Deep Learning for the Summarization of News Articles:

Reducing Bias and Misinformation

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

Abstract

With an effort to mitigate the bias and misinformation in news articles, we confront the discernible subjectivity in news platforms by developing a news filtering model that summarizes and maintains the valuable content of published material. The model prompts users to input a news topic, and in return, they receive a paragraph summary of content related to the given topic. The information outputted to the user is obtained as a result of the application of clustering and extractive summarization techniques. Our model works as follows: first we find the news topic that is relevant and of closest similarity to the user's input by clustering news articles with the Universal Encoder algorithm. Next, we iteratively summarize these clusters with a BERT-based extractive model which outputs the final summary. We then evaluate the accuracy of the output both qualitatively and quantitatively by seeking readability and measuring the loss of semantics during our overall summarization process. These evaluations show encouraging qualitative results, as the outputs successfully depict a relevant news summary from the user's input. In this paper, we also present the preliminary results of the second phase of this project: the implementation of bias detection in news sources to yield unbiased summaries to users. Future work includes full development of a fine-tuned model with current news data, and a robust algorithm to detect bias from the available news sources. Our results can be reproduced via the provided Google Colab notebooks in our 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 analyzes "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]. Thus, in an effort to confront misinformation, amongst the public we create a news filtering system that offers its users with the least biased extractive summary of a topic of their choice.

2. Background

With a daily influx of information, keeping updated on the latest news can be difficult and daunting. Major news publishers and smaller organizations dedicated to simplifying the consumption of information have created resources for their readers to digest information more quickly and easily. Small companies such as Newser and the Daily Skimm specialize on curating summaries of the most recent and relevant news for specific categories, while larger publications like BBC and The Economist provide condensed versions of their articles for the day or the past week. The methodology most of these organizations deploy to create their summaries rely on a team of journalists who hand select relevant content. With potential bias from individuals determining the significant information, the summaries that these platforms provide, although convenient for the user, are effort consuming for the content creators and may not be a holistic viewpoint of the events surrounding a news topic.

In response to reducing the effort exerted by journalists, development on automated journalism supports journalists in completing trivial and time consuming tasks. BBC's research and design team implements NLP techniques to "extract higher level, structured information, such as quotes and statistics from documents" to improve the quality of content they provide on their platform, while reducing the time journalists spend digging for material [2]. The Associated Press' use of Automated Insights, a natural language generation software, to create U.S. corporate earnings stories has boosted production of this material by fifteen-fold [3]. Narrative Science, like Automated Insights, is a natural language generation software used to create data-based reports on sports and game results [4]. However, this software has faced criticisms on the potential magnification of bias that already exists in the data [5]. Automated journalism brings about two problems: a much greater velocity of information and potential bias in the outputs. By summarizing information and filtering out bias in the data, our project attempts to leverage the techniques utilized in automated journalism to combat the issues that this field produces.

Creating unbiased summaries of news topics is a challenging task, since the ground truth of what defines bias has yet to be determined. Knowhere, a startup created by data scientists, uses machine learning to create “factual and impartial” articles, covering content from all sides of the political spectrum. Once the articles are generated, human editors review the material to ensure quality content, and their feedback plays into fine tuning the algorithm [6]. Our project aligns very closely with Knowhere’s product: a resource that provides an objective summary of the news. Knowhere offers a daily newsletter and a feed of articles in four categories (politics, business, tech, and world), yet they do not have a search tool to assist users in finding relevant material. Our technology is deployed as a search engine that allows users to immediately access information that they are most interested in across a multitude of topics. Within the architecture of our algorithm, we use state-of-the-art methods such as BERT and the Universal Sentence Encoder, along with preliminary bias filters, to find and present relevant material in a concise and objective way. In the next section, we discuss why we chose those models and how they fit into our product design.

3. Data and Methodology

In the subsections below, we describe the data utilized and the several clustering and summarization techniques applied to obtain the best performing model. We also describe our current approaches for the early stages of bias implementation within our model.

3.1. Data

The dataset we utilize is obtained from Kaggle, notably ”All The News” [7], which contains information such as title, content, and date of publication of over 140,000 articles from 15 major news platforms. We leverage the available computing capacity with 13,000 articles, all of which were acquired by performing a uniform random sample from the available data. The sample of data was later pre-processed for cleaning prior to the application of the clustering and summarization algorithms.

