Sentiment Analysis and NLP models for Identifying Biases of Online News Stations

Anuska Acharya
aachary4@gmu.edu

Grace Cox
gcox4@gmu.edu

***Abstract*—This work attempts to identify potential reporting bias surrounding recent controversial decisions for articles published from August 2019 to October 2021 within four major news organizations: FOX, CNN, NBC, and NPR. This potential bias is determined by conducting a Sentiment Analysis using NLTK's (Natural Language Tool Kit) VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Intensity Analyzer. Copious amounts of literature have been published regarding sentiment analysis and bias identification of news articles, though none employ VADER. The team determines the overall sentiment for an article using the Polarity Compound calculated by the Sentiment Intensity Analyzer, which then corresponds to a political tone indicated within the verbiage and context of the article. Upon completion of the analysis, it was found that CNN, NBC, and NPR tend to have the most negative sentiment surrounding this topic, while FOX tends to be more neutral though still on the positive side. This translates to the surprising identification of a slightly democrat tone for articles published by FOX, and a more republican tone for those articles published by NPR, CNN, and NBC.**

***Index Terms*—Machine learning, Sentiment analysis, NLP, News**

1. Problem Description

The U.S. War in Afghanistan informally began in 1999 when the United Nations declared al-Qaeda and the Taliban to be terrorist entities and, "impose[d] sanctions on their funding, travel, and arms shipments" [1]. With that being said, the United States became formally involved in the War in Afghanistan following the Terrorist Attacks on the World Trade Center and Pentagon on September 11, 2001, that were sparked by the assassination of Ahmad Shah Massoud, the commander of an anti-Taliban coalition called the Northern Alliance. The September 11 terrorist attacks on the United States resulted in a rapid increase in the number of troops on the ground in Afghanistan, up until 2012 when U.S. involvement in the middle east began to see a rapid decline (Figure 1). Although the U.S. has been decreasing the troop levels in Afghanistan since 2012, the complete removal of all American troops from Afghanistan on August 30, 2021, sparked major political upset, both in the Middle East and here in the United States.

The following diagram (Figure 2) displays the timeline of events surrounding the United States' complete withdrawal from Afghanistan, beginning with the signing of

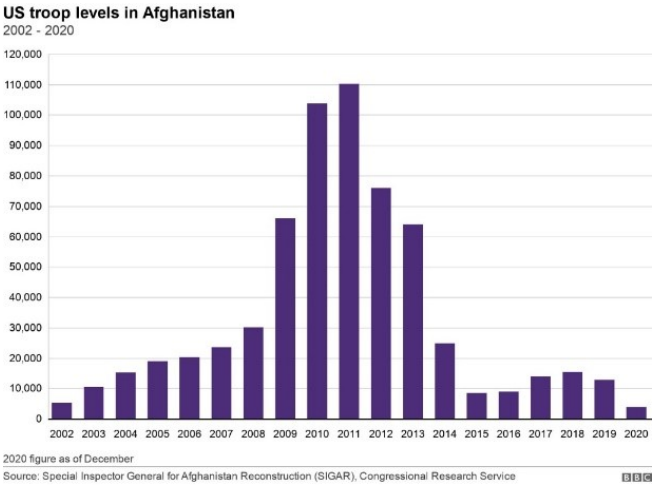

Figure 1. Levels of US Troops in Afghanistan from 2002-2020

the Doha Agreement by President Donald Trump and the Taliban in February 2020, and ending with Joe Biden's decision to pull all remaining troops from Afghanistan on August 30, 2021 [2].

Being that the United States' complete withdrawal from Afghanistan is major international news, there is value in understanding the potential biases that lie within the many news organizations reporting on this topic throughout the United States. This project will focus conducting a sentiment analysis of 30 news articles written and released by each of four major news stations, to include: FOX, NBC, CNN, and ABC, regarding the United States' withdrawal from Afghanistan earlier this year. This sentiment analysis will aid in identifying potential biases that lie within these different organizations when reporting on a major political event.

