Understanding the Afghan Conflict: Insights from Machine Learning and Text Analysis

Abuzar Royesh

06/12/2019

Contents

Introduction

As President Donald Trump contemplates a full withdrawal of U.S. troops from Afghanistan[1], many within the research and policy making community attempt to understand how the dynamics of the conflict have changed over time. Now, 18 years later, little is left of the optimism with which the war was started in 2001. How have the U.S. administration and the international community represented the struggle that is dubbed a lost war by many[2]? And how has that portrayal changed over time? There is a dearth of rigorous empirical analysis of this question. This paper seeks to utilize machine learning and text analysis to analyze the vast corpus of U.S. government and international community documents on Afghanistan. In particular, I look at more than 8,000 official English language newswires and press releases published by various sources between 2004 and 2019 to assess the change in topics and sentiment. The following sections lay out the research question, methodology, findings, and areas of future inquiry.

Research Question

How does the content and sentiment of the official English language newswires and press releases on Afghanistan vary over time?

Methodology

Data

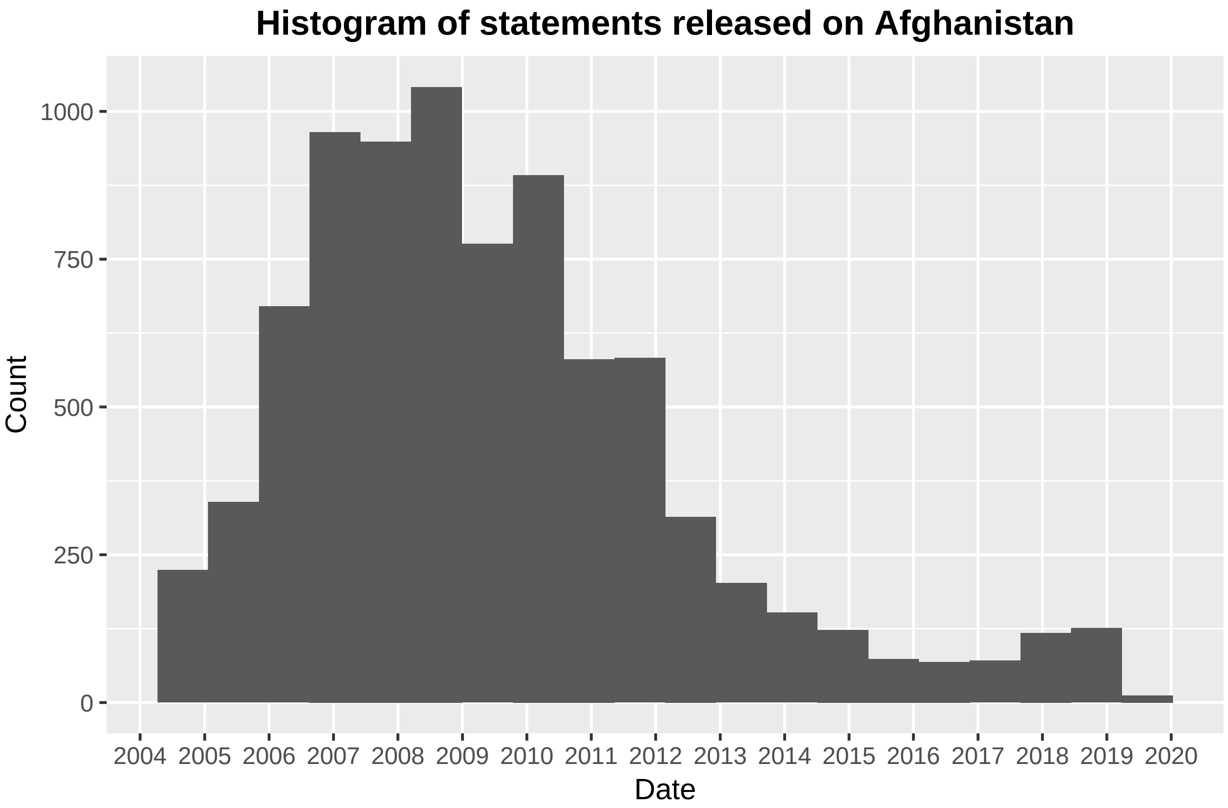
For the purpose of this research project, I downloaded all official English language newswires and press releases from Nexis Uni. The data set contained 8,281 documents dating between May 03, 2004 to April 04, 2019. From each document, I parsed the time and date of publication, word count, document type, and the name of the office that released the document. The offices that had released these documents included the following:

- The White House,
- The U.S. Department of State,
- The U.S. Department of Defense,
- Voice of America,
- NATO's International Security Assistance Force (ISAF),
- The U.K. government,
- The World Bank, and
- The United Nations.

