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](https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/?tabId=tab-b39b851c-e417-48ef-9b10-93ee21a0030e)]. 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](https://www.statista.com/statistics/785919/worldwide-number-of-digital-newspaper-subscribers/)].

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](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.)]. A common explanation for this is that the media is biased [[10](https://journals.sagepub.com/doi/pdf/10.1177/1464884918808134)]. 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](http://arxiv.org/abs/2109.00024)] 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](https://doi.org/10.1038/s41598-022-13690-4)] 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](https://doi.org/10.1155/2021/6679661)] 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](https://arxiv.org/pdf/2007.05044.pdf)] 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](https://www.kaggle.com/datasets/snapcrack/all-the-news)] 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)

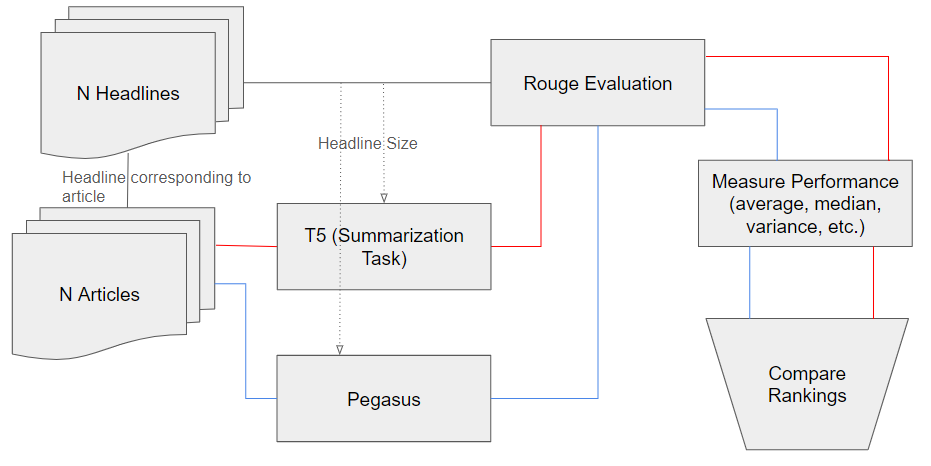
## Headline Evaluation

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.

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](https://doi.org/10.1038/s41598-022-13690-4)].

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](https://www.researchgate.net/publication/324455115_A_Stylistic_Analysis_of_Selected_Newspaper's_Stories)] 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.