2. Importance of problem

The decision by the U.S. to withdrawal from Afghanistan is a history-making decision that has ended the United States' nearly 20-year involvement in the war against terrorism, al-Qaeda, and the Taliban in Afghanistan and was met with mixed reactions of support and opposition from the general American public (Figure 3) [3]. We note that roughly 70% of individuals identifying as Democrat and Independent support this decision, mirroring the 70%

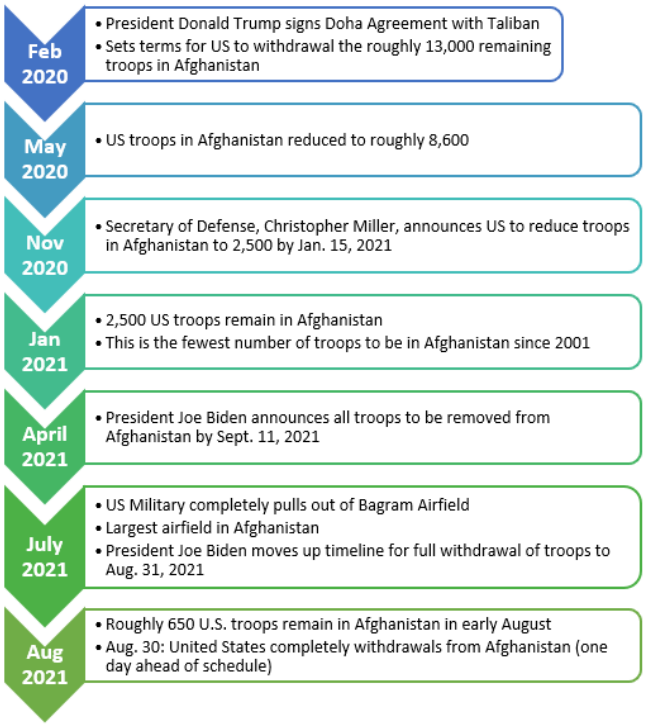

Figure 2. Timeline of the United States' Withdrawal from Afghanistan

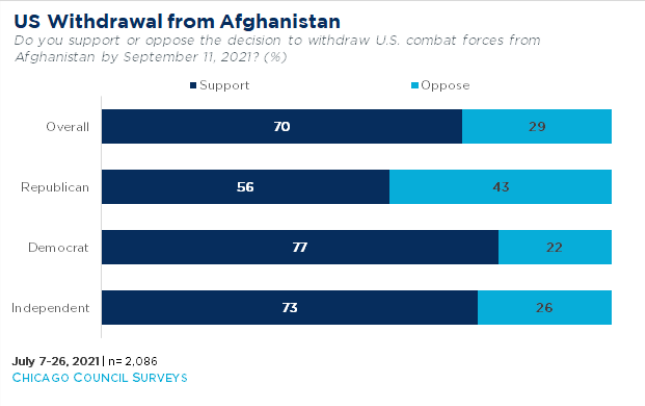

Figure 3. Political Support regarding the US Withdrawal from Afghanistan

overall support this decision received. This is contrasted by the near split in support/opposition responses seen within individuals who identify as Republican.

The Republican and Democrat parties appear to be on opposing sides of this issue, with the majority of individuals with Republican leanings believing that the United States made the wrong decision in completely pulling out of Afghanistan and the majority of Democrats believing that the U.S. decision to withdraw troops from Afghanistan was the right decision (Figure 4) [4]. Despite the publics' varying levels of support surrounding the United States' decision to withdrawal from Afghanistan, there has been an abundance of news coverage regarding this hot topic. Machine learning

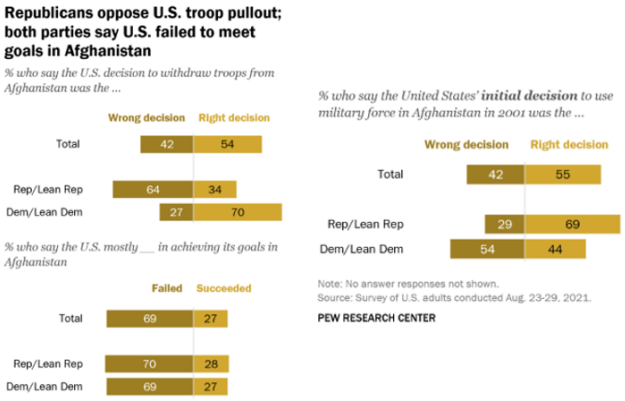

Figure 4. PEW Research Center Poll regarding US Troop Pullout

models for fake news detection and text analysis could help to better understand the online posts [**?**], [**?**], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

3. Preliminary Literature Review

Many researchers have worked in sentiment analysis on different news stations. Research conducted by Reis, Benevenuto, Melp, Prates, Kwak collected a total of 69, 907 headlines from four different news sources: BBC News Online, Daily Mail Online, The New York Times, and Reuters Online [26]. The primary purpose of the analysis was to identify the sentiment of the headlines of various news articles produced by these big and popular news sources. The research focused on identifying if such positive or negative sentiment of the headlines was able to generate more clicks. This research found that the polarity of the headlines of the article impacted the popularity of the news article leading to more clicks. The study also revealed the headline with a positive or negative tone attracted more readers compared to the headlines with a neutral sentiment.