The documents included press releases, news releases, news stories, statements, transcript of speeches, articles, newsletters, columns, hearings, fact sheets, and announcements. The minimum word count for an individual document was 35 words (two press releases by ISAF) and the maximum stood at 10,875 (A press release by the United Nations). The following figure shows the frequency of articles by date of publication.

[1] Gibbons-Neff, Thomas, and Julian E. Barnes. "Under Peace Plan, U.S. Military Would Exit Afghanistan Within Five Years." The New York Times, March 1, 2019, sec. U.S. https://www.nytimes.com/2019/02/28/us/politics/afghanistan-military-withdrawal.html.

[2] Young, Stephen B. "Why America Lost in Afghanistan." Foreign Policy (blog). Accessed June 12, 2019. https://foreignpolicy.com/2019/02/05/why-america-lost-in-afghanistan-counterinsurgency-cords-vietnam/.


Histogram of statements released on Afghanistan

Source: Documents compiled from Nexis Uni

Topic Modeling

To look at the distribution of topics over time, I ran vanilla Latent Dirichlet Allocation (LDA) on the corpus of the 8,281 documents. To do so, I used the `stm` package in R with $k = 10$ topics and extracted the parameter theta (composites versus topics matrix). After I fitted the model on the data, I used both the words with the highest probability within each topic (parameter β) and frequent and exclusive (FREX) to manually label each topic. In the cases where the highest probability words were not enough to determine a theme, I read the five documents with the highest probability for that specific topic.

I used the average theta score for all documents from a specific time period to look for the distribution of topics over time. For instance, to assess the distribution of topic models over months of the year, I averaged the theta scores for that specific topic for all documents that were released during that month of the year.

Sentiment Analysis

I also explored the change in sentiment over time. I used the `tm` package in R to remove capitalization and punctuation, remove filler words, stem the remaining words, and create a document term matrix based on a uni-gram model. I removed sparse terms (defined as uni-grams appearing in less than 2 percent of the documents) from the document term matrix. To account for the difference in document lengths, I normalized the matrix so that all documents had equal weighting regardless of word count.

Subsequently, I used a dictionary classification method using the dictionary of positive and negative words compiled by Neil Caren. I used the stemming function from the tm package to ensure that the stemming was consistent with the document term matrix, and removed all word stems that appeared in both negative and positive word lists. I used this word sentiment data set to compute a positive and negative weight for each document and calculate an overall score for each document through tallying the positive and negative scores.

To explore the change in tone and attitude over time, I averaged the document sentiment scores for the specified time interval (for instance, year or month). I looked at both the change between 2004 and 2019 and between the different months of the year (to assess whether there was seasonality). In order to ensure that seasonality was not influenced by the number of documents released in a given year, I also ran a year fixed effects regression on the data. I used the following model to isolate seasonality:

$$y_{it} = X_{it}\beta + \alpha_i + u_{it}$$

where

y_{it} is the sentiment score observed for individual document i at time t,

X is the month of the year,

β is the coefficient on the month,

t is year (time trend),

α_i is the intercept, and

u_{it} is the error term for individual document i at time t.

