What is the essence of a news article? An application of generative text summariazation models to news articles and an analysis of the biases they carry over.

MOWAFAK ALLAHAM, Northwestern University, USA - MowafakAllaham2021@u.northwestern.edu
JULIA BARNETT, Northwestern University, USA - JuliaBarnett@u.northwestern.edu

Generative models have made great strides in the area of text summarization. These models strive to extract the essence of bodies of text to produce short and comprehensive summaries. As the demand for an informed society rises concurrently with a decreasing attention span for news media, the importance of accurate and comprehensive summaries of news articles becomes vital. In this work we plan on developing a news summarizing tool that can be utilized by journalists and the public alike. In addition, we aim to highlight potential concerns of biases, false claims, or misleading information that can be carried over into these summaries by analyzing the outputs generated by GPT-3 to articles published on low credible news domains on the topic of climate change.

CCS Concepts: • **Concept** → **Concept**; *Concept*; Concept; Concept.

Additional Key Words and Phrases: generative deep models, text summarization, political journalism

ACM Reference Format:
Mowafak Allaham and Julia Barnett. 2022. What is the essence of a news article? An application of generative text summariazation models to news articles and an analysis of the biases they carry over.. In *CITY 'YY: CONFERENCE TITLE, Month DD–DD, YYYY, City, St.* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The average amount of time a user spends on a news sites has been decreasing yearly–culminating at a staggering estimation of the average user spending less than 2 minutes on news sites in 2020 [4]. Half of U.S adults get their news media from social media [8]. Among these adults, around 40% of Facebook users and 50% of Twitter users regularly get the news on these social media platforms. Accordingly, it is essential to offer the public readable and credible news in order to share an unbiased lens to domestic policies, and international affairs.

One challenge to news curators is how to present news information to the readers in a way that keeps them engaged with relevant news. One approach is to offer the readers a daily briefing, similar to the morning newsletters by the New York Times, that provides bullet-point summary of daily news events and articles published online. One challenge to achieve this objective is the laborious effort that journalists put forth to sift through relevant news articles and summarize ones that are relevant to the readers.

The objective of our research is to develop a proof-of-concept tool using GPT-3, a deep generative model, that can help journalists summarize news articles from a range of news sources and across news categories. We also plan on extending our contribution by highlighting potential risks that may occur from such task by assessing whether GPT-3 embed any political biases, false, or misleading information in summaries of articles from low credible news sources such as right-wing biased domains.

We hope that our contribution yields summaries of articles that are unbiased but also informative of any potential false or misleading information that could protect journalists and society from mistakenly being either spreaders or victims of misinformation.

Conference acronym 'XX, Month DD–DD, YYYY, City, ST

2 RELATED LITERATURE

For the purposes of this project, we will have a limited literature review. If we extend this to a conference paper (as we hope to), we will dedicate this section to generative text models and their biases, classification of political journalism, and biases and conspiracy theories present in political journalism.

Possible (and likely) options:

- GPT-3 - [3]
- BERT - [5]
- Text summarization using unsupervised deep learning - [12]
- Text Summarization Techniques: A Brief Survey - [2]
- A Survey of Text Summarization Techniques - [9]
- NLP based Machine Learning Approaches for Text Summarization - [1]
- Abstractive text summarization using LSTM-CNN based deep learning - [11]
- Get To The Point: Summarization with Pointer-Generator Networks [10]

3 DATA

To fine-tune GPT-3 for article summarization, we will be using data compiled by Nallapati et al. [7] containing a variety of news articles from CNN and The Daily Mail and the short summaries corresponding to each article. The news articles and the short summaries in this dataset have an average of 781 and 56 tokens, respectively. This dataset offers a rich stream of news content as it contains 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs which we can be utilize to fine-tune and validate text summaries generated by GPT-3. In case we need additional data, we will scrape news articles using news-please–a generic news web crawler built by Hamborg et. al. [6].

