Asch Harwood
Project Summary
December 16, 2020

# Predicting Audience Attitudes and Behavior from the News

## Background and Problem Statement

When it comes to data, the social impact world does not have the kind of access the private sector has to drive decision making and innovation. Unstructured data from news, social media, and websites, however, provides a rich, open source for developing insights that can drive transformative social change. I am all too familiar with this challenge, having committed the last decade of my professional life to applying data and technology to affect positive social change.

Understanding public narratives around civic engagement, which the media plays an important role in shaping, has been a central theme of this work. As a result, I decided to focus this work on how we can applynatural language processing and predictive modeling skills to look at how media influences the beliefs, attitudes ,and behaviors of audiences.

Based on this logic, I sought to answer the question:

> Can we use political news to predict how people might feel in the future about political or civic issues?

## Approach

To answer this question, I began with the negotiated media effects theory, which posits that media does shape our attitudes, beliefs, and behaviors. But how it does so is determined by our own lived experiences, social networks, and pre-existing beliefs.

I then conducted an intensive landscape analysis to identify similar efforts to use news to make predictions about audiences. The most promising work I found demonstrates modeling financial news to predict future broad changes in stock indexes.

Armed with a theory and promising case studies, I began gathering data with the intent of building a simple MVP-- sentiment classification for news articles. I then expanded the size and scope of my data, experimented with different audience measures for prediction, and eventually moved to answer my main research question. I approached my research in 5 phases, summarized in table 1 below in reverse chronological order.

**Table 1: Summary of phases, data, preprocessing, modelling and results**

| Phase | Feature/Source | Target/Source | Final Shape | Preprocessing | Models/*Package* | Accuracy |
|---|---|---|---|---|---|---|
| 5 | 214,000 news articles from 2015-2020, aggregated by hour, *GDELT/web scraping* | Hourly search volume for 'depression', 2015 - 2020, binarized and shifted 12 hours, *Google Trends* | 29,389 rows | -Tokenization<br>-Binarize target<br>-12 hour shift on target<br>-Train test split<br>-TF-IDF and count vectorization for logistic regression<br>-Pretrained word embedding for neural net | -Logistic regression with gridsearch/*Sklearn*<br>-Dense neural net with google news word embedding and 3 hidden layers/*Tensorflow*<br>-Other models test: RNN, CNN, LSTM, GLoVe only, Bert/*Ludwig* | Train: 67%<br>Test: 61%<br>Train: 82%<br>Test: 81% |
| 5.1 | 214,000 news articles from 2015-2020, with avg tone, polarity, pos and neg word count, and more, aggregated by hour, *GDELT/web scraping* | Hourly search volume for 'depression', 2015 - 2020, binarized and shifted 12 hours, *Google Trends* | 29,389 rows | -Tokenization<br>-Binarize target<br>-12 hour shift on target<br>-Train test split<br>-Pretrained word embedding for neural net | -Ensemble dense neural net with GLoVe word embedding and 7 tone scores with single hidden layer/*Ludwig* | Train: 75%<br>Test: 58% |