3.2. Clustering

With an effort to identify the most relevant articles to summarize from, we cluster the articles that are of closest similarity to the user’s input. Thus, in this section we elaborate on our two approaches find the top 10 articles that are most relevant to the user’s query:

- Term Frequency-Inverse Document Frequency (TF-IDF) [8].
- Universal Sentence Encoder [9].

In both of these approaches we utilize a combination of title and article content as the input to the algorithms. Additionally, we utilize a cosine similarity matrix measure to determine the article embeddings of closest similarity to the generated embedding of the user’s input. The 10 articles that resulted in scores closest to 1, were denoted as part of the cluster.

TF-IDF: The benefit of using this algorithm is the dismissal of articles that do not contain the terms present in the user’s input, as well as the penalization of the terms that are most common in the article corpus. However, despite its usefulness to quest for the articles of most relevancy, particularly with queries resembling the structure of small numbered n-grams $(n < 4)$, one of its shortcomings is the inability to capture full context and semantics of a more complex user input that could come in the form of a sentence.

Universal Sentence Encoder: We apply this algorithm to our subset data to take advantage of its great performance with text similarity tasks, and to successfully encapture various kinds user queries, as well as the semantics and context in article content. In contrast with TF-IDF, the Universal Sentence Encoder takes into account semantics and word context from each article and user given input. The performance of the encoder greatly exceeds that of TF-IDF, as we later discuss in the results section.

3.3. Summarization

After determining the best mechanism to cluster articles, we proceed to create a single summary from the top 10 scoring articles obtained from our clustering method. Here we elaborate on the two approaches taken to achieve this task.

- Similarity Matrix and Ranking approach [10].
- BERT extractive summarizer, which is a generalization of the Python’s Lecture Summarizer service [11].

To deliver the user with a summary of the top 10 most similar articles in a process consisting of 2 iterations.

- On the first iteration, each algorithm outputs 10 summaries (one for each article).
- On the second iteration of both approaches we concatenate these 10 summaries as a single text block, and use this block as an input to each these summarizing algorithms.

The final output for each of these approaches is a single summary from the 10 articles that were most relevant to the given query. Below we give the reader further details on each algorithm as applicable to our main task.

Similarity Matrix: To output the summary, we take the top 10 articles and split them into sentences to allocate them inside a cosine similarity matrix. Afterward, we generate a rank with the pagerank algorithm to determine the sentences of greater importance. Although the outputted summary satisfied our qualitative expectations, this approach turned out to be computationally expensive. During the ranking calculations within our matrix, some of the articles exceeded memory limits, and thus, were not considered for the generation of the final summary.

BERT Extractive Summarizer: We opted to proceed with our task with an extractive summarizer as it has been shown that this method is successful at collecting key phrases and sentences that best represent text content. This model is pre-trained on literature and is primarily designed for the summarization of academic lectures. However, despite the author's intention we find that article content and written lectures still comprise bodies of text that intend to deliver a message. Furthermore, we proceed to utilize this model as it creates word embeddings with the state-of-the-art BERT algorithm and clusters the embeddings with KMeans to identify the sentences closest to the centroid for the creation of a summary. In contrast to the similarity matrix approach, this model outputted a summary without computational difficulties.

After comparing all approaches in each of the clustering and summarizing mechanisms, we came to the conclusion that using the Universal Sentence Encoder along with the BERT Extractive Summarizer resulted in the best algorithmic sequence to achieve or primary task. We elaborate on the evaluation metrics for the decision of our model components in the Results and Discussion section.

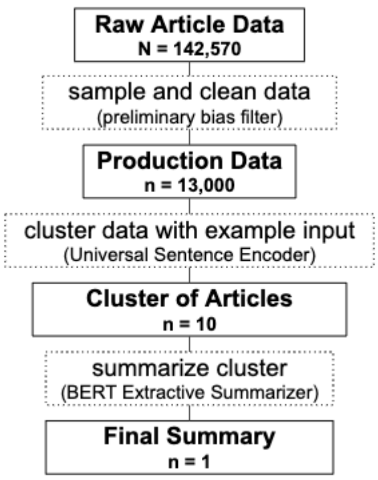

Figure 1: General workflow of our model.

3.4. Bias

It is our goal to be able to produce objective summaries for our users. We attempt to reach some level of bias detection and filtering in this project by defining sentiment as a proxy for bias and using a crowd-sourced political bias rating of publications to guide the removal of potentially biased articles from our data before sampling for production.

