Detecting Political Tendencies of Major News Media

Alexandra Li
CS & PPL
Emory University
alexandra.li@emory.edu

Yunjie Wu
CS & QTM-Psychology
Emory University
ywu449@emory.edu

Eula Wang
Undecided
Emory University
jingyu.wang@emory.edu

Abstract

Major studies in the field of news media analysis focus mostly on the categorization of media in terms of their political wings and little on differentiating their views on specific political topics. Our project seeks to examine media perspectives by using topic extraction with LDA to target more specific topics and conducting opinion mining via sentiment analysis. We attempt to present the public with a more authentic and clear view in regards to the media's political tendencies so that they can be more informed about their own bias toward major new outlets and also gain a more objective view regarding what to expect when reading news articles. Among the political topics examined, our study finds that the media generally hold negative views toward topics under coronavirus with different negativity, and have more various opinions toward topics under immigration but also choose different topics to cover. We also observe that different media hold notably more negativity toward some topics and also share their degree of negativity across different subtopics. Our result offers the public an intuitive way to understand media perspectives regarding specific political topics so that they may interpret information presented in news with a more objective attitude.

Keywords—Political Tendency, News Media Analysis, Topic Extraction, Sentiment Analysis, Opinion Mining

1 Introduction

Due to the nature of news media in current society being a mixture of subjectivity and objectivity, news outlets' reputations often precede their actual views presented in news articles, especially when it comes to political topics. It is common knowledge that most news media have their perspectives shaped by multiple factors, from government agenda to political parties to journalists' own subjectivity. Even in democratic states where media are less censored, they are far from objectivity and the public is also well aware that media are biased. Therefore, it is natural for people to have certain expectations regarding the content of an article even without reading it but only by knowing the media the article is associated with. In other words, many readers start an article with presumptions about the journalist's opinion which largely shape how they interpret the information given by the article.

Existing studies that center on labeling news outlets with left, right, and central wings and also how "factual" they are no doubt strengthen those presumptions, not only leading to the creation of stereotypes for each news media but also render the public to believe that they understand media bias. These previous studies also fail to specify the labels in terms of the specific topic to which they may apply. This simply implies that if a media is labeled as "right", it will always embrace an opinion typical of right-wing regardless of topic.

It is exactly this mindset of seemingly knowing and over-generalization of labels that generates new bias. People are now confident that they are being relatively objective, believing that they have taken media bias into consideration when trying to discern factual information. Additionally, since they "understand" a news outlet's perspective, they can avoid being manipulated by applying that understanding to every topic. For instance, when someone finds out that an article is from Fox News and is about mask regulations, one will automatically expect strong opposition to policies that require masks from the author. Even if the article takes a relatively neutral position on the issue, it is more likely for people to stick to their original belief by unconsciously not wanting to notice the true perspective of the journalist so that their original assumption can be confirmed. Instances like this demonstrate how the actual information provided can be overlooked or twisted in ways to fit people's presuppositions, all caused by vague ideas about

each news outlet's political position and unfounded connections between labels of political inclination and media perspectives on specific political topics.

Although it is unlikely to completely eliminate this bias of "I already know their biases" or its over-generalization, it is possible to somewhat reduce it by presenting the public with media outlets' actual perspectives on specific political topics(i.e., immigration, healthcare). Unlike most existing studies that focus on categorizing media, our project emphasizes the examination of political views themselves. We categorize news articles according to the specific political topic they belong to, which we refer to as a general topic, and then discover smaller and more specific topics, which we refer to as subtopics, under each general topic using a bottom-up approach. By understanding media views on a more subtle level, we try to tackle the problem of over-generalization.

We hope our study can offer people insights into what media opinions on familiar topics really are so that they are better equipped with the information required to know what they should expect when reading a news article. They will also be able to compare the results of our study to "common knowledge" regarding media bias and decide for themselves which part of it is true by understanding their own bias toward those media first. In this way, we intend to help people to embrace a more objective attitude before approaching information presented by the media and also boost their capabilities to extract useful and factual information from news in general.

Throughout our study, we continue to improve our methods to be better adapted for information extraction in terms of specific topics and also more precise tendency detection. We believe that our approach allows for the effective connection between topic extraction and opinion mining that may benefit future researchers in the field of news analysis who are interested in understanding media opinions on a deeper level.

2 Related Work

Because we intended to establish a connection between media and their views on specific political topics, we found ourselves with mainly two tasks: finding the topics that news articles evolve around, and examining media perspectives on those topics. One work that we came across during the research phase helped us understand how to perform these two tasks. It presented a model that combines the use of LDA and sentiment analysis, and proposed a theory of the "personalities" of news media—that a media outlet has a predictable attitude towards each topic (Doumit and Minai, 2012). Their models represented the views of each news media with their "meme-synergies", which are assumed to be latent structures in news media that cause a certain media to react to certain topics in certain ways (Doumit and Minai, 2012).We decided to base our model on this theory—-using the LDA for topic extraction and sentiment analysis to generate sentiment scores for examination of news media perspectives.

