Predicting Polarization in Social Media

Project Team

Stephanie Ramos Gomez	stephanieramos@uchicago.edu
Jesica Ramirez Toscano	jramireztoscano@uchicago.edu
Kelsey Anderson	kjanderson@uchicago.edu

Project Summary

A diverse set of opinions and ideas in public discourse is desirable for many reasons, including the generation of novel solutions to social issues and the health of a nation's democracy. In fact, in the early days of social media, the internet was heralded as a tool to usher in a new age where everyone's voice could be heard, counted and healthy public debate would ensue (Puglisi & Snyder, 2015). Current views, however, center on the opposite: society appears to be increasingly polarized and hostile in online interactions. Online spaces have become the target of criticism and fears about cyberbullying, trolling, fake news and the potential for vulnerable individuals to become violently radicalized.

Because isolation and antagonism prevents the kind of generative dialog social media at its best can create, we seek to understand and explore the mechanisms involved in online polarization. We will examine if identifiable biasing traits in traditional media sources may be predictive of increased polarization in user comments using a neural network text analysis of twitter posts. We will use a variety of tools and approaches found in relevant machine learning research.

Current Research

An Cross-disciplinary Overview

Social media and the news are intimately intertwined in today's public arena. The sharing and resharing of events brings news to our fingertips with lightning speed. Issues of bias and polarization are studied in a wide variety of fields. Researchers are often concerned with several aspects of bias and polarization: the bias present in news media articles, social clustering along ideological lines which reinforces increasingly skewed versions of reality (“echo-chambers”) and the erosion of meaningful civil discourse in the general public.

Social media spaces, contrary to popular belief, may be more pluralistic than expected (Tucker, et. al, 2018). People tend to have broad networks and are exposed to many viewpoints on political issues at low levels of depth through non-political threads. Explicitly political topic threads, however, are more polarized in nature, with the most emotional or sensational content experiencing higher levels of user engagement (Tucker, et. al, 2018). The people who are most active on social media sites and have the most followers also tend to hold the most extreme views (Hong & Kim, 2016). Various researchers find that organically viral content is somewhat rare. Typically a traditional news outlet, public figure or social media influencer will highlight an issue and this action sets off a cascade of subsequent conversations (Tucker, et. al, 2018).

Since news is often accessed through social media, political news is often made visible in these channels by powerful institutions like news media and that these conversations tend to be the most polarized, it seems worthwhile to look at the factors that create or mitigate toxic or polarizing threads.

Research from Computer Science

Though many fields are interested in media bias and polarization, computer science and machine learning are uniquely positioned to expand our understanding of these issues because natural language processing has the ability to scan and look for patterns in large collections of text quickly. If models can accurately predict and mitigate things like bias or toxic content, perhaps polarization can be reduced and civil discourse improved.

About Polarization

Polarization as Connectedness

An area of interest related to the openness and health of discourse on social media relates to the overall polarization in social networks. In their 2013 paper, Guerra et. al examines three potential measures of community polarization: modularity, boundary node connections and popularity of boundary nodes. They use metrics computed on node graphs created with an open source automated graph visualization tool called Gephi to analyze six different datasets from twitter, facebook and blogs.

From the graphs they calculated the degree to which separate communities exist (modularity) using a formula shared with previous literature. Guerra et. al critique modularity, however, because it fails to distinguish between communities that are simply somewhat separate (in their datasets, for example, New York football fans and New York basketball fans) from communities where antagonism exists between groups (another sports example: between the supporters of rival Brazilian soccer teams). For polarization to exist, they argue, antagonism is an important distinguishing characteristic. To better capture this, they propose two measures of connections between groups: counting the proportion of in-group versus out-of-group connections in boundary nodes and calculating correlation between the popularity of a node and it's closeness to a boundary between groups.