Publisher Bias: AllSides is a news platform that categorizes existing articles by the articles' publications' known political bias, to provide their readers with news from "all sides" [12]. The political bias rating of these publications are voted on by thousands of users to confirm the categorization on the platform: left, left-center, center, right-center, and right. Since our dataset contained articles from publications that were rated as left and right leaning, we decided to remove the articles from those publications entirely as our first layer of bias filtering.

Sentiment Analysis: As a second layer in our preliminary bias filter, we set thresholds on the polarity and subjectivity of each article. As our first approach, TextBlob is used to analyze the sentiment of our data for the computational efficiency it offers. Neural network methods that have been developed to more accurately determine sentiment would be much more computationally expensive than deploying a sentiment lexicon to quickly return results to our end users, as we hope to grow this project into a web application. The distribution of polarity and sentiment among all the publications in our dataset are presented in Figure 2.

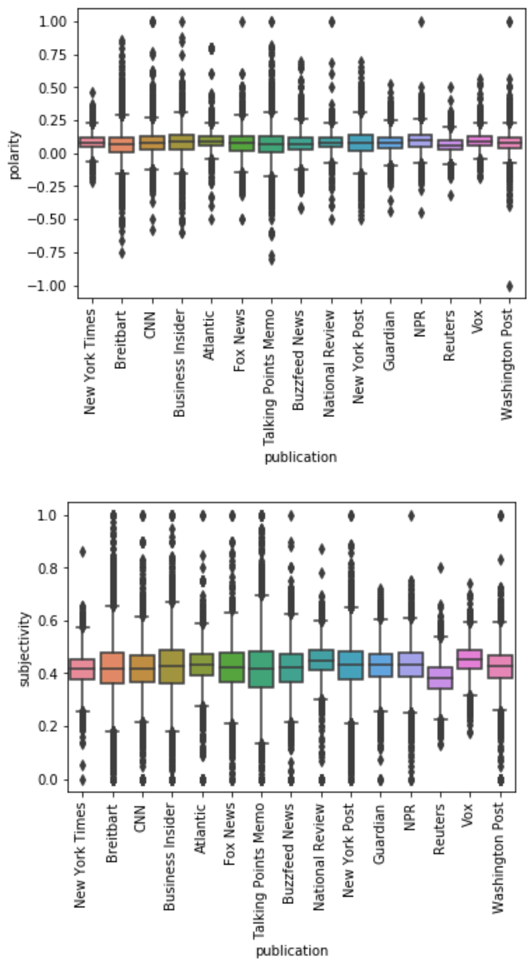

Figure 2: Distribution of sentiment and polarity.

Polarity ranges from -1 to 1, to represent negative and positive sentiments, respectively. Looking at the distributions and outliers, we decided to set thresholds at -0.5 and 0.5, so that the articles we put into production had a more neutral sentiment. Subjectivity ranges from 0 to 1, where 1 is subjective and 0 is objective. We decided to set a threshold for subjectivity at 0.75, since most of the outliers for high subjectivity began at the 0.75 level. Thus, any articles that were below a 0.75 subjectivity level and within -0.5 and 0.5 for polarity were kept to sample into production.

Balanced Coverage: As a final layer to our bias filter, we considered the frequency of certain topics among publications to help filter out news sources that may over report certain events, while keeping the publications that have more balanced coverage of the news. This aspect was not implemented into our system due to the variability among publishers with respect to the proportion of their articles that reference a specific topic, as displayed in Figure 3; it was difficult to isolate certain publishers that had a consistent imbalance in their coverage. Additionally, most publishers that do show a high proportion of topic coverage are removed in the first layer that penalizes political bias.

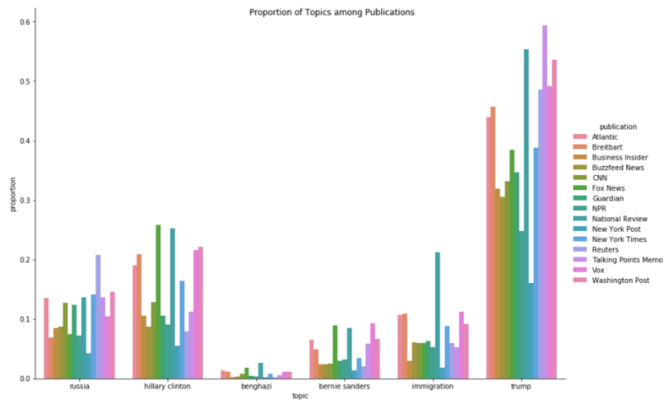

Figure 3: Proportion of topic coverage by publisher.

