optical Illusion on Winter Nights - The New York Times
Whale Featured in 'Blackfish,' Dies - The New York Times