Findings

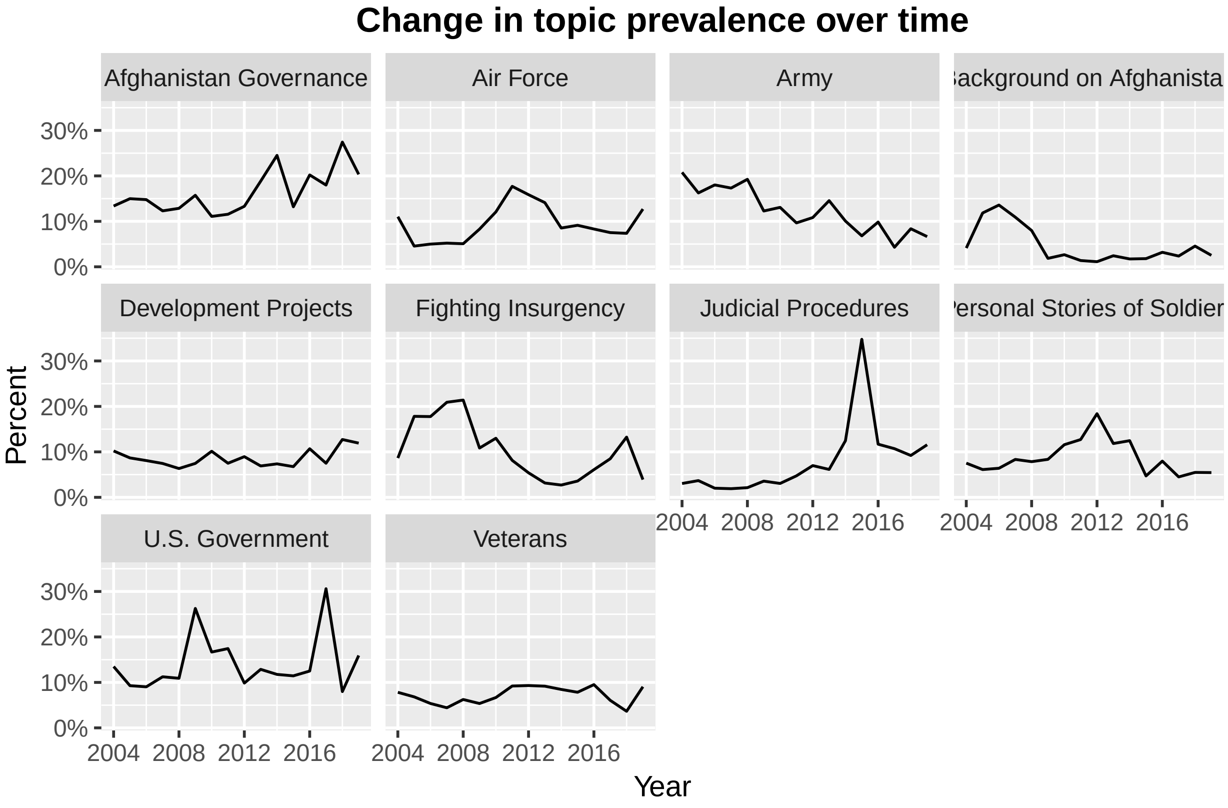
Topic Modeling

Running LDA with $k = 10$ topics yielded the topic areas summarized in the following table. The first column contains the manually generated topic titles for each topic cluster, while the second and third list the highest probability and frequent and exclusive (FREX) uni-grams, respectively.

Topic	Highest Probability	FREX
`Afghanistan Governance`	afghanistan, said, afghan, will, secur, govern, peopl	inaud, question, holbrook, think, secretari, reaffirm, reconcili
`Fighting Insurgency`	forc, afghan, coalit, provinc, kill, insurg, attack	cach, suspect, detain, small-arm, bomb-mak, milit, grenad
`Development Projects`	afghanistan, afghan, develop, will, program, project, agricultur	usda, agribusi, fas, workshop, prt, fulbright, usg
`Veterans`	servic, veteran, famili, serv, member, nation, war	iava, medal, ptsd, veteran, fmwrc, tbi, badg
`Personal Stories of Soldiers`	said, marin, soldier, time, one, work, will	rifleman, humour, sapper, laugh, lad, joke, ski
`U.S. Government`	afghanistan, troop, presid, iraq, contact, pleas, will	senat, bipartisan, rep, congressman, codel, committe, obama
`Air Force`	said, air, forc, oper, mission, support, afghanistan	airdrop, refuel, nmcb, harrier, sorti, amc, pallet

Topic	Highest Probability	FREX
Judicial Procedures	afghanistan, contract, state, investig, depart, offic, unit	attorney, guilti, plead, indict, mccaskil, conspiraci, briberi
Background on Afghanistan	afghanistan, nato, say, afghan, countri, govern, intern	scheffer, jaap, voa, mujahidin, pdpa, soviet, daoud
Army	forc, command, armi, afghanistan, oper, contact, soldier	conway, div, drum, —-, brigad, twentynin, bct

The following figure shows the change in topic prevalence over the years, highlighting a number of interesting findings that are listed in bullet point format below.



