Final Project Abstract by Nicholas Petrocelli, Johnny Boustany, David Lauerman, Max Dekle

Hypothesis:

Remote work has been a topic of popular discussion in recent years, especially gaining traction during the COVID-19 pandemic when many businesses began to accommodate work-from-home policies in order to comply with social distancing guidelines. However, some companies have been advocating for a return to the office, despite pushback from their employees. Therefore, we wanted to analyze a diverse set of opinions to see if this was actually true – is remote work really on the decline, or is it here to stay? Given a diverse set of voices and datapoints, we set out to test the hypothesis that remote work sentiment has decreased over the past few years; in other words, remote work is not as popular as it used to be.

Data:

We combined 3 unique sources for our dataset: The Guardian, The New York Times, and Reddit. 307 and 295 articles were obtained from the NYT and the Guardian, respectively, by querying their APIs for articles about 'remote work'. 548 Reddit posts were obtained by scraping the "experienceddevs" and "cscareerquestions" subreddits for posts containing the terms "remote work", "work from home", or "WFH". A specific challenge that we faced was ensuring that most of the articles found during scraping were actually relevant to remote work, especially from NYT/ The Guardian; to alleviate this, we filtered such that only posts after Jan 1, 2020 were included, and filtered out certain confounding keywords during scraping.

Our identifying attributes for these entries include the ID, URL, DATE, PLAIN_TEXT and SOURCE. The ID is a unique ID that distinguishes the articles from each other, retrieved from the source (the Guardian, Reddit and the New York Times). The PLAIN_TEXT is a cleaned version of the article contents, having removed non-alphanumeric characters, inconsistent spacing, and ensuring all text is unicode.

Findings:

We utilized the supervised pre-trained model known as RoBERTa, commonly used to evaluate Twitter sentiment, in order to generate sentiments for our data. Since the model can only evaluate strings up to 514 characters long, we utilized NLTK's sentence tokenizer to obtain a sentiment for each sentence, the collection of which was averaged to obtain a single sentiment rating per article. The model was validated with a 0.71 overall testing accuracy, with the following metrics:

	Negative	Neutral	Positive	Average
Precision	0.89	0.69	0.86	0.81
Recall	0.34	0.97	0.28	0.53
F1	0.50	0.81	0.42	0.58

The performance of the RoBERTa model on our dataset was subsequently compared to its performance on the Twitter dataset. Our dataset had a lower average recall potentially explained by a biased tendency of rating news sentences neutrally. This suggests that the model's outputs on news articles can be most trusted if it identifies significant positive or negative sentiment.

We then conducted linear regression analysis on the sentiment ratings to determine if the 2022-2023 articles are statistically different from the 2020-2021 articles. Our mean r2 value was very low at -0.004, as is our mean MSE value at 0.05.

A low MSE suggests that the predicted values are very close to the actual values. In other words, the model predicts the outcome variable with reasonable accuracy. Despite this, the low R-squared value shows that the model is unfortunately not explaining much of the variance in the outcome variable. Therefore, we can conclude that our results are statistically insignificant by virtue of our low r-squared value, and fail to prove the hypothesis that remote work has received an increasingly negative sentiment in recent years.

We additionally analyzed the following three hypotheses to gain insight into the integrity of our dataset:

- ☐ **Is the sentiment significantly different between news sources and Reddit?**
  The t-statistic was -1.334 and the p-value 0.182, which indicates that we fail to reject the null hypothesis. This means that there is no statistically significant difference in sentiments between news sources and Reddit, as the p value was not less than 0.05.
- ☐ **Is the sentiment significantly different between the Guardian and NYT?**
  The t-statistic was -0.186 and the p-value 0.852, which also indicates that we fail to reject the null hypothesis. This means that there is no statistically significant difference in sentiments between The Guardian and NYT sentiments.
- ☐ **Is the sentiment significantly different between the two subreddits?**
  The t-statistic was -1.341 and the p-value 0.180, which again indicates that we fail to reject the null hypothesis. Thus, once again there is no significant difference in sentiments between the two subreddits.