Similarly, in another research article by Islam, Ashraf, Abir and Mottailb [27] conducted sentiment analysis to detect the polarity based on sentence structure and dynamic dictionary. The research used detection of sentences and the use of a library for the classification of the news article. The library was defined using a list of reserved words. The proposed algorithm selected online news articles and extracted paragraphs, sentences, phrases, and words. The end of the sentence was detected if the sentence included the "." full stop sign. Once the end of the sentence was detected, the algorithm then searched for positive or negative words, sentences and phrases and determined the polarity of the news articles. The researchers utilized java programming languages and NetBeans IDE to develop the interface for the algorithm. The experiment included the extraction of 56 random news articles from The Independent, The Telegraph, and The Daily Star which resulted in only an 8.93% margin of error which is only 5 out of 56 articles. Further analysis

TABLE 1.

News Organizations Covered in Analysis and Number of Articles Used	
News Organization	**Number of Articles Scraped**
FOX	20
NBC	20
CNN	20
NPR	20

TABLE 2.

Sentiment Analyses to be Conducted on News Articles Covering U.S. Withdrawal from Afghanistan	
1	**2**
Positive	Democrat
Negative	Republican
Neutral	Independent

of the error showed these articles had a smaller number of sentences which resulted in difficulty in determining the polarity.

A research article by Jagdale, Deshmukh, and Shirsath [28] explained the three levels of sentiment analysis and advanced online text analysis [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41]. These levels are document level, sentence level, and entity and aspect level. In document-level sentiment analysis, the whole document is analyzed, and the polarity of the entire document is identified. Similarly, in sentence-level sentiment analysis, each and all sentences are analyzed to determine the polarity; positive, negative, or neutral. Entity and aspect level sentiment analysis is based on opinion. The research involved 2225 documents from the BBC. The methodology used in this research was tokenization, stop words, stemming, and then assigning scores based on sentiments.

4. Proposed Approach

To complete our Sentiment Analysis successfully and adequately, the team will focus on four major news organizations, previously listed, and will scrape a sufficient number of articles from the websites of each organization. The following table (Table 1) provides an overview of the four major news organizations analyzed throughout this project as well as the number of articles regarding the United States' withdrawal from Afghanistan that were scraped for use throughout our Sentiment Analysis.

Two types of Sentiment Analysis (Table 2) will be conducted throughout this project. The first is a general Sentiment Analysis that describes the overall tone of the article as either Positive, Negative, or Neutral, which will aid the team in determining the overall attitude within each news organization regarding the United States' complete withdrawal from Afghanistan. The second Sentiment Analysis will describe the political tone of the article as being Democrat, Republican, or Independent, which will aid in highlighting the overall political biases present within each news organization surrounding the United States' withdrawal from Afghanistan.

5. Proposed Method For Evaluation

The team intends to take advantage of various Natural Language Processing (NLP) techniques using programs such as Python and R to conduct a sentiment analysis on each of the 120 articles covering the United States' withdrawal from Afghanistan published by one of four news stations (FOX, CNN, NBC, NPR). The team will then use the results of this sentiment analysis to identify any potential biases surrounding major political events that are apparent within each of these major news organizations. Several text preprocessing steps must be completed on all text to successfully complete a sentiment analysis on these articles published by each of the four major news organizations. These text preprocessing steps can include tokenization, stop word and punctuation removal, lemmatization, stemming and more, and are described in detail below.

5.1. Text Preprocessing Definition

Preprocessing text data refers to the process of transforming the text input into a, "predictable and analyzable" [42] form for the task at hand and the ultimate goal of cleaning and preprocessing text data is to, "... reduce the text to only the words that you need for your NLP goals" [43]. It should be noted that different tasks require an emphasis on different steps within the preprocessing procedure. Some common types of text preprocessing techniques include tokenization, punctuation/noise removal, and text normalization (lowercase tokens, stop word removal, and lemmatization/stemming).

5.2. Tokenization

Tokenization refers to the preprocessing task of, "breaking up text into smaller components of text (known as tokens)" [44]. A token may be an entire word, a part of a word, or characters like punctuation. Tokenization is one of the most important parts of NLP preprocessing, as it defined what our models can express. In this project, tokenization was done using SpaCy [45], a Python package often used in NLP settings. SpaCy not only provides generic tokenization functions, but also allows the user to, "customize the tokenization process to detect tokens on custom characters" [46]. This custom tokenization in SpaCy could be used for words including hyphens or apostrophes that should be processed as a single token.

5.3. Punctuation & Noise Removal

Punctuation in text data does not add much, if any, value to the data and the meaning behind it and thus is typically removed from the raw textual data during the preprocessing steps. Punctuation includes characters including, but not limited to commas, periods, exclamation/question marks, hyphens, and apostrophes.