We aim to have a system that detects and filters bias for an objective summary of a user-given topic. The current methodology in place is an "early-stage" approach to this goal.

4. Results and Discussion

In the subsections below, we describe the evaluation and results for the clustering and summarization techniques applied, as well as a preliminary evaluation for the bias filter.

4.1. Clustering

To determine the best clustering method, we evaluated the clusters from the TF-IDF model and the Universal Sentence Encoder model against researcher annotations. We randomly sampled 200 observations from the data and defined five consistent news topics in the sample. Considering both the article title and content, we then labeled each article in the sample with which topic the article fell under. For articles that did not correspond to any of our five defined topics, we set the label to zero. These annotations effectively acted as cluster labels for the articles in the sample.

For each of the models, we created word embeddings of the 10 topics, and for the title and content of every article in the annotated data. We then clustered the topics with the articles that had the highest cosine-similarity, resulting in the model's predicted clusters for each of the topics. A correct prediction occurs when an article is clustered to the same topic as its annotation. For example, if an article that was labeled "2", corresponding to the second topic, ends up in the predicted cluster for the second topic, then the model made a correct prediction. As a measure of precision, we calculated the percent of correct predictions out of all predictions. Additionally, as a measure of recall, we calculated the percent of correct predictions out of all the original articles that were annotated to the corresponding cluster topic; in other words, we determine how many of the articles we annotated were assigned to the correct cluster. The results for each of the models are presented in Figure 4.

TF-IDF		Universal Sentence Encoder	
Precision	Recall	Precision	Recall
0.28	0.41	0.42	0.66

Figure 4: Evaluation results for tf-idf and Universal Sentence Encoder.

The Universal Sentence Encoder model significantly outperformed the TF-IDF model in both precision and recall. To determine the kind of text corpus that maximized the performance of the encoder, we implemented the same evaluation methodology with word embeddings but applied to either the article title, the article content, or both the article title and content. The results for each of the corpora are displayed below in Figure 5.

Title		Content		Title + Content	
Precision	Recall	Precision	Recall	Precision	Recall
0.32	0.55	0.42	0.66	0.42	0.66

Figure 5: Evaluation results for corpora on Universal Sentence Encoder.

The encoder model that used a combination of title and content performed just as well as the encoder model that used only the article content. Saving the computational cost of running on longer texts, we concluded that encoding just the article content with the Universal Sentence Encoder was the best method for our clustering model. Please refer to Figure 8 in the Appendix for a workflow diagram of this evaluation.

4.2. Summarization

To determine the best summarization method, we evaluated the summaries produced by the similarity matrix model and the BERT extractive summarizer model, for retention of the semantics of the original articles used to create the summaries. This was operationalized by clustering the summaries with the original training data to determine if the articles in these new clusters are the same as the articles in the original clusters.

To deconstruct our evaluation, let's define "cluster A" as the set of top ten articles most relevant to a user given topic. For each model, we summarize all the articles in "cluster A" for a single, final summary which we would output to the user. We then run the Universal Sentence Encoder on this summary and find the top ten articles that have the highest cosine similarity, resulting in "cluster B." To evaluate the model, we compare "cluster A" to "cluster B" and count how many articles in "cluster A" are also in "cluster B." The evaluation metric measures the percentage of articles from "cluster A" that appear in "cluster B" out of all articles in "cluster B" to proxy the semantic retention of the final summary to the articles that were used to make the summary ("cluster A"). The results for each of the models are displayed below in Figure 6.

Similarity Matrix	BERT Summarization Model
0.38	0.46

Figure 6: Evaluation results for similarity matrix and BERT extractive summarizer.

Retention of the semantics of the original articles is much stronger with the BERT extractive summarizer. Additionally, looking at the content of the summaries, the BERT summarizer also performs better qualitatively, with more relevant and comprehensive material. Thus, we decided to implement the BERT summarizer as our summarization model. Please refer to Figure 9 in the Appendix for a workflow diagram of this evaluation.

Example outputs of our model summaries can be found in the Appendix.