Source: Documents compiled from Nexis Uni

- While during the initial years (2004 to 2008), fighting the insurgency in Afghanistan and U.S. army were the most prevalent topics, over time the focus has shifted to governance in the country. Specifically, in 2014, when more than half of the U.S. troops left Afghanistan[3], the discourse was shifted from the military to governance in the country. The discussion of fighting insurgency, nonetheless, has picked up steam once again under Trump administration.
- The U.S. air force became a major topic of discussion between 2011 and 2013, perhaps as a result of a greater focus on airstrikes.
- The discussion of Army (ground fighting) has consistently waned over time, even despite the surge in number of troops between 2009 and 2014.
- During the initial years of the war, a lot of documents were released that contained historical and current background data on Afghanistan but the number has remained steadily low since.

[3] Associated Press. "A Timeline of U.S. Troop Levels in Afghanistan since 2001." Military Times, August 8, 2017. https://www.military times.com/news/your-military/2016/07/06/a-timeline-of-u-s-troop-levels-in-afghanistan-since-2001/.

- As expected, around the time of elections in the U.S. when the incumbent president is not running for office, the U.S. government becomes a central topic of discussion. This might be a result of increased debate on the decisions of the new administration on Afghanistan. A cursory look at the documents with the highest theta value for each of 2009 and 2017 is consistent with this hypothesis:

```
## [1] "Sen. Carl Levin, D-Mich., chairman of the Senate Armed Services Committee, sent
the following letter today to Sen. John McCain, R-Ariz., ranking member of the committee,
Sen. Joseph Lieberman, ID-Conn., and Sen. Lindsey Graham, R-S.C., in response to their
request for congressional testimony from senior military commanders responsible for
Afghanistan. Dear John, Joe and Lindsey: Thank you for your letter of September 18. I
agree with you concerning the importance of succeeding in Afghanistan and the need for
Congress and the American people to understand how the future of Afghanistan is linked to
our own safety here at home. At the present time, while General McChrystal has submitted
his assessment of the situation on the ground and his recommendations concerning the
strategy for Afghanistan up through the chain of command, he has not yet submitted his
recommendation as to the resources that he believes would be needed to implement the
strategy. I also understand that discussions on strategy are ongoing. Under these
circumstances I believe that it is premature to seek the military commanders' testimony
on their resource recommendations to implement a strategy before the President's senior
advisers, including Admiral Mullen and Secretary Gates, have had an opportunity to
provide their advice to the President relative to those recommendations."
```

```
## [1] "House Democratic Caucus Chairman Joe Crowley (D-NY) issued the following
statement on President Trump's remarks on the Afghanistan War. \"President Trump has no
strategy for ending the war in Afghanistan. Despite campaigning on a grand vision to
withdraw troops from this long-running military conflict, it is clear that President
Trump lacks a comprehensive plan to keep American troops out of harm's way. \"Tonight, I
had hoped for firm details on how the White House plans to extract the U.S. from this
volatile region, while still protecting American interests. Instead, we were presented
with a vague plan that will likely leave American troops in combat for years to come. The
Trump administration owes the American public, our military personnel, and the U.S.
Congress a clear answer on how it plans to end the Afghanistan war and a vision for its
newly-announced engagement with Pakistan."
```

- Finally, there is a sharp increase in the prevalence of the topic of judicial procedures in 2015. A closer exploration of the data shows that there were a flurry of activity from the U.S. Department of Justice in coordination with Special Inspect General for Afghanistan Reconstruction (SIGAR) during this time. The following are excerpts from the texts of the three documents that had the highest theta value for the topic:

```
## [1] "A Fort Campbell Army Sergeant pleaded guilty today to conspiracy to commit
bribery in connection with contracting for supplies while serving in Afghanistan.
Assistant Attorney General Leslie R. Caldwell of the Justice Department's Criminal
Division, Acting U.S. Attorney John E. Kuhn Jr. of the Western District of Kentucky,
Assistant Director in Charge Andrew G. McCabe of the FBI's Washington Field Office,
Special Inspector General for Afghanistan Reconstruction John F. Sopko, Director Frank
Robey of the U.S. Army Criminal Investigation Command's (CID) Major Procurement Fraud
Unit, Acting Special Agent in Charge Paul Sternal of the Defense Criminal Investigative
Service's (DCIS) Mid-Atlantic Field Office and Brigadier General Keith M. Givens,
Commander of the Air Force Office of Special Investigations (OSI) made the announcement.
Ramiro Pena Jr., 43, of Fort Campbell, Kentucky, pleaded guilty before U.S. District
Judge Thomas B. Russell of the Western District of Kentucky to a one-count information
charging him with conspiracy to commit bribery."
```