TABLE 3.

Raw Text Data	Lowercase Text Data
CaMeL	camel
UPPER	upper
lower	lower

5.4. Lowercase Tokens

Text data usually contains characters that have different cases, some of which may not be conducive to the Natural Language Processing procedure. In order to further normalize textual data for ease of analysis, all tokens are typically converted into a lowercase capitalization scheme. The following table (Table 3) displays a few examples for creating lowercase tokens from raw textual data.

5.5. Stop Word Removal

Stop words are commonly used words in a language that provide little to no information to the text. Some examples of stop words in the English language can include, but are not limited to: "the", "is", "a", "are" [42]. There are many packages available within Python that are capable of detecting and removing stop words from the provided text, with the most popular being the NLTK package. Stop word removal is one of few text preprocessing tasks that are used for text normalization.

5.6. Lemmatization & Stemming

Stemming refers to the NLP preprocessing task that is concerned with, "bluntly removing word affixes (prefixes and suffixes)" [2]. There are two major errors that can arise from Stemming Algorithms: Over Stemming and Under Stemming. Over stemming occurs when two words that have different stems are stemmed to the same root word; for example, the words "universal", "university", and "universe" are stemmed to "univers" which would not be correct since their modern meanings are in different domains and are generally not synonymous [47]. Under stemming occurs when, "two words that should be stemmed to the same root are not" [47]. For example, the words "alumnus", "alumni", and "alumnae" should all be stemmed to the same word, but typically are not. Textual data can contain tokens that are different forms of a certain word (i.e. walk, walked, walking) and condensing all of the tokens in a text to their root word increases the ease of analysis while at the same time reducing inflectional forms. Lemmatization refers to this process of taking each token and bringing it down into its root form. The goal of lemmatization is to "... remove inflectional endings only and to return the base or dictionary form of a word" [48]. The dictionary form of a particular word is referred to as the lemma for that word. The example below (Figure 5) displays lists of possible words contained in text data as well as the respective lemma for each list of words.

[am, is, are] → be
[walk, walked, walking] → walk
[watches, watching, watched] → watch

Figure 5. Lemmatization Examples

6. Results

Prior to conducting this sentiment analysis, date metadata was scraped from each of the 80 articles to determine the month and year that each article was published in (Figure 6). This date metadata was provided in a schema.org annotation format within @type: NewsArticle and will provide the team with further context for the sentiment analysis to be conducted. We note all articles, except for one, were published in 2021, with the majority of articles being published in August and September 2021. This coincides with the previously mentioned timeline regarding the United States' complete withdrawal from Afghanistan.

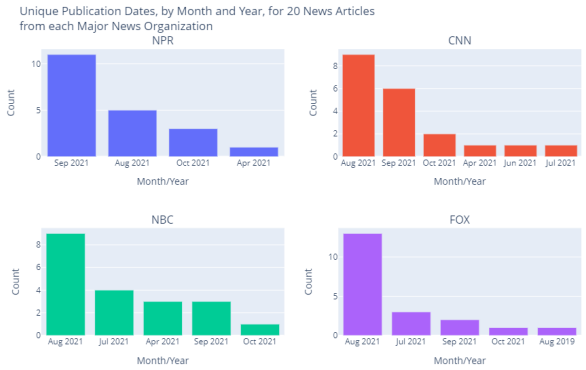

Figure 6. Publication Dates for Scraped Articles, by News Organization

This sentiment analysis of four major news organizations is based on 20 articles published by each of FOX, NBC, CNN, and NPR regarding the United States' withdrawal from Afghanistan; the links for each of the articles used within this analysis can be found in Table 1 within Appendix A. It should be noted that though the team initially explored this text data upon completion of textual pre-processing, due to time constraints, the team decided to approach the problem from a different angle. Each webpage is scraped using the BeautifulSoup and urlopen libraries within Python, then using NLTK's VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Intensity Analyzer, a Polarity Compound score indicating the articles overall sentiment is calculated. This Polarity Compound score is a normalized sum of the Positive, Negative, and Neutral scores that ranges in value from -1 to 1, with positive values indicating a positive sentiment, negative values indicating a negative sentiment, and values close to 0 indicating a more neutral sentiment. It should be noted that VADER's Sentiment Intensity Analyzer is pretrained and thus does

