Neural Networks for the Summarization of News Articles:

Reducing Bias and Misinformation

Silvia Miramontes, Emma Russon
University of California, Berkeley
s-miramontes@ischool.berkeley.edu, erusson@berkeley.edu

Abstract

To mitigate the bias and misinformation in existing news articles, we confront the discernible subjectivity in news platforms, while maintaining the valuable content of their published material. We aim to build a news search engine that takes a user-given topic of interest. Our algorithm then clusters relevant articles based off of the user's input and creates an extractive summarization from the returned clusters, all while penalizing known bias from certain publishers. In the baseline model, we use TF-IDF and Word2Vec features with k-means clustering to determine the clusters of closest similarity to the input, and then use k-nearest neighbors to classify the cluster that the input topic belongs to. Future work includes creating weights that penalize known bias and producing an extractive summary from the articles in the cluster. Our results can be reproduced via the provided Google Colab notebook or the provided Github repository.

Keywords: Neural Networks, NLP

1. Introduction

With the rise of mass media comes an increased responsibility to keep the public informed with integrity and transparency. However, to remain competitive in the media industry, platforms seek to increase readership and engagement by releasing more material, often leading to low quality and biased information being distributed to the public.

Over the last few decades, it has become evident that the information being released by news platforms has started to deteriorate, as many of their publications are often found to lack factual information. A study conducted by RAND analyses "Truth Decay" – the diminishing role of facts and analysis in the American Public life. This is evidenced by:

1. Increasing disagreement about facts and analytical interpretations of facts and data.
2. Blurring of the line between opinion and fact.
3. The increasing relative volume and resulting influence of opinion and personal experience over fact.
4. Declining trust in formerly respected sources of facts.

Although these trends are not novel in American history, it is evident that the level of disagreement on objective facts is a new phenomenon [1].

Nowadays very few readers strive to find the most accurate news sources. Even though American journalism was once known to be accurate by providing factual information to its audience, the increase in accessibility to information has led individuals to seize from looking for reliable news sources and instead seek information that parallels their personal opinions. Since most information is now provided based on recommendation systems in the consumer's mobile phones, there is a decreasing desire from consumers to critically evaluate the provided and published information.

2. Background

Here we elaborate in existing work addressing bias detection in political articles, as well as summarization tasks in NLP.

2.1. Bias Detection

Following recent presidential campaigns that would leverage various media outlets, extensive research has been published to identify political bias within speeches, statements, and news articles. Misra et al. developed a Long Short-Term Memory network trained on a dataset of ideological sentences which were manually labeled to be conservative or liberal. Their model's performance had an F1 score of 71.8% when detecting any implicit political bias in a text [2].

Another model which used multilayer perceptrons on the same training data, to predict the political bias in news articles, achieved an F1 score of 81%. This increase in accuracy could be attributed to the ability of MLP models to handle semantic parsing, sentence modeling, classification, and prediction [3].

Other existing work for bias detection used a recursive neural network to determine the political bias in sentences and a domain-informed Bayesian hidden Markov model to produce proportions of various ideologies in U.S. presidential candidate speeches [3]. Our paper aims to adopt the methodologies of better performing models, such as multilayer perceptrons, to identify the existing bias in news articles.

2.2. Summarization

Another NLP task that has been around for decades is the automation of text summaries. Assisting readers in easily digesting loads of information from medical literature reviews to film plot summaries, automatic summarization accelerates and increases the knowledge of consumers. With text automation comes the challenges of selecting the most essential content in the text and how to appropriately condense the information. There are two types of text summarization: extractive and abstractive. Extractive summarization simply selects content from the original text, with a focus on relevance and importance. Abstractive summarization provides a novel summary of the original text, with the complexities of semantics, generalization, and domain knowledge. Our model will implement extractive summarization using TextRank, which finds the sentences in a corpus that are the most similar as a proxy for relevance. If time permits, we will also seek to apply other methods such as SummaRunner and Nucleus, which have higher ROUGE scores than TextRank [4].

3. Data and Methodology

In the subsections below, we describe the data we utilize for the development of this model, as well as the methods utilized for its completion. We add a note at the end where we elaborate on the computing architecture utilized for the improvement in performance and accuracy of our algorithm.

3.1. Data

The data to be utilized is obtained from Kaggle, notably ”All The News”[5] and ”BBC News Summary” [6] for the clustering and training of our neural network, respectively.

To create clusters and determine the bias of existing articles, we will utilize ”All The News” data which contains the information and content of 200,000 articles along with their publications from 2016 to 2017. To determine the bias of these articles, we build upon existing bias detection algorithms by incorporating the known bias of certain news platforms such as Fox News and Buzzfeed.