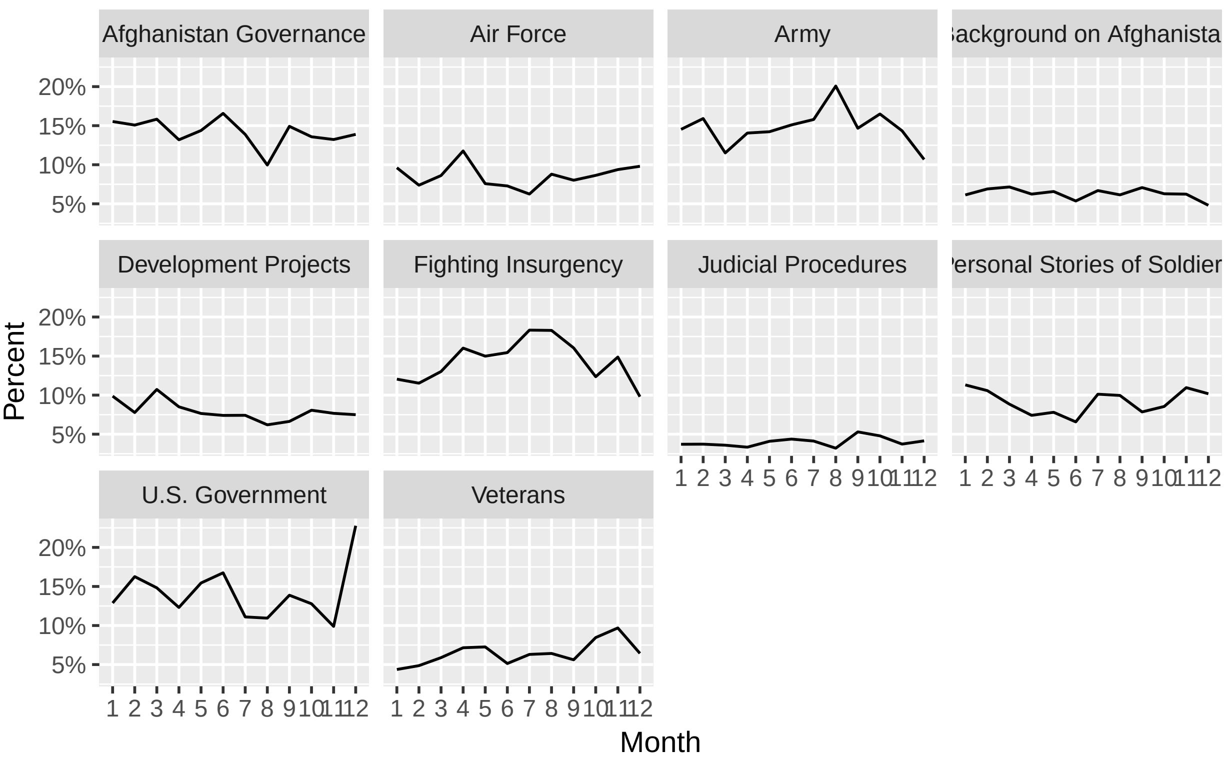
```
## [1] "A former specialist with the U.S. Army stationed at Forward Operating Base (FOB)
Gardez, Afghanistan, was sentenced today to 30 months in prison for accepting a $20,000
```

```
bribe from a truck driver in exchange for allowing him to take thousands of gallons of
fuel from the base. Assistant Attorney General Leslie R. Caldwell of the Justice
Department's Criminal Division, Acting U.S. Attorney Brian Stretch of the Northern
District of California and U.S. Attorney Michael J. Moore of the Middle District of
Georgia made the announcement. Anthony Don Tran, 28, of Stockton, California, pleaded
guilty on June 9, 2015, to bribery of a public official."
```

```
## [1] "An independent contractor for a trucking company in Afghanistan that was
responsible for delivering fuel to U.S. Army installations was sentenced to four years in
prison today for offering a U.S. Army serviceman $54,000 in bribes to falsify documents
confirming the receipt of fuel shipments that were never actually delivered."
```

A look at the change by month of the year shows some seasonality with the topics. For instance, during summer time which is concurrent with the annual Taliban offensive[4], there is more discussion around army and fighting insurgency. On the other hand, the U.S. government becomes the topic of discussion during the month of December. Given the findings of the previous section, this sharp increase can be attributed to transitions in the U.S. administration following the 2008 and 2016 elections.

