~~Don't~~ Shoot the Messenger: Understanding Intrinsic Biases in News Readership and the Evolving Role of Modern News Platforms

Mansi Ganatra and Samar Haider

To be presented on December 4, 2019

1 Introduction

The rise of the internet has resulted in it becoming the primary means of information dissemination across the world. While traditional news readership numbers appear to dwindle, the world remains hungry for information, resulting in an ever-increasing traffic of users and views on online platforms. These platforms, in their original sense, were meant to serve primarily as a medium to aid the distribution of news, yet their role has recently evolved from one of a neutral platform to that of a curator. This has resulted in many blaming them for becoming increasingly partial to certain topics and ideologies and drifting dangerously close to the role of becoming content creators themselves, which would require them to be governed by a different set of laws entirely.

However, with the sheer volume of news published online every day, there has arisen a need to tame this firehose of information in order to make it useful to individuals. One such method is to create 'recommender systems', models which take into accounts users' preferences and tailor their suggestions accordingly. While they appear to be the right solution in theory, current methods used to build recommender systems rely on collaborative filtering, a technique which uses similarity metrics between different users to suggest items to them based on the preferences of their counterparts with similar activities. This can lead to the creation of echo chambers in news readership, leading to even more polarization in society by minimizing the cross-pollination of ideas from one community to the other.

2 Proposed Work

In this project, we aim to understand the phenomenon of online news readership, with a focus on analysing the trends that occur along different dimensions to answer questions such as: what kinds of news is published more; which topics are more widely-read; what differences there are, if any, in the demographics of people who read different genres of news stories; and finally, how all of these trends have evolved over time as online news readership has increased.

Finally, we hope to set the groundwork for building a bias-aware recommender system for a news publication. Bias here refers to the biases occurring organically in news readership or biases induced by a system/platform/algorithm. We recognize that a healthy diet of balanced news, both with respect to the topics as well as the stance on those topics, is an important component in building a well-informed society that is cognizant of the wide range of events and activities around it is incredibly important in the fight against the dangers of the social media fake news phenomenon. To

this end, we aim to build a recommender system that operates on content instead of social features, and attempts to encourage diversity in news recommendation to minimize the risk of creating echo chambers. Built using data[non-personally identifiable, anonymized] shared with us by the Los Angeles Times, this will be the first bias-aware recommender system for their news publication.

3 Data

3.1 Los Angeles Times

Our primary dataset will be sourced from the Los Angeles Times, one of the largest newspapers in the U.S. by daily circulation. The data will contain highly-detailed user activity on the Los Angeles Times website. All data will be non-personally identifiable, anonymized. The Los Angeles Times website, greets its customers with a manually-curated homepage, same as the one in their printed daily newspaper. While it is unusual for a large news publication to not have a recommender system, the lack thereof provides opportunities for insights in organic user behavior free from biases perpetrated by such a system. As the data acquisition is still under process, more details on its structure and granularity will be shared once available. We plan to study this dataset from a number of different angles, including but not limited to: understanding which kind of news stories are more prevalent, which genres of news have a bigger readership, what is the demographic make up of readers of different types of stories, what sentiment is found in the textual content of these stories, how all of these factors have evolved over time, and more. We will use this analysis to inform our creation of a bias-aware recommender system.

3.2 New York Times

Depending upon the acquisition of the previous dataset, another source for data we are considering as a fallback is that from the New York Times API. The New York Times hosts a number of different APIs on their website: article metadata, article search, book reviews, user comments, geographic linked data, social media sharing data, movie reviews, news tags, and semantic terms. They also have a recommender system in effect for years now. We plan to use these APIs to collect data from before and after the time that the New York Times launched their recommender system in order to study how trends in news readership on their website changed as a result of it, and if it did so for better or for worse with respect to news diversity.

3.3 Reddit

Our final fallback data source is Reddit. Reddit serves as a news aggregator and discussion forum with a massive network of over 130,000 active communities and more than 330 million users, making it one of the most popular websites in the world. Reddit is composed of numerous smaller communities called 'subreddits' which host discussions on certain topics. Of our interest are the subreddits r/news and r/worldnews, which have around 20 million users each. We wish to understand the dynamics of this social discussion on reddit pertaining to certain topics. In particular, we hope to look at upvotes as a proxy for the popularity of posts of certain topics and mine user comments for sentiment expressed by the community for that topic or story. We also plan to analyze which news sources are most popularly quoted on Reddit, and what kind of news is each used most for. We also aim to look at differences in reception of cross-posts, posts with the same content that are shared across different subreddits, and look at the variation in responses to them across different communities. Finally, Reddit tags as 'controversial' comments that have a high number of both upvotes and

downvotes, signifying polarization in views about it. We aim to mine these controversial comments to better understand the online conversation on Reddit.