Though our overall direction has been determined, we experienced difficulty trying to connect the two tasks. We first attempted to locate key sentences given by extracted subtopics and then perform sentiment analysis on those sentences (since most sentiment analysis models work best on sentence level), but most previous studies utilized event extraction instead of LDA to identify key sentences. To be more specific, event extraction is when we try to group sentences similar to each other and then find key words which will be extracted as events (Julinda et al., 2014). It was clear that if we use this approach we would be going in circles by trying to go from key words to relevant sentences but ended up with keywords. This realization prompted us to realize that it was impractical to attempt finding key sentences when what we want is to have explicit political topics with corresponding media views, so we should instead prioritize topic extraction.

We therefore made several attempts throughout the course of our research to connect extracted topics with sentiment analysis. First, we tried to perform sentiment analysis on the whole article and then matched the article with the topic that carries the most weight and assigned the generated sentiment score to that topic. However, this deviated from our envision of viewing articles as a combination of subtopics, not one which ensues over-generalization that we wish to avoid. We then attempted to calculate the weight percentage of each topic for a sentence and then assign the sentiment score of that sentence according to those weights to different subtopics. This again could cause issues since the sentiment of a sentence does not necessarily have to be distributed evenly across all words and keywords are also not regarded as subjects of that sentiment.

Our breakthrough happened during our research for the sentiment model. The original model we used for sentiment analysis is VADER (Hutto and Gilbert, 2014), which performs sentiment analysis on sentence level so that we can only distinguish keywords from other words by giving them more weight during the allocation of sentiment score. The new sentiment model we use (Yang et al., 2021) is capable of calculating the sentiment score for a word in the context of the sentence that word belongs to, meaning that we can directly obtain the sentiment of keywords within a sentence. The detailed process of how we obtain the ultimate sentiment score for each topic will be illustrated in the approach section.

3 Approach

Our approach consists mainly of two parts: topic extractions and applying sentiment analysis to demonstrate media opinions. For the first part, we run LDA with Gibbs sampling with all articles under the same general new topic from our dataset(i.e., coronavirus) to extract several key subtopics. For the second part, we run sentiment analysis on generated keywords under the sentence's context and eventually by combining weights of keywords with sentiment result of keywords we obtain a sentiment vector for each subtopic under each general topic for the news outlet of our choice.

3.1 Topic Extraction with LDA

3.1.1 LDA with Gibbs Sampling

LDA(Latent Dirichlet Allocation) is a probabilistic topic-modeling tool used for the extraction of latent topics from news feed data (Blei et al., 2003). It utilizes unsupervised learning through training and generates a certain number of subtopics. Each of these subtopics consists of keywords with the most weight among all keywords included in that subtopic, and both the number of subtopics and the number of most weighted keywords will be determined manually before the training.

In order to find useful and meaningful subtopics under each general topic, we use a specific implementation of LDA–the usage of Gibbs sampling (Zhikai, 2016). Lda with Gibbs Sampling is an algorithm used for successively sampling conditional distributions of variables (Johansen, 2010). It iteratively draws an instance from the distribution of each variable, conditional on the current values of the other variables in order to estimate complex joint distributions.

After changing the code to suit our purposes and assigning values to necessary parameters for the LDA, we run our data to the model and have it return a list of subtopics in the form of keywords with their corresponding significance(weight). However, in order to render the result capable of representing some meanings so as to attain more clarity when our final results are interpreted(it is necessary for the meaning of the subtopic to be easily understandable), we need to perform the "filtering" step.

3.1.2 Filters

Each subtopic consists of some keywords that contribute nothing to its comprehension and those words are the target of our filters. Basically, the filters we design will filter out words that impede us from extracting useful information in order to have a keyword list with words related enough but without being synonyms. We roughly divided these "unwanted" keywords into three categories:

- Unrelated(i.e., people, America)
- Somewhat related(i.e., president, government)
- General topic related(i.e., immigration, healthcare)

Words that belong to the first category are likely to appear in any news so that they add no new meaning to the subtopic they are in. We filter out all words that belong to the first category by running the LDA first, identify keywords that are unrelated, and filter them out before running the LDA again. Somewhat related keywords refers to those that are relevant to the general topic but whether they contribute any meaning to the subtopic remains undetermined and requires manual inspection. For words like this, we will judge manually whether they are useful for interpreting the keyword list, and we again filter them out if not before running the LDA to generate a new subtopic list. The last category includes words that are directly related to the general topic, usually synonyms of words that constitute the general topic themselves. Whether these words are necessary for the understanding of subtopics is decided by comparing the subtopic before and after their filter, and corresponding changes are made based on the observation of those comparisons.

Besides the filtering method itself, it is also essential for us to demonstrate the efficacy of these filters. The exact measurement will be illustrated

in more detail in the experiment part but we follow two rules when it comes to the evaluation of subtopics generated. The first rule is that keywords from different subtopics should not have too many overlaps, because that indicates less specificity in regards to each subtopic. The second rule we follow requires us to avoid only leaving extremely specific words, such as filtering out all words that belong to the "somewhat related" category. Filtering out too many seemingly unrelated keywords will only result in missing subjects or leaving words that are unrelated to each other thus incapable of forming any meaning. This implies that some overlap between different subtopics is necessary, since we should avoid filtering out some keywords that belong to the somewhat related category. All these factors render us to draw the conclusion that we should keep track of the percentage of overlap throughout the filtering process to see if our filters really help us to obtain more meaningful subtopics.