The authors find that datasets they expected to exhibit polarization do show greatly reduced out-of-group connections on the boundary (Guerra et. al, 2013). They also find that, to a lesser extent, their hypothesis about popular nodes being internal ones for polarized networks and boundary nodes for non-polarized networks is generally true. The exception in this second metric was for US political blog posts. For news blogs, boundary nodes did tend to be in-group facing, however, among them there was a large number of popular boundary nodes. The authors explain this as the presence of news media blogs, which exist at the boundary because they are commonly referenced by both sides of a debate.

This finding related to US political blogs supports the idea that online interactions about policy and politics do contain a diversity of opinions with media outlets at the center and divergent communities drawing their content out into their own polarized spheres.

The Guerra et al. article fails to examine data at a lower level than network connections (i.e. retweets, follows or friend connections). So, while their findings are interesting, it is important to note that antagonism is semantic. A user might share a news story in order to promote it or mock it, depending on the caption or comment text they provide.

Polarization as Measuring Controversy

Controversial topics are often linked with hate speech and misinformation spreading, which is why it is important to identify controversy. Also detecting controversy provides a basis to offer the user different points of view by recommending them personalized content to read.

O. de Zarate et al. (2020) with a graph-based approach, measured controversy in 30 twitter datasets in 6 languages. To determine the “degrees of controversy”, the data goes through four phases. First, the graph building phase, in which they build a conversation graph that represents a retweet-graph of a single topic discussion. Second, the community identification phase involves using a graph-clustering algorithm, Louvain, to determine the communities involved in the discussion. In the embedding phase, each user is embedded into a corresponding vector encoding syntactic and semantic properties of the posts. To perform this step the authors selected two models among the most advanced ones, namely Fasttext and BERT, which embed texts into fixed dimension vectors encoding semantically significance and meaning. Finally, to compute the controversy score, they select some users as the best representatives of each community’s point of view. They used the HITS algorithm to estimate the authoritative and hub score of each user, and selected the central users, which correspond to the users with the 30% highest hub and authoritative score. Then the controversy score is defined by the distance between the central users and the rest of the users in each cluster.

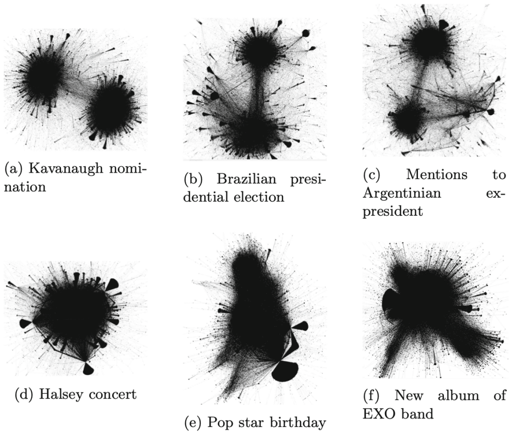

Figure 1. Controversial and non-controversial topics from O. de Zarate et al. (2020) twitter datasets. ForceAtlas2 algorithm used to layout the discussions.

This type of approach gives useful and promising results in measuring polarization, however it also requires the data being labeled to test the results. In this paper, the authors defined certain topics to be controversial, and from that they test if the data pipeline results were correct. So there isn't an established ground-truth.

For our project, we could use this approach to get "controversy" scores, as the measure of polarization in the twitter threads. It captures an important aspect of how far apart people's ideas are within a single topic, which is related to but missing from the Guerra et al. analysis of network polarization.

About Media Bias

Embedding or Neutralizing Bias

Given the popularity of using social media for news access, its unique space constraints and its rapid pace, the format lends itself to news items being summarized down to eye catching snippets. These snippets commonly contain links back to full articles or hashtags so conversations can be tracked, but the compression of information into such small bites makes bias based on filtering out contextualizing information a serious concern. Trolls and others interested in polarizing or hijacking public discourse can easily exploit this phenomena.

Recently, two similar projects used ideas of bias and sentence extraction summarization to create new versions of news media articles, one for the purpose of aggregating out all bias and the other for the purpose of injecting bias.