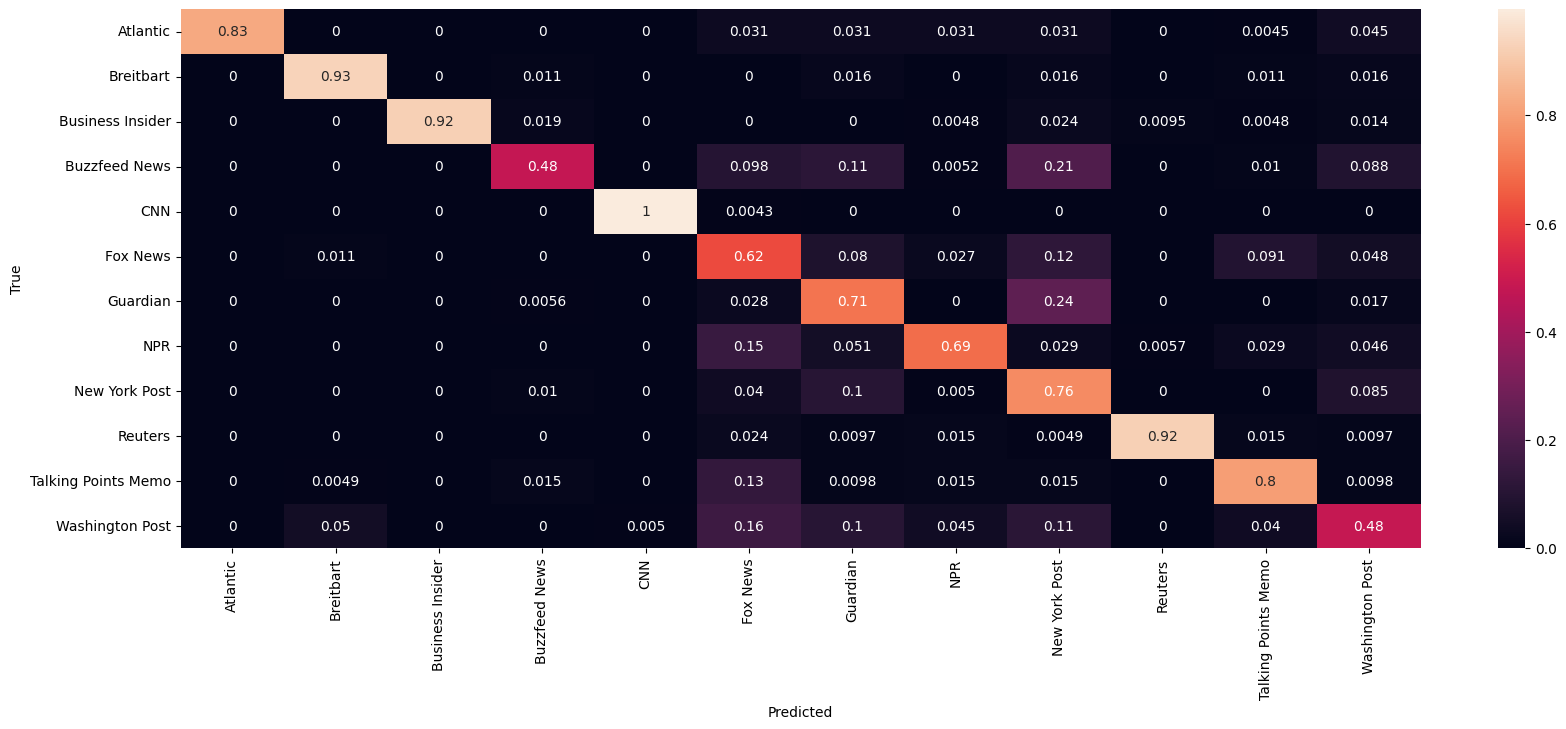


Full tabulated results are documented in the appendix, including scores for Rouge1, Rouge2, and variance scores.

## 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.



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/](http://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**