3.1.3 Parameters for LDA

As mentioned in 3.1.1, there are two parameters for running LDA that we will have to decide manually—-the number of keywords in each subtopic N and the number of subtopics K. Both of them play important roles in the interpretability of generated subtopics so we also need ways to judge if the parameters we choose are the most suitable ones. For N, we follow the idea that for those keywords with lesser weights, which will appear at the end of the keyword list, we check if deleting those words will affect the comprehension of the subtopic. If the meaning of the subtopic is influenced by that word then in most cases we keep it. In this way, we try to find an N that can enable the subtopic list to include the least number of keywords while adding more will add little meaning.

Since we hope to generate keywords that are related enough but also capable of conveying certain meanings, the value for K will be determined by comparing the result of different K in terms of the percentage of keywords that contribute to the subtopic's meaning(we calculate one percentage score for each subtopic). We then examine in which interval those percentages fall and adjust K accordingly. For a suitable K, we expect most percentages to be high and few to none percentages with an extremely low value. More detailed evaluation methods will be given in the experiment section.

3.2 Sentiment Analysis

3.2.1 Sentiment Model

To evaluate the sentiment of each subtopic, we apply an Aspect-based Sentiment Classification model (Yang et al., 2021). This model adopts the local sentiment aggregating (LSA) mechanism instead of existing dependency tree-based models, making it less expensive and therefore more affordable for our project. It is also trained with more than thirty thousand ABSA samples and outperforms the state-of-the-art on four common data sets that consist of various topics:

1. Laptop14 from SemEval-2014 Task4 (Pontiki et al., 2014)
2. Restaurant14 from SemEval-2014 Task4 (Pontiki et al., 2014)
3. Restaurant15 from SemEval-2015 task12 (Pontiki et al., 2015)
4. Restaurant16 from SemEval-2016 task5 (Pontiki et al., 2016)

We test the model further with some articles from our own data set and found its output to match that of our subjective judgment.

The model requires a sentence and a keyword as its input, and outputs a 3-dimensional vector (probability of negative, neutral, and positive respectively) that represents the sentiment towards a word under the context of that specific sentence. The input follows the required format where [CLS] indicates the start of the sentence, and [SEP] is wrapped around the target word. As shown in Figure 1, this model is capable of accurately detecting the sentiment of different words in the same sentence.

3.2.2 From Subtopic to Ultimate Sentiment Vector

In order to calculate the sentiment of an article regarding a specific subtopic, we first calculate the average sentiment vector for all keywords in this subtopic using algorithm 1.

Next, we calculate the sentiment vector for article A of subtopic t using this formula:

$$S(A,t) = \sum_{i=0}^{N-1} k_i * w_i$$

where k_i is the average sentiment for the k^{th} keyword in subtopic t in article A, and w_i is the k^{th} keyword's corresponding weight.

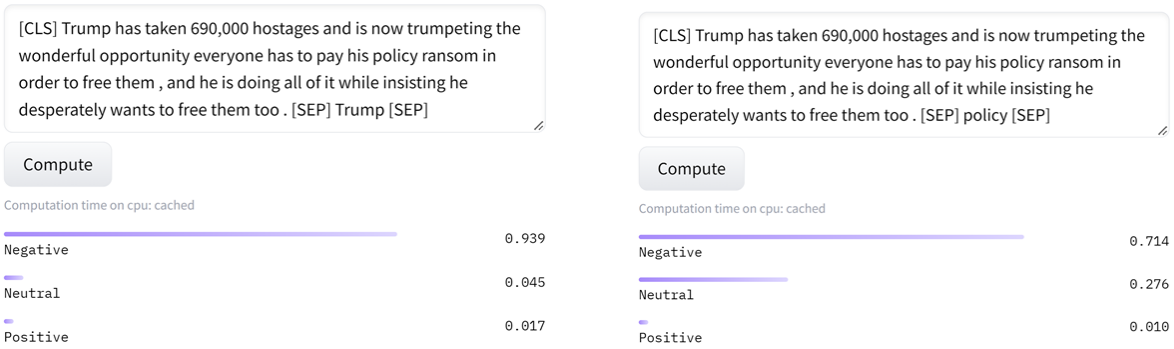

Figure 1: Demonstration of Sentiment Model

Algorithm 1 Average sentiment vector for keywords

1: **for** $keyword \in subtopic$ **do**
2: **for** $sentence \in article$ **do**
3: **if** $keyword \in sentence$ **then**
4: Calculate the sentiment of $keyword$ under the context for $sentence$
5: **end if**
6: **end for**
7: Calculate the average sentiment vector for $keyword$ in $article$
8: **end for**

For the last step, we retrieve all articles that belong to the same media and of the same general topic and calculate the average sentiment vector score for each subtopic by taking the average of all scores of that subtopic from retrieved articles.

4 Experiments

4.1 Data Set

4.1.1 Original Dataset

We use a data set named AllSides, which contains a total of 34,737 articles from 73 news outlets, and the articles are classified into 109 topics (Baly et al., 2020). Because AllSides pulls articles from various topics and media with different political tendencies(for example articles under the same topic will include media from all political wings), we consider our data set toe be well balanced in terms of their content and ideology. Our research focuses on the study of two political topics—Immigration and Coronavirus—-given the fact that they both have relatively larger numbers of articles and also are topics we are most familiar with (Table 1).