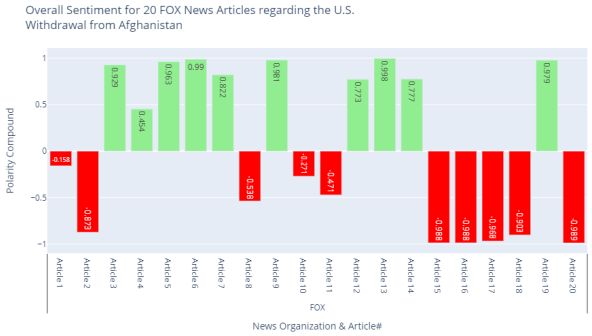

Figure 7. Sentiment Analysis for FOX News Articles

Figure 8. Sentiment Analysis for NBC News Articles

Figure 9. Sentiment Analysis for CNN News Articles

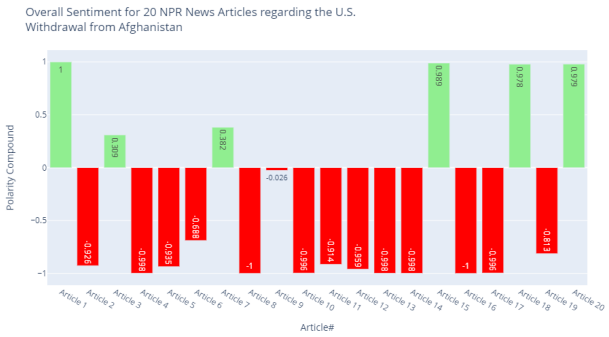

Figure 10. Sentiment Analysis for NPR News Articles

not require any training effort, making it ideal considering the time constraints placed on this project.

The following Figure 7 displays the Polarity Compound scores for each of the 20 articles regarding the United States' withdrawal from Afghanistan that were published online by FOX News. This figure displays that FOX is relatively split between articles classified as having a positive sentiment and those having a negative sentiment. With that being said, the articles with positive sentiment tend to have a stronger positive sentiment than those articles with negative sentiments, in which case we see weaker negative sentiments (indicated by polarity compound scores that are closer to 0 e.g., -0.158, -0.271).

The following Figure 8 displays the Polarity Compound scores for each of the 20 articles regarding the United States' withdrawal from Afghanistan that were published online by NBC News. This figure displays that NBC tends to have a largely negative sentiment within articles regarding this controversial political decision. There were only two articles out of the 20 that were used within this analysis that were classified as having a positive sentiment. We note that of the selected articles published, the majority have a very strong negative sentiment, indicated by Polarity Compound Scores that are either -1 or very close to -1.

The following Figure 9 displays the Polarity Compound scores for each of the 20 articles regarding the United States' withdrawal from Afghanistan that were published online by CNN News. We note that, like NBC, CNN tends to have a largely negative sentiment within the published articles regarding this monumental decision. Only four of the 20 articles were classified as having a positive sentiment. Also, we noted that the articles with a negative sentiment are typically stronger than those with a positive sentiment for the articles published by CNN.

The following Figure 10 displays the Polarity Compound scores for each of the 20 articles regarding the United States' withdrawal from Afghanistan that were published online by NPR. We note that, like NBC and CNN, NPR tends to have a largely negative sentiment within the published articles regarding this monumental decision. Only six of the 20 articles were classified as having a positive sentiment. Also, we noted that the articles with a negative sentiment are typically stronger than those with a positive sentiment, and most of the Polarity Compound scores for these articles are both very strong (i.e. close to 1) and negative.

The following Figure 11 displays the Average Sentiment for the four major news organizations FOX, NBC, CNN, and NPR. The Average Sentiment of all 20 articles is determined by calculating the average polarity compound score across all articles published by each of the major news organizations previously mentioned. This figure aids in highlighting the overall bias that is present within the articles published by FOX, NBC, CNN, and NPR regarding the United States' withdrawal from Afghanistan. We see that while FOX tends