| 4 | 52,757 news articles from 2015-2020 with a 9 month gap of missing data, aggregated by hour, *GDELT/web scraping* | Hourly search volume for 'depression', 2015 - 2020, binarized, *no shift, Google Trends* | 33,207 rows | -Tokenization<br>-Binarize target<br>-Train test split<br>-TF-IDF with ngrams<br>-CountVectorization with ngrams<br>-Standard scaling<br>-Limit max_features on vectorization | -Logistic Regression with TF-IDF/*Sklearn*<br>-Logistic Regression with CountVect, ngram=(1,3), max_features=10,000 and standard scaling/*Sklearn* | Train: 66%<br>Test: 61%<br>Train: 64%<br>Test: 60% |
| 4.1 | 52,757 news articles from 2015-2020 with a 9 month gap of missing data, aggregated by hour, *GDELT/web scraping* | Hourly search volume for 'depression', 2015 - 2020, binarized, *24 hour shift, Google Trends* | 33,183 rows | -Tokenization<br>-24 hour shift on target<br>-Train test split<br>-Pretrained word embedding for neural net | -Dense neural net with google news word embedding and 3 hidden layers/*Tensorflow* | Train: 89%<br>Test: 87% |
| 3 | 52,757 news articles from 2015-2020 with a 9 month gap of missing data, aggregated by hour, *GDELT/web scraping* | Daily average social media sentiment, 2018 - 2020, binarized on mean sentiment, *no shift, Brandwatch* | 551 rows | -Tokenization<br>-Binarize target<br>-Train test split<br>-F-IDF<br>-Gridsearch with H20 automl and sklearn | -Logistic Regression with TF-IDF/*H20*<br>-Logistic Regression with TF-IDF/*Sklearn*<br>-Neural Net with TF-IDF/*Tensorflow*<br>-GBoost, Random Forest, Ensemble/*H20*<br>- SVC, Random Forest, SGDC, XGBoost/*Sklearn* | Train: 65%<br>Test: 66%<br>Train: 73%<br>Test: 65% |
| 3.1 | 52,757 news articles from 2015-2020 with a 9 month gap of missing data, aggregated by hour, *GDELT/web scrapin*g | Daily average social media sentiment, 2018 - 2020, *no shift, Brandwatch* | 551 rows | -Tokenization<br>-Train test split<br>-TF-IDF | -Neural net regression with TF-IDF/*Tensorflow* | Rsquared: -0.001763 |
| 2 | 52,757 news articles from 2015-2020 with a 9 month gap of missing data, *GDELT/web scraping* | Individual article *tone*, 2015 - 2020, calculated by *GDELT* | 52,757 rows | -Tokenization<br>-Binarize target<br>-Train test split<br>-Bag of words (TF-IDF and count)<br>-Unigrams, bigrams | -Logistic Regression with CountVectorizer<br>-Logistic Regression with TF-IDF<br>-Logistic Regression with TF-IDF and bigrams | Train: 81%<br>Test: 80%<br>Train: 89%<br>Test: 84%<br>Train: 81%<br>Test: 78% |
| 2.1 | 52,757 news articles from 2015-2020 with a 9 month gap of missing data, *GDELT/web scraping* | Individual article *polarity,* 2015 - 2020, calculated by *GDELT* | 52,757 rows | -Tokenization<br>-Binarize target<br>-Train test split<br>-Pretrained word embedding for neural net | -Dense neural net with Google news word embedding, 3 hidden layers | Train: 86%<br>Test: 82% |
| 1 | 4,630 climate change-related news articles from 2019-2020, *GDELT/web scraping* | Individual article polarity, 2019 - 2020, calculated by *GDELT* | 4,630 rows | -Tokenization<br>-Binarize target<br>-Train test split<br>-Count vectorization | -Logistic Regression with count vectorization | Train: 85%<br>Test: 82% |
| 1.1 | *Speaker quotations* from 4,630 climate change-related news articles from 2019-2020, *GDELT/web scraping* | Individual article polarity, 2019 - 2020, calculated by *GDELT* | 4,630 rows | -Tokenization<br>-Binarize target<br>-Train test split | -Logistic Regression with count vectorization | Train: 80%<br>Test: 69% |