The data is in JSON format, with each file containing information other than the content itself,

	immigration	coronavirus
# Article	1547	1079
# Source	96	118

Table 1: Data statistics

such as the news' source(news outlet) and author name(Figure 2). Each news article has been annotated in regards to its political position(which is classified into left, neutral, and right). However, we do not plan to use this part of the data because we try to avoid labeling and generating more bias due to "common knowledge"on media bias. We will instead take use of the general topic, source, and article content, but hope to incorporate other information such as title, publish data, and author for future work.


```
{
    "topic": "immigration",
    "source": "Washington Times",
    "bias": 2,
    "url": "http://www.washingtontimes.com/news/2013/may/24/blocking-the
    "title": "PAUL: Blocking the pathway to a national ID",
    "date": "2013-05-24",
    "authors": "Sen. Rand Paul",
    "content": "The controversial immigration-reform bill that passed th
    "content_original": "ANALYSIS/OPINION:\n\nThe controversial immigrat
    "source_url": "www.washingtontimes.com",
    "bias_text": "right",
    "ID": "00JPgT4b72LdNFwS"
}
```


Figure 2: Example Entry

4.1.2 Annotated Data

For each of the 2 topics(Immigration and Coronavirus), we annotated 50 articles, 100 in total. Our annotation is based on the interpretation of subtopics generated by LDA for each general topic. For each article, we classify its sentiment towards each subtopic as negative, neutral, or positive, represented as "-1", "0", and "1". If we do not consider there to be any related information about the subtopic in the article, we label that as "X" (Figure 3). For the articles we annotated, we only have

access to the content and title of the article, with no knowledge on the source to avoid biases.

In our preliminary stage, we each annotated the same ten articles, and compared our results. We found that our results did not align very well, so we decided to develop an annotation guideline. We first need to agree on the distinction of neutral/positive and neutral/negative. Therefore, we found some sentence examples for each of the categories to establish a standard. Secondly, we realize that during our initial annotation phase, we fail to lay enough emphasis on keywords themselves but reinterpret our interpretation to form a statement. During the transformation from keyword list to subtopic interpretation, we should try to avoid creating an event, like "whether the government supports reopening of business," but to have it still remain a topic like "reopening of business." However, we initially failed to do this and often view an interpreted topic as a statement, causing instances like assigning sentiments to a subtopic that has never appeared only because relevant content has been mentioned. After combining these realizations and establishing an agreement to keep in mind the importance of not over-interpret, we annotated the rest of the articles independently.


```
{
    "topic": "Guatemalan boy detained at border dies in US custody",
    "date": "",
    "content": "An eight-year-old boy from Guatemala has died in US government custod
    "ID": "iBN9du6P7vurEAFi",
    "sub_0": "-1",
    "sub_1": "X",
    "sub_2": "X",
    "sub_3": "X",
    "sub_4": "X",
    "sub_5": "-1",
    "sub_6": "-1",
    "sub_7": "X",
    "sub_8": "X",
    "sub_9": "X"
},
```


Figure 3: Example Annotation Format

4.2 Subtopics

4.2.1 Original Subtopics and Interpretation

Table 2 and 3 show the list of generated subtopics of general topic "Coronavirus" and "Immigration" respectively, both with 10 subtopics and 10 keywords for each subtopic after we determined that K=10 and N=10 give us the most desirable result. Keywords for each subtopic are placed in descending order in regards to their weights so words on the left are more important during interpretation. Based on the two original lists and this rule of "left significance, " we obtained in Table 4 and 5 two lists of interpretation of subtopics, on which our annotation relied. These interpretations will also assist us during our final result phase by demonstrating what specific aspects media views are referring to.

4.2.2 Filter Evaluation

We use these 2 following ways to measure the efficacy of keyword filters:

- Percentage of subtopics generated that have at least another subtopic with at least 1 common keyword
- Percentage of subtopics generated that share more than 1 keywords with at least another subtopic

The first measurement works to show the overlap between subtopics. Since it is both natural and necessary for subtopics that belong to the same general topic to share "somewhat related" keywords(so that some meaning can be extracted), more than half of the subtopics are expected to have other subtopics that share the same keyword. The second measurement indicates whether that overlap has become too much(over-overlap) for the specificity of every subtopic to be presented, so one of our most important missions is to reduce this measurement through the application of filters. Table 6 demonstrates these two kinds of measurement after applying one more type of filter each time in the order of unrelated words, somewhat related words but little contribution, and directly related words. For example, the second column is the measurement after only filtering out unrelated words and the fourth contains those after filtering all three categories of unwanted keywords. We observe a gradual decrease of over-overlapping as shown by the second measurement, which suggests more specific meaning from each subtopic. We also see a certain level of overlap to be maintained throughout the filtering process as manifested by the first measurement, which like mentioned earlier, is an indication of the existence of "somewhat related" words necessary for subtopic's construction of meaning.