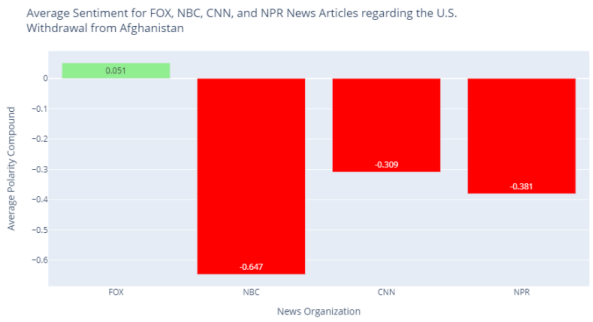

Figure 11. Average Sentiment Analysis by News Organization

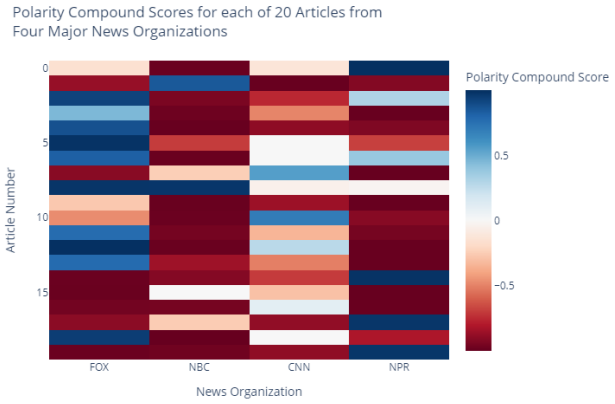

Figure 12. Political Tone Analysis for Individual Articles by News Organization

to be relatively neutral when reporting on this topic, there is still a slightly positive overall bias present. In contrast, NBC, CNN, and NPR all have a negative bias, with NBC having a strong negative bias when reporting on the United States' decision to withdrawal completely from Afghanistan, indicating their strong disapproval of this decision, and CNN & NPR having a relatively moderate negative bias'.

In addition to determining the overall general sentiment, we want to determine whether the articles published by each of these four major news organizations tend to have a Democrat, Republican, or Neutral leaning. Figure 12 depicts the Polarity Compound scores for each of the 20 articles for ABC, CNN, NPR, and FOX, with blue shades indicating a more Democrat leaning, red shades indicating a more Republican leaning, and white shades indicating a Neutral leaning. It should be noted that the political leaning of an article is determined by the value of its polarity compound calculated by VADER. Polarity Compound values closer to 1 indicate a democrat tone and those values closer to -1 indicate a republican tone. CNN appears to have the most neutral articles (though still Republican in leaning), with only two articles having strong Democrat leanings, while NBC and NPR are clearly the most Republican leaning in nature as there are more deep red shades present in Figure 10. One surprising result from both Figure 11 and Figure 12 is that FOX typically tends to be more Republican in nature, while NBC and CNN tend to lean more Democrat. This result could be explained by the fact that there was a change of presidency during the process of the United States' withdrawal from Afghanistan. Another explanation for this surprising result for FOX News stems from the information displayed above in Figure 3, which shows that Republicans are fairly split regarding the United States' decision to completed withdrawal from Afghanistan, with only 56% of surveyed republicans supporting this decision.

7. Future Work

One potential area of future work regarding discovering biases within articles published by major news organization regarding the United States' complete withdrawal from Afghanistan includes creating a model that is unique to the verbiage and context of these articles. Due to the time constraints of the course, the team was unable to both train and employ a Natural Language Processing model with adequate accuracy. Since the team decided to use the Polarity Compound scores for each article to represent the articles political tone, another area for future work would entail creating dictionaries and lists of keywords that are representative of each political leaning (Democrat, Republican, or Independent).

8. Project Timeline

The following table (4) contains the proposed timeline for this project, spanning the roughly 14 weeks of the semester. It should be noted that the estimated due dates listed below are subject to change as the project progresses. The project timeline is divided into four sections: Project Proposal, Project Milestone 1, Project Milestone 2, and Project Final Report & Code. The details on these sections and what these four sections entail is listed below.

References

[1] C. on foreign relations. "the u.s. war in afghanistan". [Accessed December 9, 2021]. [Online]. Available: https://www.cfr.org/timeline/us-war-afghanistan

[2] E. Kiely and R. Farley. "timeline of u.s. withdrawal from afghanistan". [Online]. Available: https://www.factcheck.org/2021/08/timeline-of-u-s-withdrawal-from-afghanistan/

[3] D. Smeltz and E. Sullivan. (August 9, 2021) "us public supports withdrawal from afghanistan". [Online]. Available: https://www.thechicagocouncil.org/commentary-and-analysis/blogs/us-public-supports-withdrawal-afghanistan

[4] T. V. Green and C. Doherty. "majority of u.s. public favors afghanistan troop withdrawal; biden criticized for his handling of situation". [Online]. Available: https://www.pewresearch.org/fact-tank/2021/08/31/majority-of-u-s-public-favors-afghanistan-troop-withdrawal-biden-criticized-for-his-handling-of-situation/

TABLE 4.

Group	Item	Estimated Due Date
Project Milestone 1	Detailed Description of Problem and Motivation	09/23/2021
	Detailed Literature Review	09/28/2021
	Describe/Refine Proposed Approach	09/30/2021
	Preliminary Results	10/03/2021
	Final Project Milestone 1	**10/04/2021 @ 11:59PM**
Project Milestone 2	Detailed Description of Problem and Motivation	10/10/2021
	Detailed Literature Review	10/15/2021
	Describe/Refine Proposed Approach	10/20/2021
	Preliminary Results	10/30/2021
	Final Project Milestone 2	**11/06/2021 @ 11:59PM**
Project Final Report & Code	Project Presentation	12/01/2021
	Final Project Report & Code	**12/06/2021 @ 11:59PM**

[5] M. Heidari, J. H. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2020, pp. 480–487.