However, considering the three p-values as a set, we can say that there is the least difference in sentiments between the Guardian and NYT compared to the other pairings, which makes sense intuitively as both are news sites and commonly publish similar articles.

Socio-Historical Context and Impact Report:

Socio-Historical Context

This project is inseparable from the context of the COVID-19 pandemic and its economic effects. The pandemic has caused a number of substantial shifts in the structure of tech jobs, with one-third of all jobs in the US moving to full-remote in 2020 according to the National Council on Compensation Insurance [1]. It is this phenomenon that made collecting our data possible, as without it many fewer news articles about WFH would have been written and far fewer workers would have posted to Reddit about their experiences. However, there are additional historical factors that influenced our data, the most significant being the general decline in mental health during the pandemic. According to the WHO, the COVID-19 pandemic caused a 25% increase in the level of anxiety and depression worldwide [2]. This would make the overall sentiment on social media sites such as Reddit more negative and would likely influence the sentiment of journalists as well, even if only on a subconscious level.

Additionally, political tensions have risen in recent years, evidenced with Joe Biden's election win, Donald Trump's ongoing criminal investigation, and Boris Johnson resigning from office. These events could potentially indicate a loss in confidence in world leaders, which is exacerbated by the unpredictable

actions of tech moguls Elon Musk and Mark Zuckerberg. Further, the Ukrainian invasion, Turkish earthquake, and Iranian protests may mean that workers are more focused on spending time with their families and communities and prefer work from home setups. The ever-changing oil prices, interest and inflation rates are also a factor in impacting the cost of commutes as well as monetary spending.

There are three primary groups of stakeholders for our project. The first is the general rank-and-file of tech and office workers, many of whom are currently working remotely. If our research aligns well with the true opinions of these workers, they would stand to benefit from their views being amplified and more widely known. Conversely, however, they could be harmed if our project misrepresented their views and contributed to a consensus for policies that they do not want. The second group of stakeholders are tech businesses and other holders of leases on office space. The rise in WFH policies has caused office leases to essentially become dead weight on businesses; in fact, according to the Wall Street Journal 2022 was a record year for office lease expirations as many businesses cut costs, putting landlords and banks at risk [3]. These entities clearly stand to benefit from our work showing a general negativity bias around WFH, as this could be used as justification to move workers back into the office. As soon-to-be tech workers ourselves, it is particularly important that we do not accidentally harm tech workers with our conclusions. A third group of stakeholders are those impacted by the daily commute of workers; bus drivers, train operators, street cleaners, even vendors who cater to the breakfast and coffee needs of the morning rush. More individuals traveling to work creates more jobs and opportunities for these individuals, as well as raising the demand for oil, new cars, and bicycles. In summary, some stakeholders benefit from work-from-home policies, while others benefit from in-person trends, but generally stakeholders do not benefit from both.

Much work has already been done in the polling sphere to track opinions of WFH policies, both from the perspective of rank-and-file workers and business owners or managers. As with all opinion polling, the results are not set in stone; however, the general consensus seems to be that WFH policies are much more popular with workers than owners, albeit with some nuances on both sides. For instance, the McKinsey Institute reported on their 2022 “American Opportunity Survey” that 87% of American workers would work remotely at least part-time if given the opportunity [4]; meanwhile, in 2021 the BBC reported that “In the US, a whopping 72% of managers currently supervising remote workers would prefer all their subordinates to be in the office…” [5]. This research has generally given the impression among tech workers that WFH policies are in danger of being revoked due to being disliked by management, and it is this impression that led us to pursue this topic for our analysis. However, it is important that we present our data in an unbiased manner, and not let our own biases influence our data analysis, presentation, and interpretation methodologies.