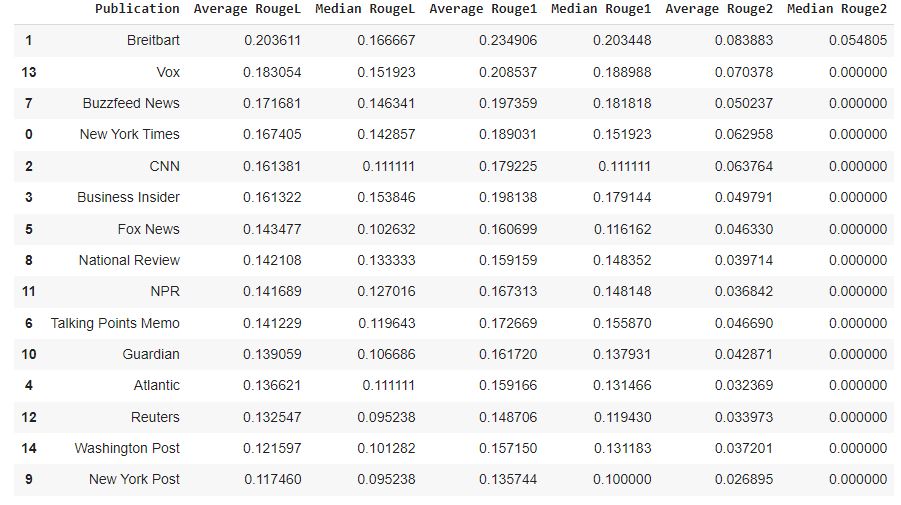
# 

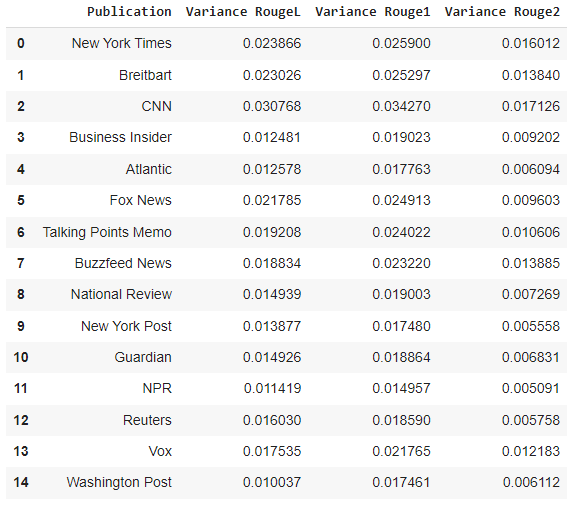
Model 1: (12 news outlets with 1000 articles per outlet) - Architecture and Parameters

# 

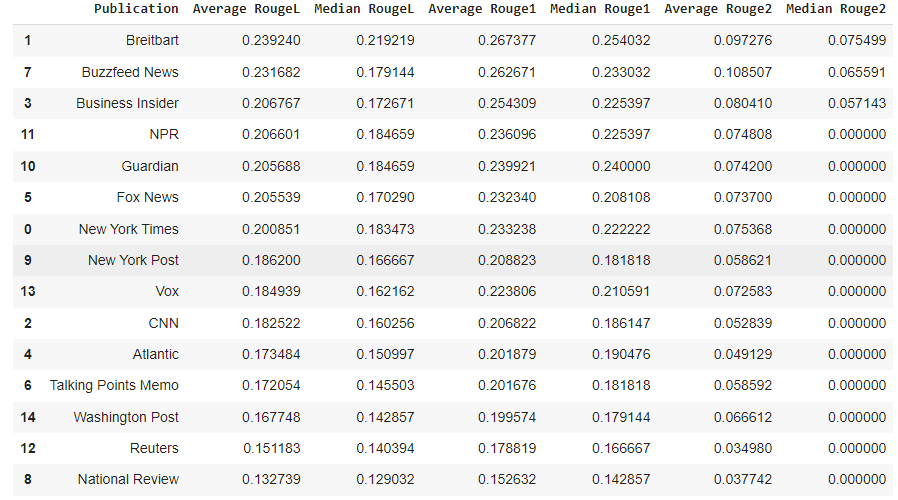
Model 2: 6 News Outlets Classification with 1000 news articles per outlet (identical model architecture to Model 1)