In the neutralizer case, Lee et al. used media bias ratings from allsources.org (which is a largely survey-based measure, but include multiple metrics) to label stories as left, right or center on the basis of their publisher (2021). The authors used a web crawler to extract 1,740 full news articles that were then grouped manually into triplets each containing a left, right and centrist text relating to the same news story. The authors trained a bias score based on these groups and tested their model on a dataset of 300 articles from an earlier study which used a sentence-level scoring metric to determine bias (the BASIL dataset). For their model, the authors used two scores: a measure of the number of biased spans in the article and a measure of the relevance of the article based on the amount of unique information from a concatenation of all three sources to the final neutralized summary text. They combine these into a total score that equally weights being free of bias and full of content. Then Lee et al. used a neural network with a Bidirectional and Auto-Regressive Transformer (BART), and alternatively tested a T5 algorithm, to learn and predict which sentences/spans should come next in a logical news article that maximizes neutrality.

Lee et al.’s outcomes were evaluated on the amount of bias their summaries were able to eliminate and the content relevance remaining (2021). With different training and tuning, they found the presence of informational bias could be drastically reduced, though lexical (word choice) bias was harder to eliminate. The algorithms they chose tended to favor very short summaries, which compromised their relevant content scores in the process.

In the bias injecting case, Jachim et. al describe creating a news summarizer that intakes a full length article and, given a desired political position, will create a summary that extracts actual quotes to create social media ready politically slanted summary text of a desired target length (2021).

The authors web scraped articles from Vox and Breitbart to establish a baseline of liberal and conservative bodies of text on which to train their machine learning model (Jachim et. al, 2021). They used term frequency inverse document frequency (TF-IDF) to identify words, bigrams and trigrams that were more likely to be important from each news source. They also found that the dataset was unevenly sampled, so used synthetic minority oversampling (SMOTE) to make new vectors to even out the number of important items between classes. Then they trained a random forest classifier to identify the probability a sentence is liberal or conservative on the basis of term frequency.

The class probability scores fed into a sentence extraction summarizer with an exponent parameter that would allow the user of the summarizer to specify how much slant they wanted the summary to contain. The summarizer itself was based on earlier work which found a sentence scoring mechanism that combines TF-IDF, sentence position and sentence similarity to the article's title could create good extraction summaries for news articles (Ferreira et al, 2014). So Jachim et al. essentially tampered with the summarizer scores, giving more weight to sentences that appeared more like those found in liberal or conservative media sources to create bias in the result.

To analyze their results of their bias-injecting summarizer, the authors used a set of linguistic scores called the Linguistic Inquiry and Word Count (LIWC), which gives scores related to the tone, analytic thinking, clout and authenticity displayed in a given text. They found that all summaries (original formula, biased liberal and biased conservative) had high levels of analytical thinking and clout. They all had relatively low levels of authenticity (because, they conjecture, summaries tend to use recognizable phrases rather than unique compositional styles). They also found that conservative summaries tended to be more negative in tone and liberal summaries tended to be more positive compared to original summaries.

Both sets of researchers use existing media provider slant ratings to help their models learn what text content is uniquely relevant to liberal or conservative spin. Though trained on these weakly labeled datasets, neither study evaluated their success solely on a measure of accuracy related to them. Instead, they performed extensive descriptive analysis of their results and reported patterns they found relevant to their original research questions.

The idea of using other dimensions in classifying a body of text beyond liberal/conservative or negative/positive is interesting. LIWC is a dictionary tool, used in Jachim et al. 's work, that scores a corpus on the basis of words and phrases present within it. Ultimately it is the content of the summary that makes it polarizing, not just the source, and capturing that may require something more nuanced than binary classification. For our project, we will look for a novel way to capture the impact of the context in order to better understand patterns in what exactly is polarizing about biased materials.

Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media

Salminen et al. (2018) collected comments from a YouTube channel and Facebook page of a major online news organization and created training data by identifying hateful comments using human judgment. Then they created a taxonomy of different types and targets of online hate, and trained machine learning models to automatically detect and classify the hateful comments in the full dataset.