4.2.3 Choosing the Best Subtopic Number K

The next parameter we have to decide is the number of subtopics generated K. And we utilize the average percentage of related words contained in a subtopic as well as the distribution of percentage of related words across subtopics to determine if K needs adjustment and how. We observe in Table 7 that when K=8, there is an extremely uneven distribution of related words percentage, which suggests that some subtopics are stuffed with related

0	virus + disease + experts + cdc + data + pandemic + spread + control + percent + social
1	social + media + times + news + company + coronavirus + facebook + video + twitter + online
2	economic + crisis + pandemic + workers + economy + americans + percent + financial + america
3	trump + election + biden + campaign + voters + democratic + voting + sanders + mail + vote
4	coronavirus + testing + covid + vaccine + tests + test + flu + fauci + drug + study
5	trump + coronavirus + house + white + senate + bill + congress + business + democrat + billion
6	stay + reopen + restriction + business + distancing + coronavirus + social + guideline + governors
7	masks + mask + risk + family + contact + person + wearing + wear + children + hand
8	coronavirus + china + virus + outbreak + countries + chinese + trump + spread + wuhan + report
9	covid + patients + medical + hospital + deaths + death + nursing + city + doctors

Table 2: Subtopics under Coronavirus

0	illegal + law + enforcement + legal + deportation + laws + status + criminal + security + policies
1	percent + voter + undocumented + party + bush + support + presidential + election + reform
2	court + judge + executive + government + ban + decision + legal + countries + security + supreme
3	bill + house + senate + republicans + reform + legislation + security + sen + gop
4	obama + congress + house + action + citizen + amnesty + white + executive + program + republi
5	children + families + asylum + parents + child + separated + process + seekers + family + illegally
6	border + mexico + migrant + patrol + security + agent + southern + government + caravan
7	workers + visas + visa + citizens + born + foreign + based + jobs + green + skilled
8	ice + detention + family + agency + custody + officials + center + agents + customs + county
9	trump + wall + daca + democrats + white + government + funding + dreamers + donald + congress

Table 3: Subtopics under Immigration

0	Covid data, CDC
1	Social Media regarding Covid
2	Economic crisis caused by the pandemic in the U.S
3	Covid and American election
4	Covid Testing and Vaccine
5	Government Bill regarding Covid
6	Reopening of business
7	Wearing mask and way to reduce risk of affection
8	Covid outbreak in China
9	Medical care and hospitalization of Covid patients

Table 4: Subtopic Interpretation(Coronavirus)

0	Laws on deportation & illegal immigration
1	Immigration issue during election & presidential campaign
2	Court's involvement in immigration
3	Government bill and reform
4	Obama and government on granting immigrants citizenship & amnesty
5	Family separation & seeking asylum
6	Mexico and border
7	Work visa
8	ICE & immigrant detention
9	Trump on building walls & daca

Table 5: Subtopic Interpretation(Immigration)

Filter words type / Measure	Unrelated	Related but contribute no meaning	Directly related to general topic
Have at least another subtopic with at least 1 common keyword	77%	86%	65%
Share more than 1 keywords with at least another subtopic	62%	30%	~20%

Table 6: Evaluation for filters(Immigration, K=10, N=10)

words while others include words too general to draw any meaningful message. Both situations are difficult for subtopics to be meaningful and require an increase of K so that meaningful keywords are distributed in a more balanced way. In the case of K=12, we also see an uneven distribution of related words percentage but also nearly no percentage to be higher than 80%. This indicates that nearly all subtopics fail to generate specific meanings because K is too large so in situations like this we should decrease K. We ultimately determine that K=10 gives us the best result, with the highest average percentage of related words and a relatively even distribution of meaningful keywords per subtopic.

4.2.4 Error Analysis

There are several steps during the generation of subtopics that may cause error and require improvement in the future. The first is the number of iterations we choose for training can be higher. We currently do 50 iterations for the generation process due to time constraint but increasing the number of iterations no doubt will output subtopics with less randomness. The second choice we made that may lead to error is that in order for subtopics to be more distinguishable, we assign 0 weights to all words that do not appear in the keyword list. For example, when we choose the number of keywords per subtopic to be 10, the 11th keyword will be assigned 0 weight during calculation of sentiment score for a subtopic while in reality it is still a word with 11th highest weight within that subtopic. We hope to gain more insights into solving this issue in future research. The last issue we wish to look more into is that when we trained the LDA model with our data, we found that some general topics generate more distinguishable subtopics compared to others. The only explanation we came up with is that this was caused by the nature of the political topic—-Coronavirus, for example, is a topic more capable of being divided into independent aspects as compared to Immigration. We therefore seek to design different filter algorithms that may accommodate general topics' complex nature in the future.

4.3 Sentiment Analysis

4.3.1 Result and Evaluation

We divide the annotated data (4.1.2) into 3 groups

1. development set: 25 immigration + 25 coronavirus
2. immigration evaluation set: remaining 25 immigration
3. coronavirus evaluation set: remaining 25 coronavirus

in which the articles under immigration and coronavirus are randomly divided into two halves.

We then use the development set to build an algorithm to classify the sentiment vector of an article for a subtopic to one of the sentiment classes we defined (-1, 0, 1, X). We discover that due to the nature of news articles, the score for neutral is generally high so that we can not simply choose the highest among the three values—-using the ratio of neutral and either negative or positive gives us a better result. We define cases in which the sum of all three scores is under a certain threshold as irrelevant. The final algorithm we designed for classification achieves a matching percentage of 75.6% for the development set, where "a match" is defined as a situation in which the category our algorithm classifies a subtopic into is the same as our annotation (Algorithms 2).