4.3. Bias

To evaluate the preliminary methods we have implemented for filtering potentially biased articles in our data, we compared the sentiment of multiple summaries produced from an unfiltered dataset (potentially biased) and from a filtered dataset (likely unbiased) with example user inputs. While we acknowledge the minor redundancy in our evaluation, we believe that the addition of filtering on news publications based on the AllSides rating added another layer that distinguishes our procedure from the evaluation. The results for each of the datasets are produced below in Figure 7.

Unfiltered Dataset		Filtered Dataset	
Polarity	Subjectivity	Polarity	Subjectivity
0.1043	0.4587	0.0868	0.4358

Figure 7: Evaluation results for unfiltered and filtered datasets.

The summaries from the filtered dataset had a lower average polarity and subjectivity, meaning that these summaries were more neutral and objective than the summaries from the unfiltered dataset. Thus, we decide to keep our preliminary implementation of bias detection and filtering within our project design. Please refer to Figure 10 in the Appendix for a workflow diagram of this evaluation.

5. Conclusion

Through the application of the Universal Sentence Encoder and the BERT Extractive Summarizer, we have developed a model that takes in an arbitrary queried news topic, and in turn, delivers a summary that best aligns with the given input. As an additional feature, we include the early-stages of bias implementation in our model for unbiased and more informative summaries. Please refer to Figure 11 in the Appendix for summaries from our model of example user inputs that vary across multiple news categories.

6. Next Steps

To improve our entire model, our future work involves the domain fine tuning of the BERT summarization model with news article data that offers users with ground truth, such as the BBC News dataset in Kaggle. Additionally, to make our sumamarization iterative process more effective, we seek to implement trigram blocking to minimize sentence repetition in the final outputted summary. Lastly, we seek to develop a full implementation of bias detection in news articles. Ideally, we would evaluate the bias in our outputs through annotated training data. However, developing a ground truth for bias is an ongoing challenge in the field of natural language processing. Our use of sentiment as a proxy for bias is admittedly naive, yet it remains appropriate for the scope of the preliminary methods deployed in this project with respect to bias. Thus, future work to build upon the bias filter and its evaluation include other approaches taken by academics, such as the usage of annotated and "fact-checking" data, bias detection with BERT and LSTM models, as well as the qualitative evaluation of these approaches.

A Note on Performance As mentioned previously, we reduced the size of our production data for our model to 13,000 uniformly randomly selected observations in order to accommodate for memory restrictions hosted our cloud servers. This may have impacted the accuracy and strength of our clusters and summarizations. We plan to overcome this by expanding the size and memory capacity of our clusters, as well as the allocation of GPUs in our computing system.

7. Appendix

Reproducibility All our notebooks are also hosted in Google Colab as an effort to keep our work accessible to users. The work of this project is avaialable in on GitHub: `www.github.com/s-miramontes/News_Filter`. All generated datasets available upon request.

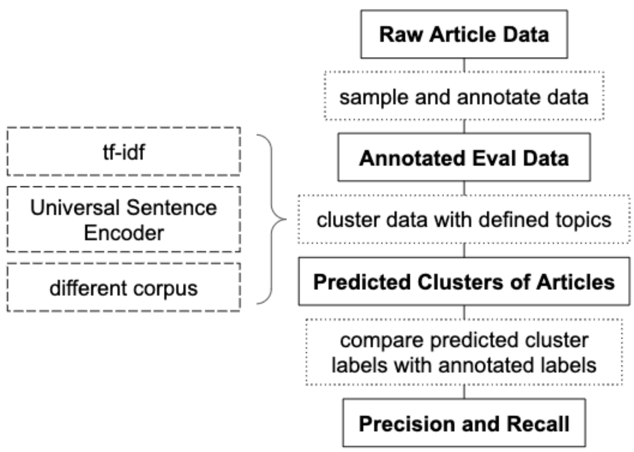

Figure 8: Workflow of clustering evaluation.

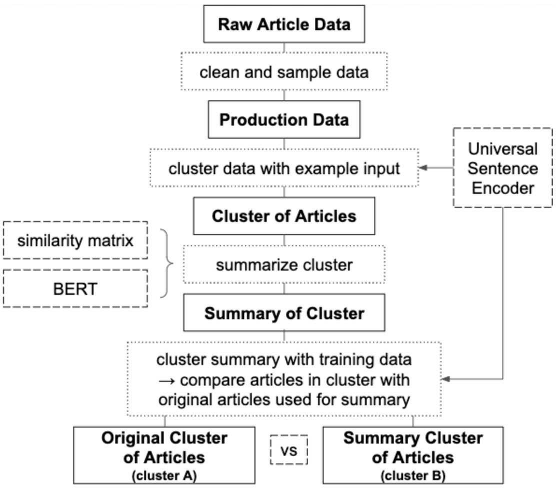

Figure 9: Workflow of summarization evaluation.