To assess the quality of GPT-3 article summaries on low credible sources, we first employed the analytics platform NewsWhip and a web scraper, we retrieved the full-text and engagement data for 15,178 articles published on Facebook and Twitter in 2021 (with a combined 20,089,856 likes, shares, and comments on Facebook

and Twitter). All articles were confirmed to be discussing climate change based on a set of keywords (n-grams) found in the headline or lead paragraph of each article. All articles were categorized as right-wing or low credible domains based on Media Bias Fact Check (MBFC) and NewsGuard credibility score.

4 METHODOLOGY

We will exclusively be using Python for data collection, fine-tuning, and generating article-summaries. The distribution of tasks and responsibilities are outlined in Section 6 (Project Plan). The expected time to prepare and clean the data is about one week and coding is about 2-3 weeks as we will need to do some experimentation with GPT-3 to optimize its hyper parameters to generate readable and cohesive summaries.

In regards to the software packages, we expect to rely on OpenAI python package, Tensorflow, Pandas, and Newspaper3k. However, we expect to do research on what additional packages we may need as we progress through the project.

5 PROJECT PLAN

By our meeting in mid November, we will have:

(1) identified which platform to use to host the data on and fine-tune GPT-3 model.
(2) converted all of our data in a useable format: the metadata associated will be stored properly, and the corpora will be stored as objects ready to be summarized in Python scripts.
(3) completed a first iteration of the text summarization model (seeking any feedback on our approach to this implementation and we will adjust accordingly).

As of December 6, 2022, we will have:

(1) successfully trained a model summarizing news articles into short summaries
(2) applied this model to a new corpus of news articles with varying levels of credibility.
(3) assessed GPT-3 for the presence/or absence of false or misleading claims in news summarizes to low credible sources in the context of climate change.
(4) qualitatively analyzed the differences amongst the generated summaries with respect to political orientations and credibility of media sources.
(5) written up a fully fleshed out data, methodology, and results sections of this proposal in web-accessible format for presentation purposes.

Distribution of labor:

- We will be working jointly on most aspects of this project, but will be primary owners of different sections:
- Julia will be primarily responsible for the data collection, cleaning, and pre-processing.
- Mowafak will be primarily responsible for the GPT-3 component of the generative summaries (in addition to summaries to low credible source in the context of climate change).
- Julia will be responsible for the the qualitative analysis of the generated summaries and assessment of the presence or absence of politically biased, false, or misleading information on climate change.
- Julia will be responsible for the maintenance of the github page and presentation of the project results on it.

6 FUTURE EXTENSION

A future direction we plan to take this in (beyond the scope of this project) is to analyze whether political biases are emphasized or minimized when summarizing these articles. We will do so by labelling 1000 of the summaries manually as left bias, left-center, unbiased, right-center, or right bias as defined by media fact checker (a prominent academic political tool) and then using this to train a classification model. Once we have predictions for the entirety of the dataset we will analyze the differences between the original classification of the news outlet and the classification of the summary: does the summary allign with the original classification or does it approach the extremes?

REFERENCES

[1] Surabhi Adhikari et al. 2020. Nlp based machine learning approaches for text summarization. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 535–538.
[2] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268* (2017).
[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Social Media and News Fact Sheet. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[4] US Comscore Media Metrix MultiPlatform. 2021. October-December 2014-2020.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[6] Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A Generic News Crawler and Extractor. In *Proceedings of the 15th International Symposium of Information Science* (Berlin). 218–223. https://doi.org/10.5281/zenodo.4120316
[7] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems* 28 (2015).
[8] Matsa K. E. Liedke J. 2022. Social Media and News Fact Sheet. Pew Research.
[9] Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining text data*. Springer, 43–76.
[10] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1073–1083. https://doi.org/10.18653/v1/P17-1099
[11] Shengli Song, Haitao Huang, and Tongxiao Ruan. 2019. Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications* 78, 1 (2019), 857–875.
[12] Mahmood Yousefi-Azar and Len Hamey. 2017. Text summarization using unsupervised deep learning. *Expert Systems with Applications* 68 (2017), 93–105.

ACKNOWLEDGMENTS

Acknowledgements go here.