[6] W.-Y. S. Chou, A. Oh, and W. M. P. Klein, "Addressing health-related misinformation on social media," vol. 320, no. 23, p. 2417, Dec. 2018. [Online]. Available: https://doi.org/10.1001/jama.2018.16865

[7] L. Cui and D. Lee, "Coaid: COVID-19 healthcare misinformation dataset," *CoRR*, vol. abs/2006.00885, 2020. [Online]. Available: https://arxiv.org/abs/2006.00885

[8] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, 2017, pp. 963–972. [Online]. Available: https://doi.org/10.1145/3041021.3055135

[9] J. H. Fetzer, "Disinformation: The use of false information," vol. 14, no. 2, pp. 231–240, May 2004. [Online]. Available: https://doi.org/10.1023/b:mind.0000021683.28604.5b

[10] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2020, pp. 322–329.

[11] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–6.

[12] E. Dolgin, "COVID vaccine immunity is waning — how much does that matter?" *Nature*, vol. 597, no. 7878, pp. 606–607, Sep. 2021. [Online]. Available: https://doi.org/10.1038/d41586-021-02532-4

[13] U.S. Food and Drug Administration. "covid-19 vaccines". [Accessed November 6, 2021]. [Online]. Available: https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines

[14] M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, "Ontology creation model based on attention mechanism for a specific business domain," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–5.

[15] N. Sallahi, H. Park, F. E. Mellouhi, M. Rachdi, I. Ouassou, S. Belhaouari, A. Arredouani, and H. Bensmail, "Using unstated cases to correct for COVID-19 pandemic outbreak and its impact on easing the intervention for qatar," *Biology*, vol. 10, no. 6, p. 463, May 2021. [Online]. Available: https://doi.org/10.3390/biology10060463

[16] M. El-Harbawi, B. B. Samir, M.-R. Babaa, and M. I. A. Mutalib, "A new QSPR model for predicting the densities of ionic liquids," *Arabian Journal for Science and Engineering*, vol. 39, no. 9, pp. 6767–6775, Jun. 2014. [Online]. Available: https://doi.org/10.1007/s13369-014-1223-3

[17] M. Heidari, H. James Jr, and O. Uzuner, "An empirical study of machine learning algorithms for social media bot detection," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–5.

[18] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," vol. 359, no. 6380, pp. 1094–1096, Mar. 2018. [Online]. Available: https://doi.org/10.1126/science.aao2998

[19] Q. Su, M. Wan, X. Liu, and C.-R. Huang, "Motivations, methods and metrics of misinformation detection: An NLP perspective," vol. 1, no. 1-2, p. 1, 2020. [Online]. Available: https://doi.org/10.2991/nlpr.d.200522.001

[20] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2020, pp. 1–6.

[21] S. Akon and A. Bhuiyan, "Covid-19: Rumors and youth vulnerabilities in bangladesh," 07 2020.

[22] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0285–0291.

[23] M. Fernandez and H. Alani, "Online misinformation." ACM Press, 2018. [Online]. Available: https://doi.org/10.1145/3184558.3188730

[24] H. Zhang, A. Kuhnle, J. D. Smith, and M. T. Thai, "Fight under uncertainty: Restraining misinformation and pushing out the truth." IEEE, Aug. 2018. [Online]. Available: https://doi.org/10.1109/asonam.2018.8508402

[25] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0255–0261.

[26] J. dos Reis, F. Benevenuto, P. O. S. V. de Melo, R. O. Prates, H. Kwak, and J. An, "Breaking the news: First impressions matter on online news," *CoRR*, vol. abs/1503.07921, 2015. [Online]. Available: http://arxiv.org/abs/1503.07921

[27] M. U. Islam, F. B. Ashraf, A. I. Abir, and M. A. Mottalib, "Polarity detection of online news articles based on sentence structure and dynamic dictionary," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*. IEEE, Dec. 2017. [Online]. Available: https://doi.org/10.1109/iccitechn.2017.8281777

[28] V. S. Shirsat, R. S. Jagdale, and S. N. Deshmukh, "Document level sentiment analysis from news articles," in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. IEEE, Aug. 2017. [Online]. Available: https://doi.org/10.1109/iccubea.2017.8463638

[29] M. Heidari, S. Zad, M. Malekzadeh, P. Hajibabaee, , S. HekmatiAthar, O. Uzuner, and J. H. J. Jones, "Bert model for fake news detection based on social bot activities in the covid-19 pandemic," in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*. IEEE, 2021.