Algorithm 2 Classify sentiment vector to sentiment class

for $sentiment\ vector$ = [neg, ne, pos] **do**
  if (neg+ne+pos) < 0.2 **then**
    Classify to class X (irrelevant)
  else if (neg<.05 **and** ne<.15 **and** pos<.05) **or** ne>4*neg **or** ne>4*pos **then**
    Classify to class 0 (neutral)
  else if neg>pos **then**
    Classify to class -1 (negative)
  else if neg<pos **then**
    Classify to class 1 (positive)
  end if
end for

We evaluate the 2 evaluation sets with this classification algorithm, and the percentage of matching is 74.6% for coronavirus and 77.3% for immigration (Table 8).

4.3.2 Error Analysis

The matching percentage for both general topics is far from desired and we summarize these possible reasons for our failure to obtain higher accuracy. First, for some subtopics that we rated as 1,0 or -1, the sentiment vectors generated from the model are [0 0 0], which means no keywords with

K	Average percentage of related words per subtopic	Percentage of subtopics with <=50% related words	Percentage of subtopics with 50%-80% related words	Percentage of subtopics with >=80% related words
12	~50	41.7	~50	<10
10	79	<10	70	~30
8	63	30	50	20

Table 7: Evaluation for subtopic number K(Immigration, N=10)

General Topic	% of matching
coronavirus	74.6
immigration	77.3

Table 8: Sentiment Model Evaluation

this subtopic have actually appeared in the article. This suggests that we may need to consider increasing the number of keywords in LDA. Second, the current mapping algorithm is too simple and incapable of showing the true power of the sentiment model, since we are not able to achieve a really high matching percentage even for the development set. Building more complicated algorithms or combining the 3-dimensional sentiment vector into one compound score may improve the model's performance. Last, for all the parameters tested, the matching rate for immigration is constantly higher than that of coronavirus, implying that there may be some hidden factors that differentiate those 2 general topics which require further inspection.

4.4 Final Result

While we require all articles under immigration and coronavirus for both the training of LDA and the building of classification algorithms (2) as well as the evaluation for sentiment analysis, we only calculate the final sentiment score for media with at least 20 articles under a general topic. This is because we believe that 20 is the minimum number that enables us to draw conclusions about media tendencies.

Our result (Appendix A) includes 14 media for each general topic, with 11 of them being the same. These tables offer a straightforward way to approach a media's view regarding a single topic and we hope the public may gain the information they need by finding certain media's views on specific topics and therefore hold more objective altitudes in regards to their expectations for each media. We are also able to discover some major patterns and trends by comparing the same media's view across different subtopics and different media's views on the same subtopic, and find some consensus reached by media across subtopics.

For coronavirus, we observe that it is a general trend for media believed by the public to be "left" to exhibit a more negative view, though the trend is not absolute meaning that one should not assume negativity to be a necessary consequence for what they believed as left media. In respect of the same media across different subtopics, we find that all media share overall negative views while the degree of negativity differs. When we adopt a holistic view, we discover that different media tend to share their views in regard to the degree of negativity. For example, most news outlets hold more negative views on the topic of Covid data/ CDC(subtopic 0) than on the topic of Social Media on Covid(subtopic 1), and the subtopic with the most negative views across all media is subtopic 8 (most media hold the most negative view in regards to this subtopic), Covid outbreak in China.

Unlike Coronavirus to which nearly all media hold negative attitudes toward all subtopics, we observe a variety of opinions when it comes to immigration, with more cases of positivity, neutral, and irrelevance. This probably indicates that immigration is also a more controversial topic. We also observe an uneven covering of subtopics by media—-there exist some subtopics to which most media have low scores for all three sentiments meaning that they are hardly mentioned, like subtopic 7(Working visa). There are also some subtopics that are covered by some but not others. Subtopic 8(ICE & detention), for instance, is more covered by Vox, the Guardian, and BBC than any other media, which shows that media have a tendency to choose the topic that they want to cover, and it is a choice that is observable statically.

5 Conclusion

Our research utilizes LDA combined with sentiment analysis to conduct topic extraction for

the generation of subtopics and opinion mining for analysis of media perspectives. We connect subtopics with sentiment scores by performing sentiment analysis for subtopic's keywords within the context of specific sentences to ultimately obtain sentiment scores for each subtopic of the same general topic from the same news outlet. Our study provides a way for the public to raise their awareness regarding media views on specific political topics to understand their own bias toward media, so that they can hold more accurate expectations when skimming through news articles and therefore avoid misinterpretation of opinions. We observe overall negative views with different negativity degree across subtopics from coronavirus for all media, and more opinion variety for immigration, with some nearly uncovered subtopics and an uneven distribution of subtopic covering which indicates an observable difference when it comes to media choosing the topic to cover. We also find that different media tend to share their views in terms of degree of negativity, that they reach a consensus on viewing certain subtopic as more negative, which is especially notable for coronavirus. We hope these findings can facilitate people in their understanding of media's political tendencies so that they can embrace a more objective attitude by being more informed about whether common beliefs about media views truly align with media's actual perspectives.