They collected data from a major online news and media company that posted several videos per day on YouTube and Facebook and received thousands of comments per video. These comments were collected using YouTube and Facebook APIs. They pulled 137,098 comments from videos posted on YouTube and Facebook, in the period of July-October 2017, and focused only on English comments. They explored hate in the dataset by building a dictionary based on public sources of hateful words and a qualitative analysis. Searching with that dictionary, they found that 22,514 comments (16.4%) contained hateful content.

Then they proceeded to assign labels to a portion of the data. They considered linguistic attributes when annotating. Swearing, aggressive comments, or mentioning the past political or ethnic conflicts in a non-constructive and harmful way, were classified as hateful. To divide the comments into groups depending on their target and type, they used manual open coding, which uses human judgment for defining the categories. The taxonomy has 13 main categories and 16 subcategories (29 in total). The main categories include targets and language, 9 describing targets (political issues, religion, racism, etc.) and 4 the type of language (accusations, humiliation, swearing, promoting violence).

To perform the supervised classification, they developed three sets of feature categories. The feature categories are n-gram, semantic and syntactic, and distributional semantic features. Before feature extraction, preprocessing was performed, removing stop words from comments and stripping the tokens of any trailing special characters or space. They experiment with a set of machine learning algorithms to perform multilabel classification of the dataset. In this experiment, they used the 5,143 labels annotated using the taxonomy. This dataset was split into training and testing (33% for testing) the classification models. Five different models were used: Logistic Regression, Decision Tree, Random Forest, Adaboost, and Linear Support Vector Machine (SVM). For each model, they tuned the parameters using scikit-learn's grid search method in Python.

The average F1 score of the 21 categories and subcategories is used to report the overall performance as it combines both the precision and recall. The average precision of the best model, SVM, was 0.90, and recall 0.67 (avg. F1 score = 0.79). The recall score indicates the model is struggling to classify some categories, most notably religion (recall=0.3). Then they apply the classifier to the full dataset to classify the comments by the main categories of hate.

Most common is hate against the Media (17.0% of instances), mostly against the particular news outlet which is referred to as “propaganda,” “fake news,” and “lies.” Another popular target is the Armed Forces (16.9%), particularly the police. The most typical language type is humiliation (31.5% of language observations) but swearing (29.3%) is also common. Somewhat alarming is the share of promoting violence (18.0%), clearly indicating that many comments are toxic.

This paper provides evidence that news media posts generate hateful comments and helps to justify the idea of using the text of a news media tweet to predict comment polarization.

About the Impact of Media Bias

Media bias detection and bias short term impact assessment

Media has the power to influence the opinions of masses which in turn conditions and influences their day-to-day activities. Aggarwal et al. (2020) used Tweets from popular media outlets and journalists in India to investigate how biased opinions are channeled and propagated via social media into a network of people following/consuming it.

On November 8, 2016, the government of India announced the Demonetization of Rs. 500 and Rs. 1000 currency notes to curb the menace of black money. This provoked wide-ranging reactions from people, with some appreciating the move and some criticizing it. This study uses the Twitter API to fetch real time tweets from the verified handles of 10 media outlets and 10 journalists. A total of 4,475 tweets are collected for the subject of “Demonetization.” Some pre-processing was done in order to make the tweets suitable for further computation: URLs were replaced with a tag “URL”, targets were replaced by the username, punctuation marks occurring at the start or end of a word were separated from the word to enable easier tokenization.

Sentiment scores for the tweets were determined by analyzing the tweets collected using VADER. NLTK-VADER is a popular tool for sentiment analysis of social media texts. The sentiment scores obtained by VADER are used for determining the polarity of text. VADER provides a polarity score (PS) ranging from -1 (strongly negative) to +1 (strongly positive). This range is split into 3 regions, namely positive, negative and neutral. PS values ranging from -1 to -0.33 are classified as negative PS values while PS values ranging from 0.33 to 1 are classified as positive PS values. PS values lying between -0.33 and 0.33 are referred to as neutral PS values.