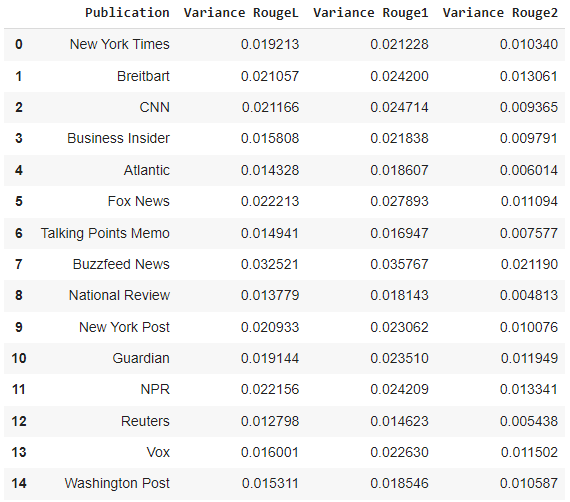
**T5 Full Results**





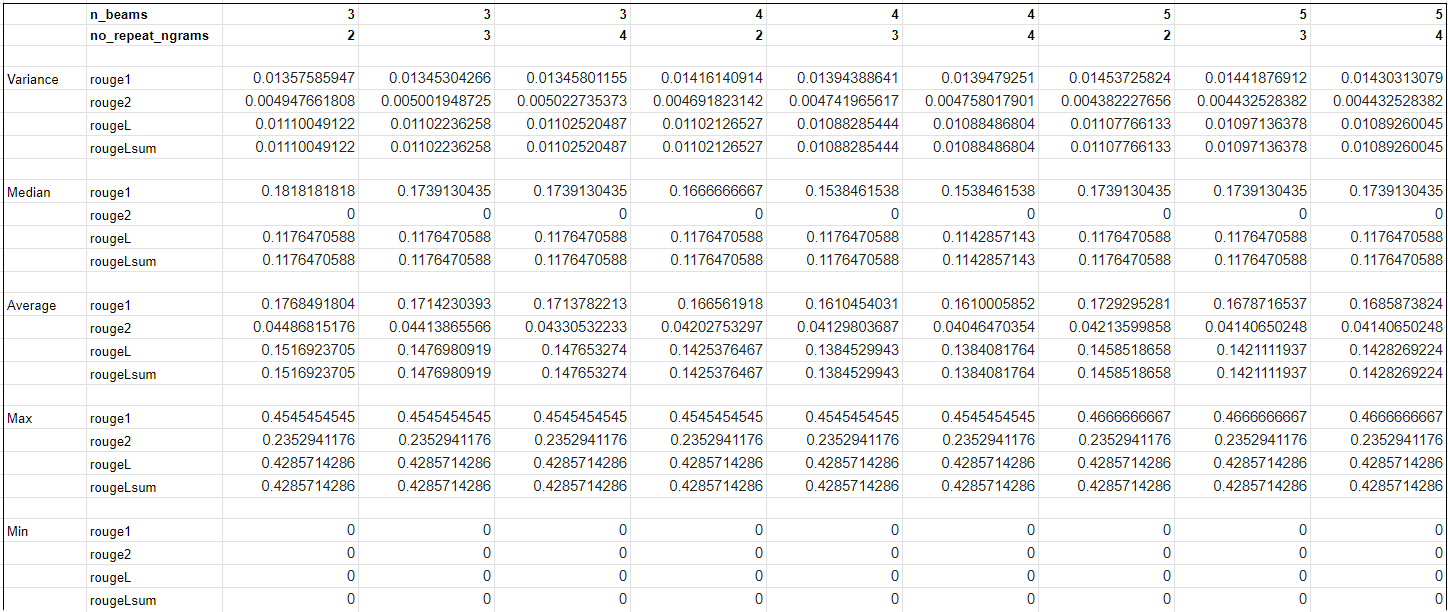
**Pegasus Full Results**