For future research, we intend to not only focus on the article itself, but also emphasize more on what media choose to present to the audience, since as shown in our findings news media are selective of content they decide to report. Another thing that we wish to improve is to find a reasonable way to combine the three values of a sentiment score into a single compound value to make media views to be presented in a more intuitive way. The last issue we will attempt to solve is to further ameliorate our method to better adapt to the transformation from subtopics to article sentiment scores and then to subtopic sentiment scores, so that our model can improve its overall performance to establish a better connection between media and their perspectives on various political topics.

References

R. Baly, G. Da San Martino, J. Glass, and P. Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. *EMNLP*.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Sarjoun Doumit and Ali Minai. 2012. Online news media bias analysis using an lda-nlp approach.

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

A.M. Johansen. 2010. Markov chain monte carlo. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education (Third Edition)*, third edition edition, pages 245–252. Elsevier, Oxford.

Silvia Julinda, Christoph Boden, and Alan Akbik. 2014. Extracting a repository of events and event references from news clusters. In *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, pages 14–18, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Heng Yang, Biqing Zeng, Mayi Xu, and Tianxing Wang. 2021. Back to reality: Leveraging pattern-driven modeling to enable affordable sentiment dependency learning.

Zhang Zhikai. 2016. Lda-gibbs-sampling.

A Appendix Final Result

News Media	0	1	2	3	4	5	6	7	8	9
Associated Press	0.207	0.099	0.15	0.096	0.118	0.14	0.103	0.087	0.241	0.108
	0.196	0.163	0.184	0.151	0.169	0.292	0.211	0.124	0.263	0.141
	0.03	0.029	0.05	0.045	0.039	0.046	0.076	0.021	0.03	0.021
BBC News	0.229	0.084	0.11	0.095	0.122	0.149	0.092	0.06	0.315	0.147
	0.215	0.19	0.161	0.099	0.196	0.303	0.189	0.113	0.331	0.178
	0.026	0.021	0.024	0.017	0.034	0.031	0.046	0.027	0.035	0.025
CNN (Web News)	0.24	0.068	0.14	0.085	0.11	0.153	0.13	0.082	0.239	0.089
	0.218	0.174	0.165	0.131	0.208	0.323	0.261	0.143	0.226	0.163
	0.046	0.035	0.052	0.041	0.048	0.059	0.08	0.042	0.034	0.031
Fox Online News	0.174	0.069	0.117	0.077	0.075	0.14	0.075	0.038	0.192	0.061
	0.222	0.196	0.245	0.193	0.201	0.415	0.219	0.076	0.289	0.102
	0.039	0.044	0.067	0.028	0.04	0.053	0.05	0.02	0.037	0.024
National Review	0.213	0.114	0.169	0.113	0.111	0.155	0.095	0.082	0.246	0.103
	0.171	0.154	0.163	0.128	0.168	0.249	0.167	0.1	0.249	0.123
	0.032	0.029	0.055	0.037	0.032	0.041	0.05	0.022	0.031	0.017
NPR Online News	0.195	0.069	0.126	0.064	0.109	0.125	0.081	0.063	0.189	0.073
	0.241	0.196	0.186	0.104	0.204	0.317	0.201	0.094	0.264	0.119
	0.039	0.034	0.036	0.03	0.045	0.051	0.041	0.013	0.025	0.018
Politico	0.217	0.058	0.145	0.099	0.099	0.168	0.117	0.08	0.226	0.075
	0.171	0.157	0.164	0.169	0.216	0.357	0.211	0.131	0.218	0.126
	0.038	0.026	0.05	0.055	0.036	0.061	0.067	0.03	0.028	0.033
The Atlantic	0.195	0.079	0.153	0.082	0.087	0.142	0.078	0.068	0.188	0.08
	0.198	0.151	0.17	0.072	0.167	0.222	0.161	0.098	0.211	0.113
	0.04	0.034	0.038	0.023	0.043	0.038	0.039	0.018	0.025	0.018
The Guardian	0.188	0.082	0.132	0.091	0.089	0.16	0.08	0.047	0.208	0.074
	0.219	0.195	0.239	0.176	0.194	0.378	0.215	0.083	0.281	0.103
	0.04	0.044	0.066	0.028	0.04	0.053	0.052	0.023	0.038	0.022
The Hill	0.177	0.068	0.113	0.071	0.071	0.138	0.087	0.04	0.169	0.049
	0.248	0.206	0.292	0.244	0.221	0.468	0.257	0.087	0.293	0.102
	0.038	0.024	0.064	0.02	0.035	0.046	0.052	0.018	0.027	0.022
USA TODAY	0.234	0.087	0.124	0.082	0.109	0.148	0.087	0.078	0.266	0.121
	0.243	0.2	0.183	0.096	0.211	0.293	0.203	0.127	0.299	0.168
	0.043	0.029	0.037	0.021	0.036	0.038	0.048	0.03	0.03	0.031
Vox	0.321	0.141	0.308	0.119	0.163	0.237	0.138	0.14	0.274	0.121
	0.281	0.196	0.269	0.084	0.214	0.255	0.199	0.135	0.234	0.144
	0.077	0.068	0.08	0.033	0.051	0.057	0.079	0.053	0.038	0.026
Washington Post	0.217	0.078	0.195	0.102	0.133	0.205	0.124	0.075	0.208	0.09
	0.22	0.15	0.201	0.102	0.205	0.279	0.222	0.081	0.207	0.167
	0.039	0.051	0.068	0.043	0.051	0.066	0.087	0.025	0.032	0.036
Washington Times	0.19	0.086	0.142	0.081	0.08	0.149	0.079	0.056	0.202	0.07
	0.185	0.148	0.188	0.091	0.174	0.279	0.163	0.089	0.226	0.105
	0.037	0.032	0.057	0.034	0.043	0.053	0.046	0.014	0.029	0.017