## Feature Data

For my text data, I relied mainly on the GDELT Project, which "monitors the world's news media from nearly every corner of every country in print, broadcast, and web formats." While GDELT does not provide article text, I used it to search for relevant news articles and then extract metadata and source URL from which I could scrape text. GDELT also calculates 6 tonal scores for each news article it references, which I used for targets in phase 1 and 2 of my project.

Identifying a target useful for predicting future human attitudes or behaviors was a much more formidable task. In table 2 below, I've provided a summary of each dataset.

**Table 2: Summary of text and target data**

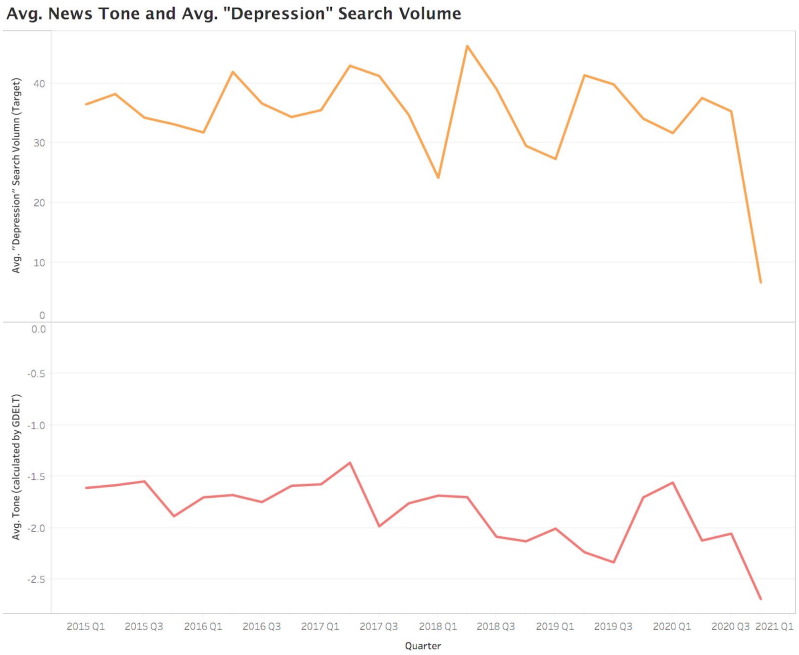
| Source | Description | Accessed | Shape | Feature/Target | Phase |
|---|---|---|---|---|---|
| GDELT | -Massive event dataset that provides access to its "global knowledge graph," where each row represents a news article.<br>-Does not provide actual article text but provides publish date, url, and thematic and tonal scores. | -Google Big Query and GDELT API and python package | -2015-2020 and yielded 214,000 usable news articles | -Tonal scores used as feature and target | All |
| Web scraping | -Scraped news article text based on GDELT urls from 12 American news sources | -Applied Newspaper3k package to scrape text from URLs | -2015-2020 and yielded 214,000 usable news articles | Feature | All |
| Brandwatch | -Provides average sentiment scores by predefined time range for Twitter, Reddit, and Youtube posts/comments | -Approximated query for GDELT and Big Query and extracted via Brandwatch csv export | -2018-2020 and yielded 551 rows of average daily sentiment via | Target | 3, 3.1 |
| Google Trends | -Provides hourly average search volume for word depression | -Google Trends API and pytrends python package | -2015-2020 and yielded 66,000 rows | Target | 4, 4.1, 5, 5.1 |
| Gallup/Pew | -Monthly polling data related to American opinions about government, politics, economics, and life satisfaction | -Manual download via Gallup and Pew websites | NA | NA | NA |

## Target Data

I eventually settled on Google Trends data for searches of the word "depression." Researchers have used Google Trends as a proxy for measuring American attitudes and behaviors before, including Seth Stephens-Davidowitz, author of Everybody Lies, who refers to this data as "digital truth serum." For me, Google Trends search volume for "depression" represented a measurable behavior --the act of seeking out information--as well for a proxy of American's wellbeing, which is theoretically influenced by the news they read.

## Visualizing the Data

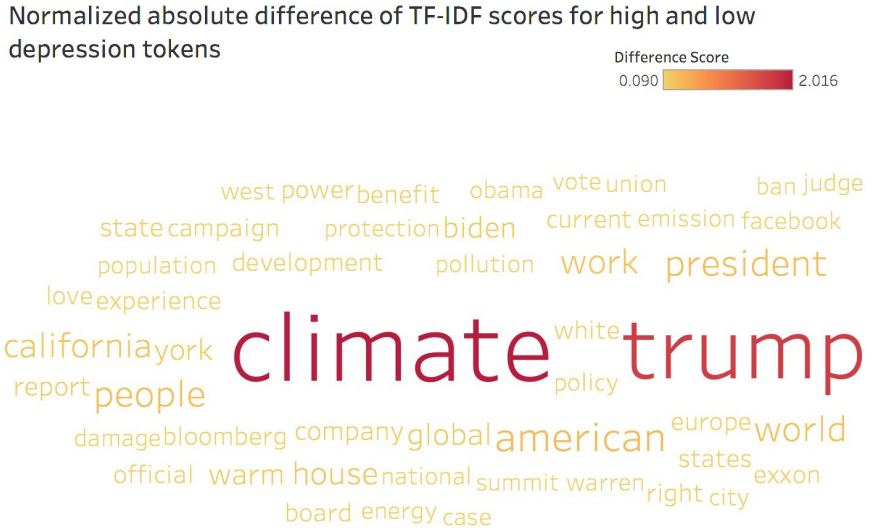
While my most successful modeling utilized pretrained word embeddings, which are very difficult to visualize, I used sentiment scores provided by GDELT to “peek” into the data.




The top line chart shows average search volume for depression from 2015 - 2020. This data has no statistically significant slope, despite the sharp drop in volume at the end of the chart. The second chart, which shows average news sentiment over time, has a moderate and statistically significant downward slope. Political news has been getting more negative over time.