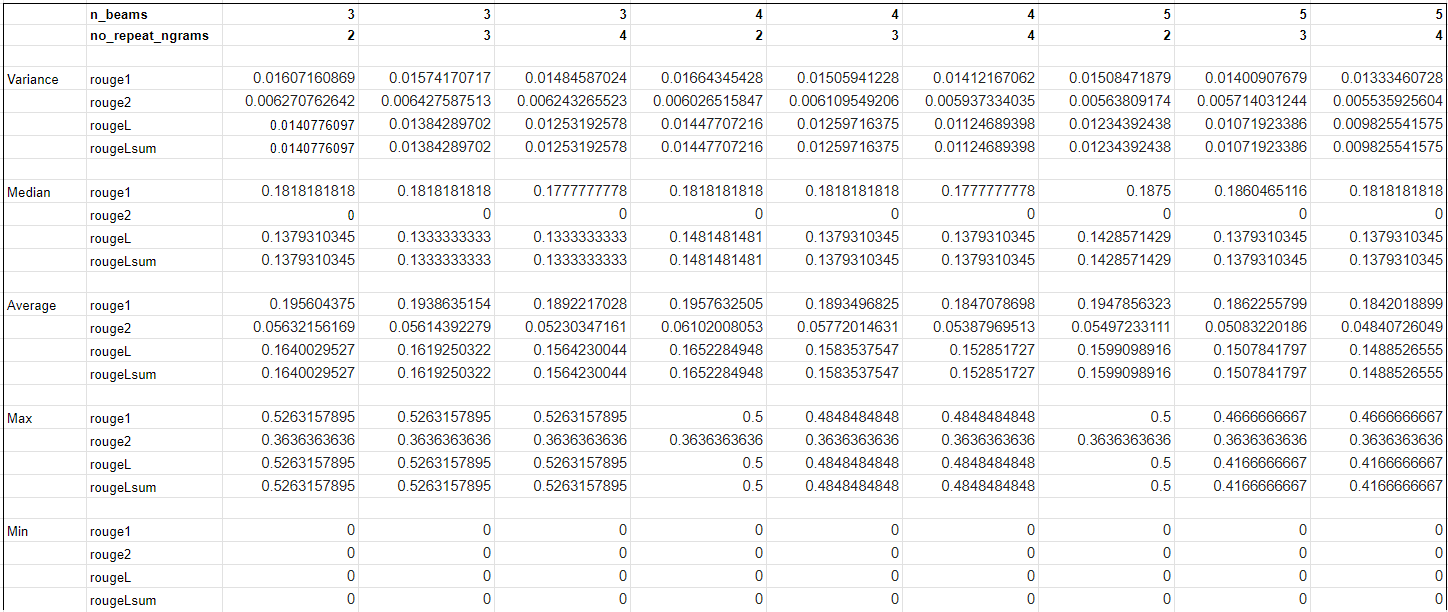


**Parameter Determination Tables**

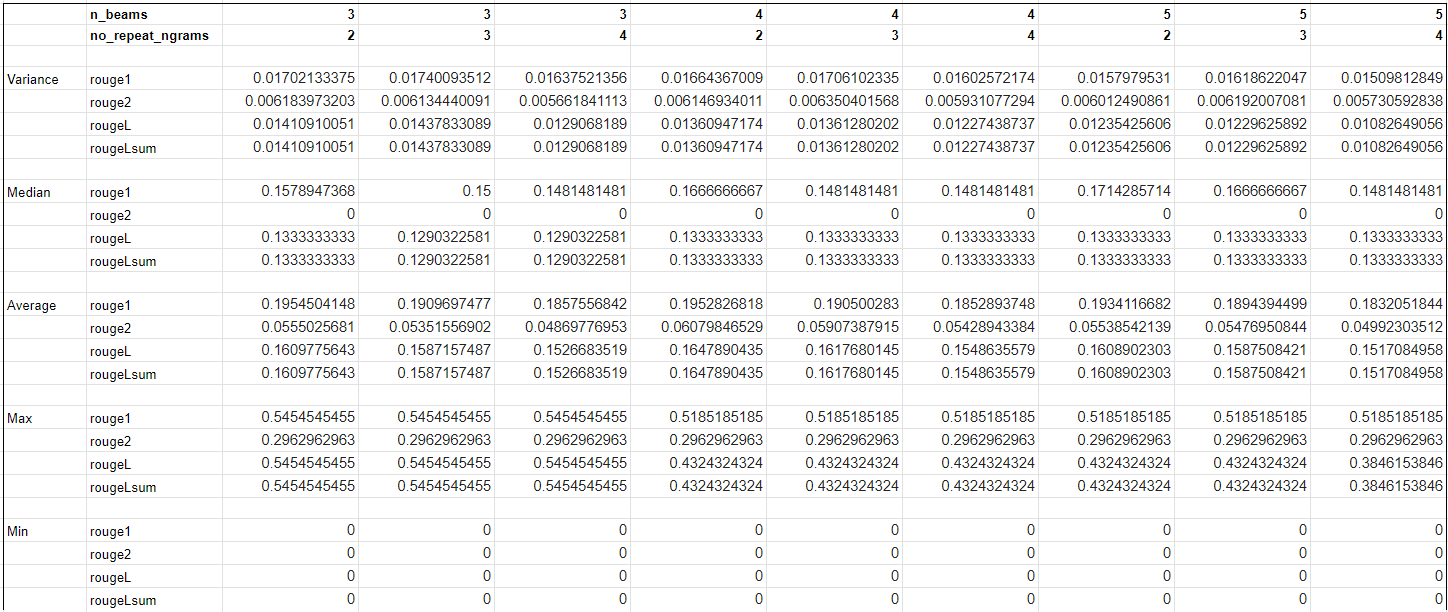
T5 Token Range 0