The polar tweets form the basis for analyzing conditioning by media outlets. For each polar tweet, a list of user-ids of the users who ‘react’ to the given tweet is also extracted. The behavior of ‘reactors’ to polar tweets is observed to study the conditioning of opinions by the media outlets. Each user has prior opinions on issues. The prior polarity (p) of a person is calculated by the exponential moving average of the polarity score of its previous tweets. However, a polar tweet/news report by a media outlet can affect the user’s opinion on the issue. The polarity in the opinion of an user might show a massive swing (reversing one’s opinion) or a minor one (a small strengthening of prior opinion). The post-event polarity (p’) is obtained by taking the arithmetic mean of the polarity scores of its tweets post the event for a period of 24 hours. An event E is said to have conditioned a consumer C if the value of p and p’ are significantly different. The two opinions p and p’ are assumed to be significantly different if the difference of p and p’ is 20% more than the standard deviation in the PSs of C prior to event E.

A large fraction of tweets by media outlets/journalists are observed to be polar or subjective in nature. More than 40% of the tweets analyzed were found out to be polar. Also, a fraction of as high as 41% of reactors were conditioned through such ‘biased’ polar tweets. This shows that an alarming number of people get conditioned by polar, subjective and biased news every day, something, which goes past the naked eye of humans around the world.

This paper provides a useful example of using VADER to assign tweets from popular media outlets that will serve as a guidance to construct the labels for our dataset. It also provides evidence that there is a short term impact of media bias.

About Sentiment Analysis of Short Texts

Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts

Sentiment analysis of short texts is challenging because of the limited contextual information they usually contain. Wangrt al. (2016) describe a jointed CNN and RNN architecture, taking advantage of the coarse-grained local features generated by CNN and long-distance dependencies learned via RNN for sentiment analysis of short texts.

- MR: Movie reviews with one sentence per review. Classification involves detecting positive/negative reviews

- SST1: Stanford Sentiment Treebank - an extension of MR but provided five kinds of labels, very negative, negative, neutral, positive and very positive
- SST2: Same as SST1 but with neutral reviews removed and binary labels

Due to the fact that CNN can extract local features of input and RNN (recurrent neural network) can process sequence input and learn the long-term dependencies, they combine both of them in sentiment analysis of short texts. The model consists of the following parts: word embeddings and sentence-level representation, convolutional and pooling layers, concatenation layer, RNN layer, fully connected layer with softmax output.

First, they perform unsupervised learning of word-level embeddings using the word2vec method and also test random initialization. Then they use a convolutional layer to apply a matrix-vector operation to each window of size w of successive windows in the sentence-level representation sequence. The same weight matrix is used to extract local features for each window of the given sentence. Using the matrix over all word windows of the sentence, they extract the n-grams feature vectors of size $l – w + 1$. They take these features as the input of the recurrent neural network and apply LSTM and GRU. The features generated from RNN form the penultimate layer and are passed to a fully connected softmax layer whose output is the probability distribution over all the categories.

They take cross entropy as the loss function that measures the discrepancy between the real sentiment distribution and the model output distribution of sentences in the corpora. The entire model is trained end-to-end with stochastic gradient descent.

Models with pre-trained vectors from word2vec and max pooling perform best among all the models. The classification accuracy is raised by 0.7% on MR and 1.8% on SST2, when implementing CNN-GRUword2vec model compared with the existing models. Models with pre-trained vectors all perform better than the others with randomly initialized vectors on all three corpora. They find that the jointed architecture of CNN and RNN model performs better than the CNN and RNN models alone in sentiment classification of short texts. This paper provides an innovative approach to building a classification model for short texts that will be useful in structuring the neural network approach for our model.