Ethical Considerations

One of the primary technical challenges of this project was collecting sufficient data, and it is this data collection that provides the greatest ethical challenges as well. For one, our data sources themselves contain societal biases; Reddit especially so, as Reddit itself has a skewed user base, as 47% of its users are from the U.S. and about twice as many of its users are male than female [6]. Our selection of news sources also demonstrates a bias on our part, as we focused specifically on news sources we considered “mainstream” or “legitimate”, which are inherently subjective terms. It is possible that selecting different

news sources would have changed the outcome of our analysis - for example, more tech-focused news sites might have had more positive sentiment towards WFH policies.

Another societal issue that we should address is the reasoning behind individuals voicing their opinions about remote work online. By nature, people are more likely to speak on things they are highly passionate about, whether that be in strong support or in strong opposition to remote work. Therefore, our dataset might contain a lot of sentiment strongly to one side or another, and not many hovering around the neutral range. To combat this, one solution might be to collect data of our own, whether that be polling college students on campus or asking workers in our own communities what their stance is on the issue.

Another ethical concern is whether we used the data in a manner consented to by the authors of the content. In a strictly legal sense we are, as all of the text we analyzed was publicly available, public-facing content. On the other hand, none of the Reddit users whose posts we collected explicitly agreed to be a part of a data analysis project. To try to thread this needle, we deliberately selected Reddit as our social media source, as it actively discourages users from posting personal information. This distinguishes it from other sites such as Facebook and Twitter which encourage users to use their real names and to put other identifying information on their profiles. While we think this is sufficient to mitigate the potential privacy issues with this project, any larger-scale effort using our methods would benefit from explicitly trying to scrub personal information from the collected data.

The final ethical quandary we must discuss is potential misinterpretations of our work. One potential misconception would be to use the fact that we did not reject the null hypothesis when comparing the average sentiments of news sources and reddit forums as proof that social media and the news are definitively aligned on the issue. Another similar misconception would be to say that the negative sentiment bias we discovered is proof that everyone is generally negative about the idea of working from home. The common thread between these is making strong claims of "proof" when in fact our analysis was largely inconclusive, and we acknowledged that more data collection and study is needed. The best way for us to prevent these misconceptions is to be upfront and clear about both the limitations of our work and all of the potential confounding factors, most saliently the general decline in mental health - and therefore sentiment - caused by the events of 2020.

Works Cited:

1. Coate, Patrick. "Remote Work Before, During, and After the Pandemic." *National Council on Compensation Insurance (NCCI)*, National Council on Compensation Insurance (NCCI), 25 Jan. 2021, https://www.ncci.com/SecureDocuments/QEB/QEB_Q4_2020_RemoteWork.html.
2. "COVID-19 Pandemic Triggers 25% Increase in Prevalence of Anxiety and Depression Worldwide." *World Health Organization (WHO)*, World Health Organization, 2 Mar. 2022, https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide.
3. Putzier, Konrad, and Peter Grant. "Record High Office Lease Expirations Pose New Threat to Landlords and Banks - WSJ." *WSJ*, The Wall Street Journal, 12 Apr. 2022, https://www.wsj.com/articles/record-high-office-lease-expirations-pose-new-threat-to-landlords-and-banks-11649764801.
4. "Is Remote Work Effective: We Finally Have the Data | McKinsey." *McKinsey & Company*, McKinsey & Company, 23 June 2022,

https://www.mckinsey.com/industries/real-estate/our-insights/americans-are-embracing-flexible-work-and-they-want-more-of-it.

5. Savage, Maddy. “What Bosses Really Think about Remote Work - BBC Worklife.” *BBC - Homepage*, BBC, 13 Sept. 2021, https://www.bbc.com/worklife/article/20210908-what-bosses-really-think-about-remote-work.
6. Bianchi, Tiago. “Reddit.Com Desktop Traffic Share 2022 | Statista.” *Statista*, Statista, 13 Jan. 2023, https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/.