The ”BBC News Summary” is a data collection of articles from various news sections (e.g. sports, business, politics) and contains five extractive text summaries for each of the articles, creating a total of 2225 original documents and 11125 summaries. This dataset will be used to train the extractive summarization model.

3.2. Clustering

We use TF-IDF and Word2Vec features with k-means clustering to determine the clusters of closest similarity. For the purposes of the baseline model, we randomly sample and preprocess data in ”All The News” to meet memory restrictions and fit the model. We then use k-nearest neighbors to classify the cluster that the user-given topic belongs to, which will be the resulting top 5 recommendations of closest similarity to the user's input.

TF-IDF: Term frequency-inverse document frequency is a numerical statistic for significance of a word in a corpus. The statistic considers the number of times a term appears within and across documents. When applied to our corpus of articles from ”All The News,” the tf-idf vectorizer creates a matrix of n-gram features for every document in the corpus. At this level, we feed in the article titles. The benefit of this method is that it incorporates the relevance of certain terms throughout the entire corpus, while a shortcoming is that it doesn't capture the full context or semantics of the document.

Word2Vec: This is a shallow, two-layer neural network trained to reconstruct linguistic context of words. Word2Vec takes as input a large corpus of text, in this case a set of news articles. The algorithm later produces a multidimensional vector space. One of the benefits of this model is that the word vectors within the generated vector space are positioned such that the sentences that share common contexts in the corpus are located close to one another in the space. Additionally, embedding vectors created using Word2Vec allows us to take into account latent semantic analysis within the inputted sentences.

K-Means: This clustering method allows us to group the outputted observations from either Word2Vec or TF-IDF with aims to minimize the total intra-cluster variance. For our baseline model, the k-means method clusters titles that are similar with respect to the features developed from the TF-IDF vectorizer and Word2Vec network.

K-NN: The K-Nearest Neighbor algorithm aims to find the k closest training examples in the feature space by minimizing the Euclidean distance. This allows us to predict which cluster the user-given input belongs to, and thus, return the top five articles in the corpus that match the topic of interest, which will later be used for summary extraction.

3.3. Bias and Summary

At this point, we have not yet implemented bias detection and text summarization in our model.

Our intention is to detect the bias in the top 5 clusters from the users input, and penalize those news platforms that are predominately biased based on their article content. We hope that by penalizing biased information, the outputted extractive summarization will be objective to the user. In section 5, we discuss our upcoming approaches to achieve this stage on this project.

3.4. A Note on Performance

As mentioned previously, we reduced the size of our training data for the baseline model to 10,000 randomly selected observations in order to accommodate for memory restrictions on hosted and local servers. This may have impacted the accuracy and strength of our clusters and classifications. We will overcome this by migrating to Google Cloud Platform (GCP), which can host and process our data at a significantly higher memory. Additionally, with GCP we will be able to connect to the Google Colab platform, which will aid in the reproducibility of our work. To look at our code, please reference the github repo on branches ‘emma‘ and ‘silvia‘.

4. Results and Discussion

The general architecture for the baseline model is to first preprocess the data and create a matrix with features for each document, which in this case is an article title. We then feed the matrix into a clustering algorithm, and then classify the cluster that the user-given input belongs to. To build the matrix that we fit the models with, we use tf-idf and Word2Vec. The results of each method are printed below in Table 1 and Table 2.

Even though the within-cluster sum of squares (‘Inertia‘) is much higher for the output of the tf-idf vectorizer, the baseline model outputs of the top five article titles in each cluster looks more promising with tf-idf produced features.

Cluster 0
Bernie Sanders knows he’s going to lose. Here’s how you can tell. The States Fight Back Supreme Court Set To Hear Church-State Case In Gorsuch’s First Week Lifesaving Flights Can Come With Life-Changing Bills London’s East End ’Like Baghdad’, As Cockney Culture Dies Out
Cluster 1
A Doorwoman in a Doorman’s World Russia Should Be Barred From Rio Olympics, World Anti-Doping Agency Says Who’s Encouraging Anti-Semitism? Confronting Anti-Semitism In Russia, In Words And Then Music Which are the most corrupt cities in the world?
Cluster 2
Bill Clinton congratulates Donald Trump It’s Donald Trump’s Party Now Donald Trump Gets Defensive Cruz: ’I Don’t Intend To Insult Donald Trump’ The Republicans Giving In to Donald Trump
Cluster 3
Did Hillary Clinton Have to Be First? Why Do They Hate Hillary Clinton? On Hillary Clinton’s Pandering Hillary Clinton is almost certain to be president Who’s more electable: Bernie Sanders or Hillary Clinton?
Cluster 4
There Is No Trump Campaign How Trump could still win Trump Wins. Now What? Fixing Immigration, Trump or No Trump Trump: ‘We’re not going into Syria’
Inertia: 9851.55