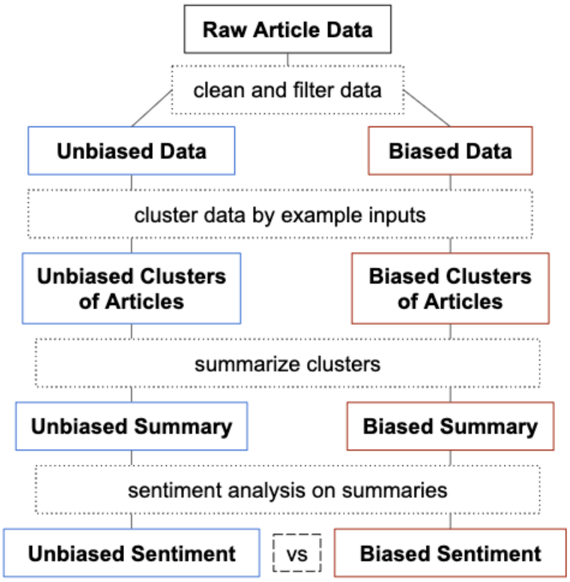

Figure 10: Workflow of bias evaluation.

8. References

[1] "Truth Decay: Fighting for Facts and Analysis." RAND Corporation, RAND Organization.
[2] "Natural Language Processing." BBC R&D, BBC.
[3] Colford, Paul. "A leap forward in quarterly earnings stories". The Definitive Source. The Associated Press.
[4] "Program Creates Computer-Generated Sports Stories". NPR. NPR. Retrieved 6 June 2014.
[5] Hefferman, Margaret. "The post was written by a human". CBS News. CBS. Retrieved 6 June 2014.
[6] "Knowhere News." Knowhere News.
[7] Thompson, Andrew. "All the News." Kaggle, 20 Aug. 2017
[8] Sparck Jones, K. (1972), "A Statistical Interpretation of Term Specificity and its Application in Retrival", Journal of Documentation, Vol. 28 No. 1, pp. 11-21.
[9] Daniel, et al. "Universal Sentence Encoder." ArXiv.org, 12 Apr. 2018.
[10] Dubey, Praveen. "Understand Text Summarization and Create Your Own Summarizer in Python." Medium, Towards Data Science, 23 Dec. 2018.
[11] Miller, and Derek. "Leveraging BERT for Extractive Text Summarization on Lectures." ArXiv.org, 7 June 2019.
[12] "AllSides Media Bias Ratings." AllSides.