[30] A. Ain, "The WHO is right to call a temporary halt to COVID vaccine boosters," *Nature*, vol. 596, no. 7872, pp. 317–317, Aug. 2021. [Online]. Available: https://doi.org/10.1038/d41586-021-02219-w

[31] E. Callaway, "COVID vaccine boosters: the most important questions," *Nature*, vol. 596, no. 7871, pp. 178–180, Aug. 2021. [Online]. Available: https://doi.org/10.1038/d41586-021-02158-6

[32] A. Weatherton. "health expert says booster shot could be needed after getting covid-19 vaccine". [Accessed June 8, 2021]. [Online]. Available: https://www.13newsnow.com/article/life/booster-shot-may-be-needed-after-covid-19-vaccine/291-49a8966c-3d91-48ad-99a0-02905c5593cc

[33] M. Malekzadeh, P. Hajibabaee, M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, "Review of graph neural network in text classification," in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*. IEEE, 2021.

[34] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[35] P. Hajibabaee, M. Malekzadeh, , M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, "An empirical study of the graphsage and word2vec algorithms for graph multiclass classification," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021.

[36] D. S. Khoury, D. Cromer, A. Reynaldi, T. E. Schlub, A. K. Wheatley, J. A. Juno, K. Subbarao, S. J. Kent, J. A. Triccas, and M. P. Davenport, "Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection," *Nature Medicine*, vol. 27, no. 7, pp. 1205–1211, May 2021. [Online]. Available: https://doi.org/10.1038/s41591-021-01377-8

[37] J. Havey. "pharma research progress hope,". [Online]. Available: https://catalyst.phrma.org/a-year-and-a-half-later-the-biopharmaceutical-industry-remains-committed-to-beating-covid-19?utm$_c ampaign = 2021 - q3 - cov - innutm_m edium = pai_s rh_c pc - ggl - adfutm_s ource = gglutm_c ontent = clk - pol - tpv_s cl - geo_s td - usa - dca - pai_s rh_c pc - ggl -$

[38] P. R. Krause, T. R. Fleming, R. Peto, I. M. Longini, J. P. Figueroa, J. A. C. Sterne, A. Cravioto, H. Rees, J. P. T. Higgins, I. Boutron, H. Pan, M. F. Gruber, N. Arora, F. Kazi, R. Gaspar, S. Swaminathan, M. J. Ryan, and A.-M. Henao-Restrepo, "Considerations in boosting COVID-19 vaccine immune responses," *The Lancet*, vol. 398, no. 10308, pp. 1377–1380, Oct. 2021. [Online]. Available: https://doi.org/10.1016/s0140-6736(21)02046-8

[39] J. H. Kim, F. Marks, and J. D. Clemens, "Looking beyond COVID-19 vaccine phase 3 trials," *Nature Medicine*, vol. 27, no. 2, pp. 205–211, Jan. 2021. [Online]. Available: https://doi.org/10.1038/s41591-021-01230-y

[40] E. C. Fernández and L. Y. Zhu, "Racing to immunity: Journey to a COVID-19 vaccine and lessons for the future," *British Journal of Clinical Pharmacology*, vol. 87, no. 9, pp. 3408–3424, Jan. 2021. [Online]. Available: https://doi.org/10.1111/bcp.14686

[41] S. Zad, M. Heidari, P. Hajibabaee, and M. Malekzadeh, "A survey of deep learning methods on semantic similarity and sentence modeling," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021.

[42] K. Ganesan. (April 2019) "all you need to know about test preprocessing for nlp and machine learning.". [Online]. Available: https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

[43] Code Academy. "text preprocessing". [Online]. Available: https://www.codecademy.com/courses/text-preprocessing/lessons/text-preprocessing/exercises/introduction

[44] ——. "natural language processing/text preprocessing". [Online]. Available: https://www.codecademy.com/learn/dscp-natural-language-processing/modules/dscp-text-preprocessing/cheatsheet

[45] spaCy. "spacy 101: Everything you need to know". [Online]. Available: https://spacy.io/usage/spacy-101

[46] Real Python. "tokenization in spacy". [Online]. Available: https://realpython.com/natural-language-processing-spacy-python/#tokenization-in-spacy

[47] T. Srivastava. (August 6, 2019) "nlp: A quick guide to stemming.". [Online]. Available: https://medium.com/@tusharsri/nlp-a-quick-guide-to-stemming-60f1ca5db49e

[48] stanford. (2009) "stemming and lemmatization". [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html