Plan of Action

Our goal for this project is to use deep learning to see what, if any, components of news media posts lead to increased social media polarity. Our model will learn from the tweet text of three major news outlets (one liberal, conservative and centrist source based on allsides.org ratings) and attempt to predict two different measures related to polarization: variance in sentiment and controversy. From an analysis of these predictions, we hope to demonstrate attributes of media behavior that lead to increased emotionality and divergence of opinions online. We seek a variety of news outlets in order to see if perceived political bias on the part of the outlet creates a difference in response beyond what is embedded in the text of a tweet.

Our measure of sentiment will be based on the open source VADER sentiment scoring tool used in Aggarwal et al. (2020). We will define a polarizing tweet as one that creates large variance in the sentiment scores of subsequent comments. This effectively captures polarization in emotionality of a body of responses.

Our measure of controversy will be based on the graphical distance of text encodings used in Ortiz de Zarate et al. (2020). This effectively captures polarization in the meanings expressed in a body of responses.

As we will examine if media bias may be predictive of increased polarization in user comments, the following is our pipeline proposal:

1. **Pull data from Twitter AP**I:
    a. Tweets that occurred during the month of March 2021.
    b. Posts that have 20+ comments.
    c. From news outlets, that according to allsources.org, represent large readerships and are center, left and right leaning.
        i. Center: Reuters
        ii. Left: New York Times
        iii. Right: Fox News

2. **Preprocess Data**
    a. Clean form stop words and special characters using spacy library
    b. Obtain bi-grams
    c. Lemmatize words

3. **Clustering tweets to find bodies of similar political topics**
    a. Using LDA or FastText and KMeans obtain the main topics of the requested tweets.
    b. Filter tweets on one to three topics for analysis, depending on corpus size.

4. **Building polarization scores**
    a. First approach: Obtain variance of sentiment score.
        i. Pass each selected comment tweets through VADER for a sentiment score.
        ii. Aggregate sentiment scores to find standard deviation and IQR.
    b. Second approach: Obtain controversy score.
        i. From a graph-based algorithm following Ortiz de Zarate et al. (2020)

5. **Feature set**:
    a. Link original news article to the level of polarity comments.

6. **Training phase**:
    a. Train multiple neural networks models (maybe different hidden layers) to predict the level of polarity in the comments based on original post text.

7. **Test phase**:
    a. Compare the predicted polarization values with the original polarization scores obtained in step 4.
    b. Show the MSE of the different applied models.

8. **Final Project Report**
    a. Write results, implications and limitations of our work.

Mid-Term Presentation Goal

For the mid-term presentation we plan to have completely constructed our dataset and decided on neural networks approaches to build and test. Also, we might have filtered the comments on the topics that we want to study. We will present our approach and a descriptive analysis of our dataset.

Final Presentation Goal

For the final presentation we will have selected our optimal model and analysed our prediction results. We will present our findings and their relevance to our research question: can elements of a news media post predict online polarization?

Team-member Specific Plans of Action

Tasks assigned to Kelsey

Kelsey will be responsible for pulling the tweets from the “right” leaning media outlet, Fox News, twitter feed.

After all tweets from all sources have been downloaded and combined into a single dataset, she will extract the response tweets and pass them through the open source VADER sentiment scoring tool (Hutto & Gilbert, 2014). She will then aggregate sets of scores to calculate standard deviation and interquartile ranges for the body of comments related to each original news media post.

Once topic clustering and controversy scores are completed, Kelsey will combine topic, controversy and sentiment scores into one dataframe to create the final feature set.

After the final feature set is constructed, Kelsey will take one neural network approach or algorithm to build a pipeline for and train a model on. The final algorithms for comparison in this step will be decided through future group conversations.

In the final stages of the project, presentation writing and final analysis tasks will be discussed and assigned as appropriate in light of our findings.

Tasks assigned to Jesica

Jesica will pull tweets from the “center” leaning media outlet, “Reuters”, twitter feed.
To approximate a measure of polarization, Jesica will follow Ortiz de Zarate et al. (2020) in which using a graph-based approach one can get “controversy scores” from embedded comments of twitter. The comments will be cleaned and embedded with the FastText Model, the encodings will be useful to calculate how far the comments are from a “central” comment. With these we will get an overall score of controversy for each news article.
Once the dataset is ready for training, Jesica will train a type of neural network model to predict social media polarity.
Jesica will also do general tasks as preprocessing data, mid-quarter presentation, writing findings and final analysis for the project.