Table 1: Results from TF-IDF Vectorizer

Cluster 1
EPA to pull back on fuel-efficiency standards for cars, trucks in future model years GOP Senator To Angry Constituents: Schedule Your Protest In Advance! Cleveland police shooting of Tamir Rice: city to pay $6 million after 12-year-old’s death This woman chose to go homeless in San Francisco instead of paying high rent Trump tours a hulking aircraft carrier to promote hike in military spending
Cluster 2
Louisiana’s Democratic Governor Robs Kids of School Choice Protesters Crowd Outside Fox News’ GOP Debate In Detroit (PHOTOS) Mary Tyler Moore, TV legend, has died at 80 ICE Deports MS-13 Gang Member Wanted for Violent Crimes in El Salvador Fifa hands authorities 20,000 pieces of evidence as internal inquiry concludes
Cluster 3
How Malay Spends His Sundays: Cooking at Home, Music in the Studio - The New York Times On Twitter, Hate Speech Bounded Only by a Character Limit - The New York Times Japanese Boy, 7, Left on Mountain by Parents Is Found Alive - The New York Times Warren Buffett and Dan Gilbert Unite in Bid to Acquire Yahoo - The New York Times Donald Trump, After Difficult Stretch, Shows a Softer Side - The New York Times
Inertia: 565.48

Table 2: Results from Word2Vec

In Table 1, we see that the model focuses on similar n-grams such as presidential candidate names like ”Hillary Clinton” and ”Donald Trump”. When feeding the topic ”Hillary Clinton emails” into the k-NN model, the input is classified to Cluster 3, which makes sense when considering the top five articles in the cluster. However, another test user-input ”Hillary Clinton defends handling of Benghazi attack” is assigned to Cluster 0, which all in all is not a very coherent cluster of article titles. At this point, we see that the tf-idf version of the baseline model is able to create clusters with varying levels of accuracy. We will move forward by running the model on all the available data and tuning cluster size to produce better results.

In Table 2, the clusters formed by Word2Vec encourages us to implement other approaches to take full advantage of the semantics considered in this model such as incorporating the entire corpus of data. Nevertheless, there are existing similarities within the clusters in regards to specific words present in some sentences, such as cities, or the usage of words referring to violent acts. Both of the user-input topics, ”Hillary Clinton emails” and ”Hillary Clinton defends handling of Benghazi attack,” are classified to Cluster 3 when fed into the k-NN model, which doesn’t make much sense when considering the top five articles in the cluster. Our efforts are to obtain the best matching clusters to the user’s input, and surprisingly the Word2Vec model is not performing to expectation. Below we discuss the next steps to take for the improvement of our clusters, and the application of bias to the summarization models.

5. Next Steps

Prior to moving forward to the calculation of bias from the generated clusters and the summarization, we seek to improve the clusters based on the user’s input. One idea is to also consider the article content for context as opposed to only running TF-IDF and Word2Vec on article titles. By doing so, we would diminish the potential bias existing within each title, as each news platform may utilize certain keywords to attract audiences. Additionally, by utilizing all the available 200k articles by taking advantage of the extensive memory of GCP’s virtual machines and deep learning environments, we are confident on the improvement of our clustering methods.

With the improvement of our clustering methods, we will proceed to detect the bias in the article content of the top clusters to then generate an extractive summarization that best resembles the user’s input. Our goal is to have the algorithm with our results publicly available to anyone interested in seeking unbiased and objective information.

6. References

[1] “Truth Decay: Fighting for Facts and Analysis.” RAND Corporation, RAND Organization.

[2] Arkajyoti Misra and Sanjib Basak. 2017. Political Bias Analysis, 2017.

[3] Vu, Minh. “Political News Bias Detection Using Machine Learning.” Pdfs.semanticscholar.org, 2019.

[4] Ghaoui, Laurent El. “Unsupervised Extractive Summarization: A Comparative Study.” Medium, Towards Data Science, 10 June 2019.

[5] Thompson, Andrew. “All the News.” Kaggle, 20 Aug. 2017

[6] Sharif, Pariza. “BBC News Summary.” Kaggle, 6 May 2018