Change in topic prevalence by month



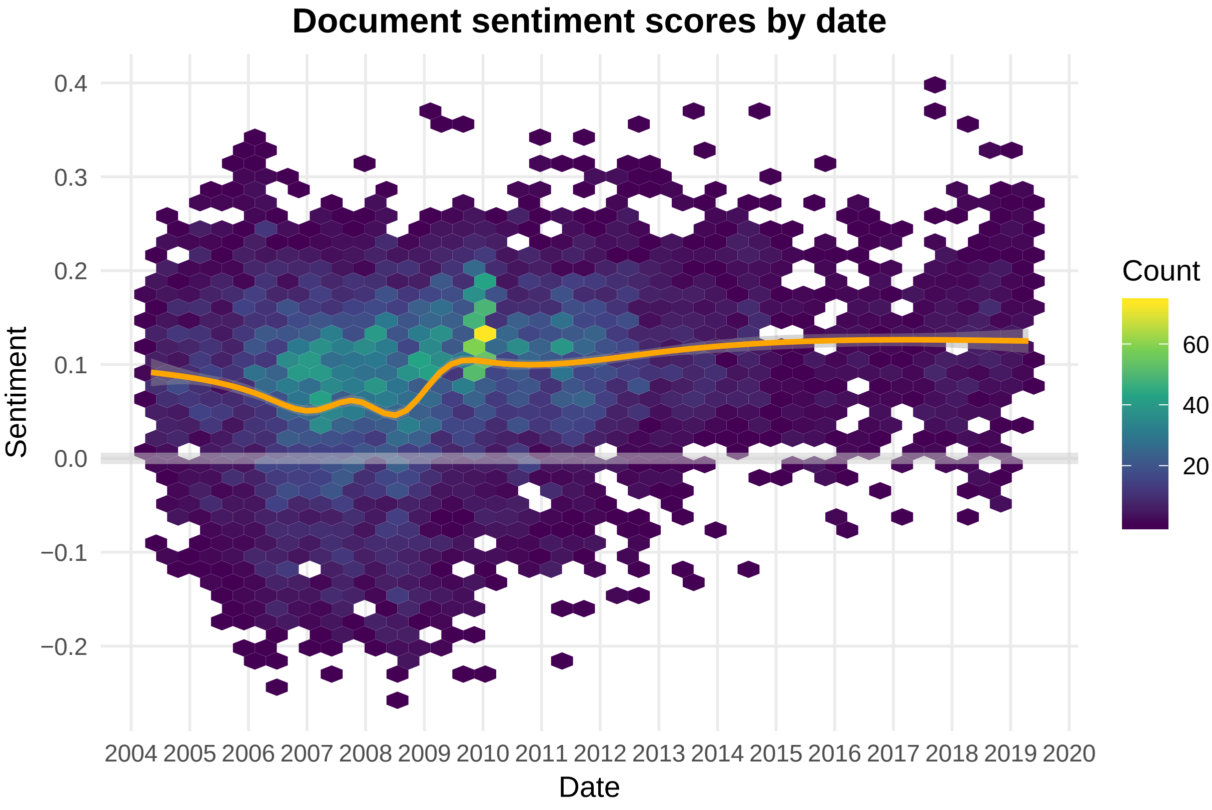
Source: Documents compiled from Nexis Uni

Sentiment Analysis

The following graph shows the distribution of document sentiments over time. The vast majority of the documents have sentiment scores above zero, indicating an overall positive tone. However, there is a lot of variation over time, especially between 2006 and 2010, when a lot of documents were published pertaining to Afghanistan.