Tasks assigned to Stephanie

Stephanie will pull tweets from the “left” leaning media outlet, “New York Times”, twitter feed.

Stephanie will also be responsible for preprocessing the data. In order to make the tweets suitable for further computation it will be necessary to remove HTML tags, remove extra white spaces, remove special characters, lowercase all texts, remove stop words, lemmatization.

Then, she will cluster the tweets into 10 topics using nearest-neighbor algorithm. Each article in the corpus is replaced with a TF-IDF vector, which is effectively an abstract numerical representation of its relative information content. TF-IDF vectorization is a popular method for topic mining. These groups will be used in the evaluation stage, where we will check the accuracy of the predictions for each topic.

Finally, she will train a machine learning model to make predictions on the dataset. Possible models are: Random Forest, SVM and CNNs.

Evaluation of Success

Evaluating the machine learning algorithm is an essential part of any project. To evaluate the success of our model we will calculate the Mean Squared Error (MSE) of the testing set. MSE takes the average of the square of the difference between the original values and the predicted values. As we take the square of the error, the effect of larger errors becomes more pronounced than the effect of smaller errors, hence the model can now focus more on the larger errors.

$$MeanSquaredError \; = \frac{1}{N}\sum_{j=1}^{N}(y_j - \widehat{y}_j)^{\;2}$$

We will calculate the MSE for all the observations in the test set over both sentiment variance and controversy score predictions as well as for each topic and each media source.

After optimizing our model, if a suitably predictive model is found, we will examine feature importance to draw larger conclusions about tweet content and its relationship to online polarization.

References

- Aggarwal, S. Sinha, T., Kukreti, Y., Shikhar, S. (2020). Media bias detection and bias short term impact assessment. Array, Volume 6, 100025, ISSN 2590-0056, https://doi.org/10.1016/j.array.2020.100025.

- Ferreira, R., Freitas, F., Cabral, L. d. S., Lins, R. D., Lima, R., França, G., Favaro, L., & Simske, S. J. (2014). A Context Based Text Summarization System. 11th IAPR International Workshop on Document Analysis Systems, Tours, France; IEEE. https://ieeexplore.ieee.org/document/6830971

- Guerra, P., Meira Jr., W., Cardie, C., & Kleinberg, R. (2013, June 28). A Measure of Polarization on Social Media Networks Based on Community Boundaries. Proceedings of the International AAAI Conference on Web and Social Media; AAAI Publications. https://ojs.aaai.org/index.php/ICWSM/article/view/14421

- Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

- Jachim, P., Sharevski, F., & Pieroni, E. (2021, April 1). “TL;DR: Out-of-Context Adversarial Text Summarization and Hashtag Recommendation”. ArXiv.Org. https://arxiv.org/abs/2104.00782

- Lee, N., Bang, Y., Madotto, A., & Fung, P. (2021, April 1). Mitigating Media Bias through Neutral Article Generation. ArXiv.Org. https://arxiv.org/abs/2104.00336

- Ortiz de Zarate, J., Di Giovanni, M., Zindel E., and Brambilla, M. (2020) Measuring Controversy in Social Networks Through NLP. C. Boucher and S. V. Thankachan (Eds.): SPIRE 2020, LNCS 12303, pp. 194–209, 2020. https://doi.org/10.1007/978-3-030-59212-7_14

- Salminen, J., Almerekhi, H., Milenković, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. J. (2018). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In 12th International AAAI Conference on Web and Social Media, ICWSM 2018 (pp. 330-339). (12th International AAAI Conference on Web and Social Media, ICWSM 2018). AAAI press.

- Wang, Xingyou and Jiang, Weijie and Luo, Zhiyong (2016). Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts. Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2428--2437). https://www.aclweb.org/anthology/C16-1229