****

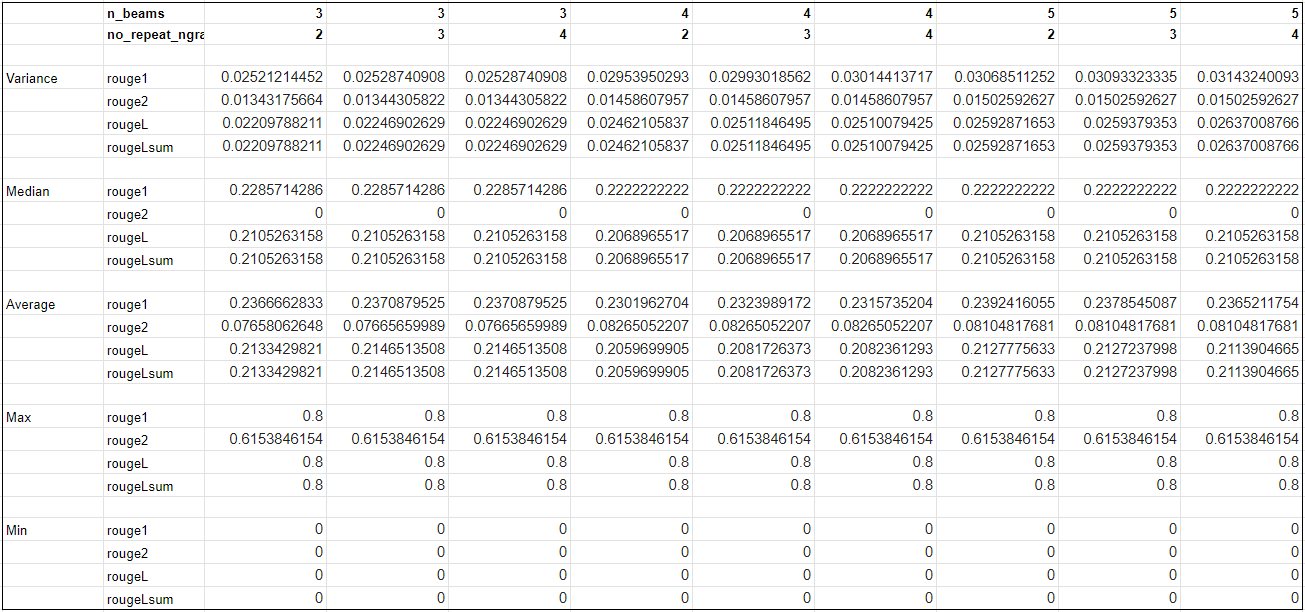
T5 Token Range 5



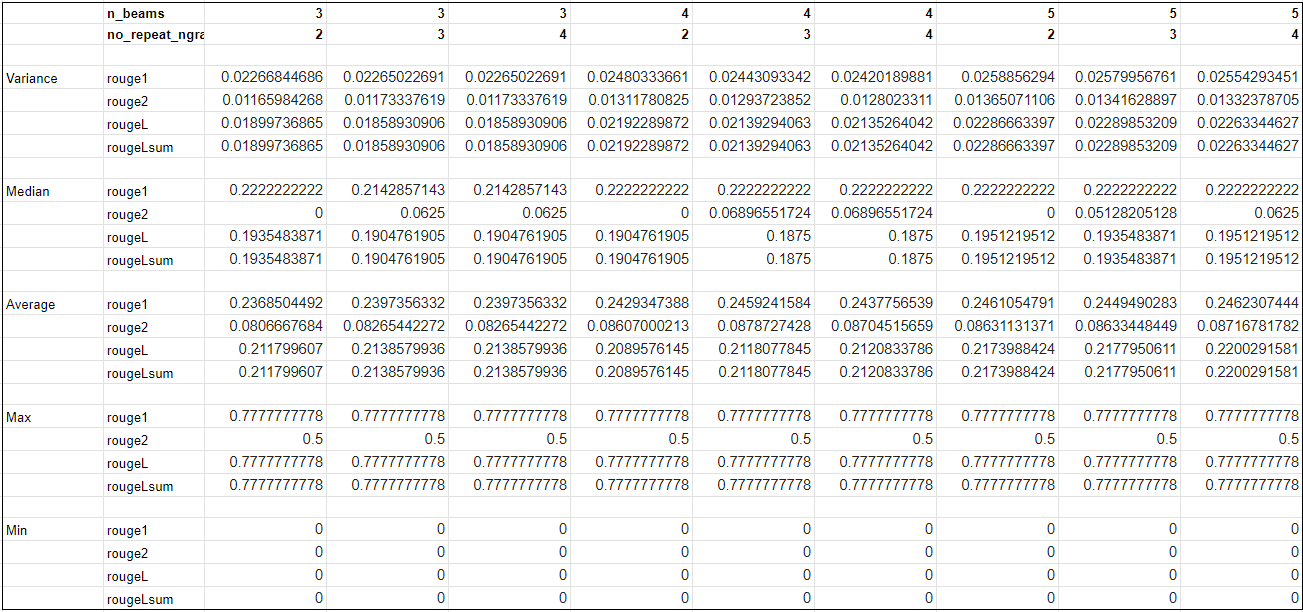
T5 Token Range 10