I also applied term frequency-inverse document frequency (TF-IDF) preprocessing separately to the tokens associated with the 2 classes, high and low depression search volume. I then applied a standard scaler to the tf-idf scores and calculated the absolute difference for each shared token from each class. The resulting difference gives us an indication of the relative importance of each token for each class. By far, the most important words are “climate” and “trump”, where climate has a higher TF-IDF score when associated with low depression volume and “trump” is more strongly associated with high depression. The table below provides the TF-IDF scores for each class as well as the difference

| Tokens | Avg. Dif | High Depression | Low Depression |
|---|---|---|---|
| climate | 2.02 | 14.65 | 16.67 |
| trump | 1.67 | 23.15 | 21.48 |
| american | 0.33 | 7.65 | 7.97 |
| president | 0.27 | 13.39 | 13.12 |
| people | 0.26 | 12.23 | 12.49 |
| world | 0.25 | 7.31 | 7.07 |
| california | 0.22 | 6.78 | 6.57 |
| work | 0.21 | 7.39 | 7.18 |
| house | 0.19 | 7.22 | 7.03 |
| global | 0.17 | 5.96 | 5.79 |
| biden | 0.16 | 5.42 | 5.58 |
| york | 0.14 | 4.69 | 4.54 |
| warm | 0.14 | 4.74 | 4.59 |
| report | 0.14 | 7.92 | 8.06 |
| state | 0.13 | 11.74 | 11.61 |
| company | 0.13 | 7.13 | 7.00 |
| states | 0.12 | 5.23 | 5.35 |
| campaign | 0.11 | 6.08 | 6.19 |
| right | 0.11 | 6.05 | 6.17 |
| power | 0.11 | 5.26 | 5.37 |



## Data Preprocessing

Data preparation involved largely gathering text and target data via a combination of web scrapers (for text data), APIs, and dashboard exports, all of which are described in table 2. It then required stopword, punctuation, and whitespace removal as well as lemmatization, which I did using the natural language processing package, spaCy.

My Phase 5 dataset ultimately included 29,000 rows, where each row represents one hour of aggregated news tokens as features. The target is average search volume for depression 12 hours later, binarized into high or low search volume based on average search volume from 2015 to 2020.

## Modeling

I experimented with multiple text processing steps, including count and tf-idf vectorization with and without bigrams and trigrams--and multiple models, such as logistic regression, random forest, xgboost, svm, and neural nets.

For phases 1 and 2 of my project, I focused on article level sentiment classification, with binarized GDELT sentiment scores as my target. In this case, my best model was in phase 2, classifying binarized sentiment for 52,00 articles with logistic regression using TF-IDF. I achieved a train accuracy of 89 percent and a test accuracy of 85 percent. Achieving relatively decent results on my “mvp,” I moved onto the future audience behavior prediction problem.

For my final model, my goal was to use hourly aggregated news tokens to predict binarized search volume 12 hours later. My highest performing model, by far, was a feed forward dense neural network. I used a pretrained Google news word embedding layer, which is ideal for my dataset because it likely includes sources I used and thus has “prior knowledge” about my data. My final model then used three hidden dense layers with descending nodes (16, 8, 4) and dropout on the final layer. This architecture works because it strikes a balance by generalizing over my “noisy” data set without overfitting.

## Findings

I evaluated the model with a simple accuracy score on a test data set. My final train accuracy was 82 percent and my final test accuracy was 81 percent, which means my model correctly predicted high or low search volume 81 percent of the time. Given that this was a rather challenging prediction task, I consider this an extremely promising result.

The implication here is that there does seem to be a relationship between political news and how audiences feel after reading the news. It also seems to suggest that we can predict, to some degree, how political news from today affects people tomorrow, even if it is a rather simple way of describing those behaviors.

## Business Applications

While this is very much a prototype, it provides a potentially new template to assess the impact of media and narratives on audiences, which is important in the civic change space. Covid is a great example of how this might be helpful. With some creative tinkering, we could use an approach like this to analyze how covid is presented in the media, predict how it might impact people, and then use those insights to head off destructive narratives.

## Next Steps

Immediate next steps include additional evaluation, including calculating precision and recall, and building a light streamlit app that gathers the previous hour's news and then makes a prediction 12 hours later.

Longer term goals involve improving the robustness and specificity of the model. This includes collecting more data. Google Trends data is available back to 2004 and it may be possible to extract news data via GDELT over the same time period.

I would also like to explore more detailed predictions, such as making it a multiclass prediction or even a regression problem. There is also significant room for additional hyperparameter tuning as well as bringing in GDELT metadata and sentiment scores into the model.

Finally, I intend to present this to a number of colleagues in the civic change and international development space to further develop concrete use cases that I could potentially adapt this approach to.

## Additional Materials for Download

Video description of **final model and results** here (Youtube video)

**Final trained model** here (tensorflow saved_model.pb with weights)

**Final aggregated and preprocessed dataset** used to train the model here (CSV).

**Jupyter notebook** used to train the final model here (IPYNB)

**Conda virtual environmen**t to configure your own environment if you want to test the model yourself here

Collection of **jupyter notebooks for scraping, preprocessing, eda, and modeling** here and description of this collection here

Lightly processed, unaggregated **political news articles** with GDELT tone scores, March 2015 - October 2020 here