"Russian interference with election"	THE MASSIVE leak of documents from the campaign of Emmanuel Macron failed to prevent his landslide victory Sunday in the French presidential election. An appropriate response must begin with full investigation and disclosure. We call on President Putin to immediately order a halt to this activity. Intelligence Committee members receive classified briefings they can't speak about in public. WASHINGTON — A day after the release of a damning intelligence report on Russia's efforts to influence the American election, Donald J. Trump called on Saturday for a closer relationship between the two nations, saying only "stupid" people or "fools" would think this was unwise. Ukrainian computer specialists mobilized to restore operations in time for the elections. Says electors potentially undermining democracy pic. Toward the beginning of the hearing, Comey said that he has no doubt that Russia attempted to interfere in the 2016 election and that Russian government officials were aware of the meddling. And it was an active measures campaign driven from the top of the government. ' " A declassified version of a on Russian hacking has concluded that Russian President Vladimir Putin ordered a campaign to influence the US presidential election with the aim of hurting Hillary Clinton' s chance of winning. " ' Multiple officials told the Post that if proven, the Russian attempts at sabotage may not be intended to swing the election any particular way, but rather to generally promote chaos and mistrust that could make the US seem less stable and trustworthy in international affairs, potentially diminishing its authorities.'
"Immigration and customs enforcement"	Reports of immigration sweeps this week sparked concern among immigration advocates and families, coming on the heels of President Donald Trump's executive order barring refugees and immigrants from seven nations. But in reality they're going after anyone they can get their hands on — period. The language in the order says that any unauthorized immigrant convicted of any crime can be deported. This was used most pervasively, and attracted the most attention nationally, in Phoenix. Apprehensions are down: Despite the growing number of officers patrolling the border, fewer crossers are being caught. Criminal deportation because of or something else? Whatever the reason, Garcia de Rayos, 35, said she has no regrets. I think this is a direct result of the new executive orders that are being put into actions by President Trump calling them ' enhancing public safety,' which really appears only to be attacking immigrant communities and people of color," her attorney Ray Maldonado said. She came illegally to the United States in the with her parents when she was 14. After her conviction she appealed a court order to voluntarily deport and lost. He claims to have found a weapon, which discharged randomly and killed Steinle. You'll never convince me that it hasn't been about the way that we've handled this issue." In a joint letter, more than 60 law enforcement heads are appealing to Trump in all but name to soften his aggressive drive to enlist police officers in the highly contentious job of deporting millions of immigrants living without permission in the country.'
"Ariana Grande Manchester bombing"	Manchester, England (CNN) Monday's attack outside an Ariana Grande concert at Manchester Arena killed at least 22 people, including children, and was carried out by a lone suspect carrying a bomb, Manchester Police said. " The incident happened shortly after Grande had left the stage, according to eyewitnesses. Eyewitness Karen Ford had taken her daughter to the concert. Taxis and local people offered free rides to those affected. Initially, said they thought popping balloons had set off a panic. In an emailed statement, Chris Upton, the head teacher at Tarleton Community Primary School, said the learning community was still coming to terms with Saffie's death. " They also wanted to send a message: We are not afraid. [Keavy Smith, 17, was there that fateful night and remembers calling her mother, Angie, screaming and crying after the bomb went off. Almost everyone seemed to be holding someone else's hand. Authorities have identified the bomber as Salman Abedi, a Briton of Libyan descent, who lived in Manchester.'
"UC Berkeley student protests"	They gathered near the White House, disheartened and dismayed. " In Austin, protesters blocked a highway Wednesday afternoon. The voted for Hillary Clinton and was disappointed. In Des Moines, Iowa it was hundred of high school students who left class to protest of election results. He doesn't respect women, Black Lives Matter, Latinos. In New York on Wednesday morning, groups of Trump supporters cheered his victory outside Trump Tower. Furious at a lecture organized on campus, demonstrators wearing outfits smashed windows, threw rocks at the police and stormed a building. He travels around California on a tour bus airbrushed with his likeness. There are racists, sexists, piggery of various kinds who will say really terrible things. In an interview Friday, Peter Sittler, a sophomore at Berkeley and vice president of the organization that sponsored Yiannopoulos's visit, the Berkeley College Republicans, told me the school's administration, from Chancellor Nicholas Dirks on down, "worked tirelessly to plan [the event] and make sure it went through. " The is an movement widely seen as espousing white supremacist and views. The University of California at Berkeley is bracing for massive protests and potential violence in an open, public space known as Sproul Plaza after learning that conservative commentator Ann Coulter plans to give a speech there Thursday afternoon. [CNN) The showdown continues between the University of California, Berkeley, and Ann Coulter over whether the conservative firebrand will speak at the famously progressive school.'
"Suicide Squad movie"	Los Angeles (CNN)"Suicide Squad" once seemed to offer Warner Bros. and DC Entertainment a shot at redemption after the creatively botched "Batman v Superman: Dawn of Justice." Yet this tale hits its own snags, generating only sporadic moments of fun amid chaos. It's a concept with roots in "The Dirty Dozen" (down to the promise of shortened or commuted sentences) although to be honest, this Hellacious wouldn't stand much of a chance against the Man of Steel. Frankly, they look mismatched even against the underwhelming threat they face in the movie, the Enchantress (Cara Delevingne) an ancient sorceress with a hazily defined scheme to destroy the world. Suicide Squad" does yield moments of irreverence, though. The Lobster, A stunning exercise in dystopian absurdism and comedy. In 2004, Catwoman was awarded a GLAAD media award for its portrayal of the character Holly Robinson as openly gay. Batman v Superman: Dawn of Rubbish, This is the challenge that Suicide Squad sets for itself early, and it succeeds just about as poorly as you might imagine. That it took DC so long to figure out what Marvel has known from the start is no small amazement.)'

Figure 11: Summaries of example user inputs.