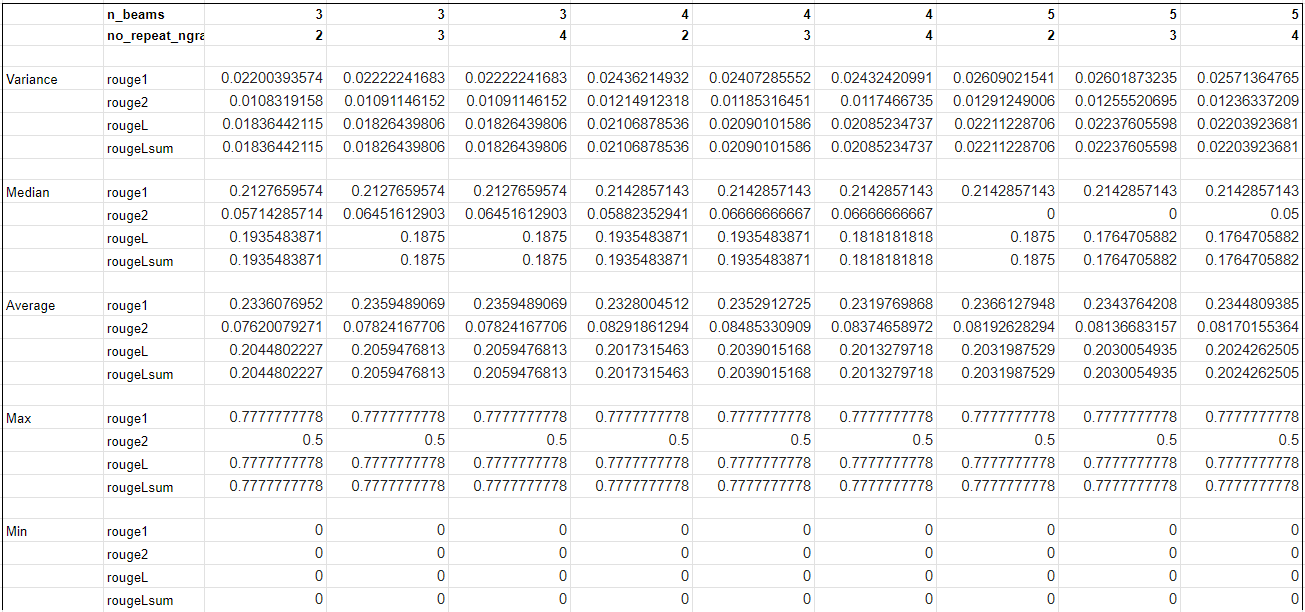
Pegasus Token Range 0



Pegasus Token Range 5



Pegasus Token Range 10



**Data Cleaning New York Times Example**