Table 9: Media perspective in respect of coronavirus subtopics
[blue | grey | red —— negative | neutral | positive]

News Media	0	1	2	3	4	5	6	7	8	9
BBC News	0.194	0.054	0.144	0.084	0.101	0.213	0.187	0.02	0.105	0.378
	0.14	0.07	0.197	0.193	0.157	0.208	0.384	0.053	0.115	0.236
	0.027	0.017	0.032	0.025	0.026	0.047	0.039	0.005	0.019	0.069
CNN (Web News)	0.156	0.095	0.079	0.169	0.182	0.087	0.083	0.058	0.046	0.124
	0.165	0.217	0.133	0.359	0.257	0.104	0.246	0.104	0.081	0.106
	0.057	0.078	0.035	0.089	0.103	0.033	0.088	0.029	0.014	0.042
Fox News	0.188	0.063	0.077	0.143	0.177	0.09	0.101	0.037	0.051	0.14
	0.171	0.103	0.111	0.278	0.195	0.11	0.242	0.083	0.083	0.118
	0.048	0.054	0.03	0.067	0.075	0.026	0.05	0.023	0.016	0.037
National Review	0.235	0.087	0.178	0.095	0.137	0.13	0.134	0.091	0.057	0.309
	0.139	0.068	0.128	0.115	0.106	0.118	0.227	0.08	0.059	0.142
	0.069	0.034	0.042	0.038	0.029	0.056	0.027	0.042	0.016	0.08
New York Times - News	0.091	0.064	0.046	0.12	0.143	0.067	0.062	0.03	0.027	0.104
	0.134	0.104	0.09	0.228	0.172	0.091	0.188	0.074	0.04	0.076
	0.039	0.06	0.03	0.063	0.07	0.031	0.045	0.023	0.011	0.033
NPR Online News	0.153	0.077	0.102	0.12	0.153	0.082	0.073	0.059	0.068	0.117
	0.161	0.14	0.137	0.225	0.2	0.134	0.186	0.13	0.084	0.112
	0.05	0.058	0.04	0.062	0.09	0.049	0.039	0.038	0.018	0.032
Politico	0.112	0.091	0.07	0.192	0.173	0.063	0.081	0.033	0.032	0.109
	0.149	0.158	0.107	0.367	0.224	0.066	0.235	0.076	0.055	0.088
	0.042	0.071	0.023	0.091	0.071	0.021	0.048	0.026	0.011	0.028
The Guardian	0.215	0.073	0.162	0.08	0.184	0.223	0.197	0.027	0.137	0.339
	0.186	0.11	0.229	0.176	0.187	0.239	0.375	0.071	0.163	0.219
	0.058	0.039	0.049	0.038	0.043	0.048	0.032	0.029	0.018	0.064
The Hill	0.172	0.048	0.123	0.125	0.242	0.17	0.116	0.035	0.054	0.317
	0.118	0.143	0.216	0.304	0.292	0.162	0.319	0.07	0.095	0.338
	0.025	0.036	0.037	0.032	0.03	0.031	0.056	0.015	0.018	0.077
Townhall	0.268	0.057	0.125	0.125	0.2	0.113	0.143	0.042	0.066	0.206
	0.147	0.076	0.16	0.218	0.186	0.137	0.283	0.047	0.086	0.173
	0.043	0.029	0.028	0.044	0.041	0.03	0.047	0.013	0.015	0.063
USA TODAY	0.222	0.07	0.154	0.111	0.155	0.141	0.143	0.055	0.084	0.291
	0.19	0.116	0.213	0.265	0.24	0.22	0.339	0.088	0.134	0.237
	0.053	0.035	0.044	0.044	0.058	0.048	0.046	0.026	0.018	0.065
Vox	0.247	0.101	0.197	0.137	0.212	0.169	0.153	0.072	0.137	0.346
	0.185	0.113	0.207	0.208	0.193	0.211	0.257	0.106	0.134	0.167
	0.054	0.049	0.057	0.049	0.059	0.064	0.035	0.035	0.027	0.069
Wall Street Journal - News	0.205	0.082	0.128	0.136	0.182	0.103	0.098	0.062	0.056	0.167
	0.221	0.144	0.203	0.296	0.232	0.171	0.252	0.125	0.106	0.162
	0.056	0.059	0.05	0.071	0.069	0.045	0.064	0.038	0.024	0.052
Washington Times	0.286	0.068	0.129	0.146	0.235	0.102	0.133	0.041	0.069	0.146
	0.242	0.103	0.159	0.279	0.232	0.115	0.237	0.076	0.099	0.117
	0.083	0.044	0.045	0.053	0.083	0.036	0.048	0.022	0.021	0.041

Table 10: Media perspective in respect of immigration subtopics
[blue | grey | red —— negative | neutral | positive]