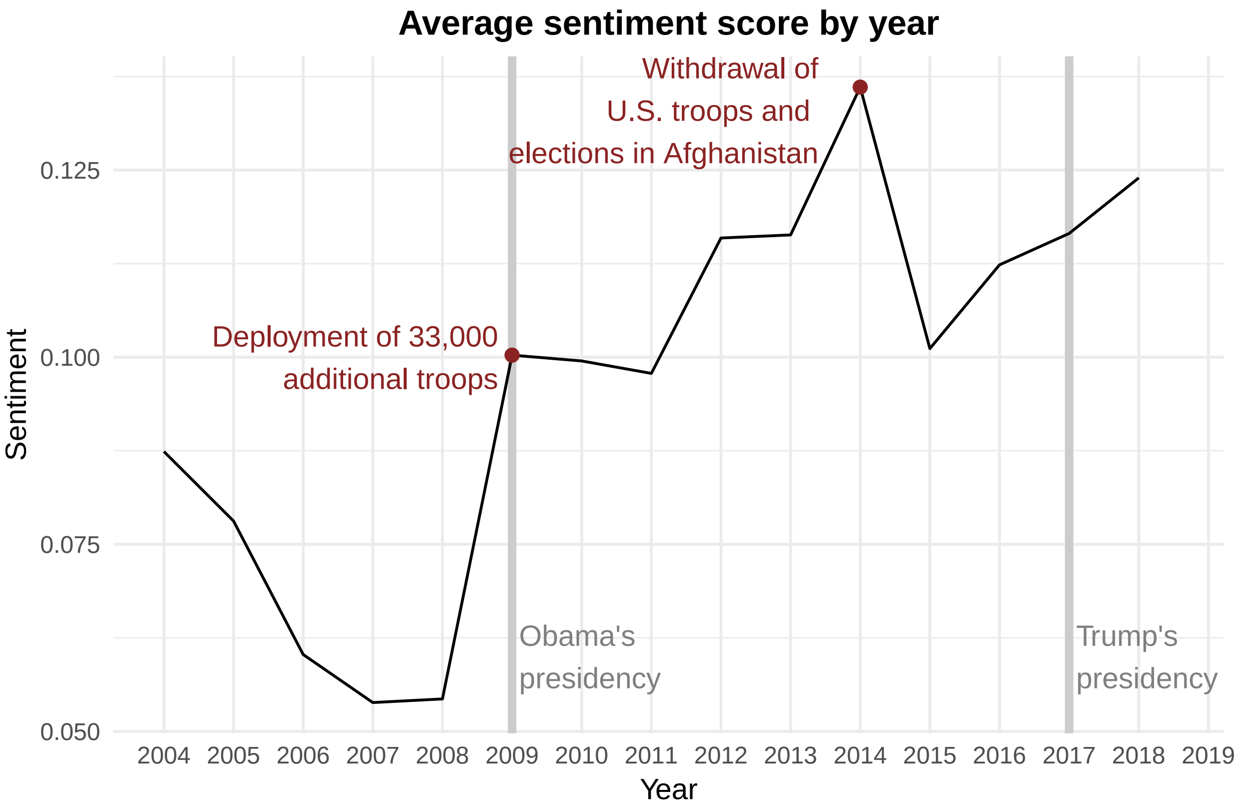
[4] Mashal, Mujib. "Taliban Announce Spring Offensive, Even as Peace Talks Gain Momentum - The New York Times." The New York Times, April 12, 2019. https://www.nytimes.com/2019/04/12/world/asia/taliban-spring-offensive-afghanistan.html.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Source: Documents compiled from Nexis Uni

To understand the variation in sentiment scores further, I looked at the average score for all documents for each year between 2004 and 2019 (Figure below).



Source: Documents compiled from Nexis Uni

This reveals a number of interesting findings. Under George W. Bush, the documents consistently took on a negative tone. The negative tone in the documents could be attributed to one or both of the following reasons: 1) The outlook on the Afghan war become more pessimistic over time, and 2) official documents focused primarily on the military side of the intervention in Afghanistan. In the latter case, words associated with military are mainly classified as negative rather than positive, regardless of which side is winning the war. A random sampling of the documents (n = 2) appears to support the second postulation.

```
## [1] " The U.S. Department of Defense's Combat Joint Task Force 82, Operation Enduring
Freedom, issued the following news release: Multiple militants were killed and one was
detained by Coalition forces during an operation to disrupt militant activities in Kapisa
province, Thursday. The force searched a compound in Tag Ab District targeting a Taliban
commander smuggling weapons and foreign fighters into Afghanistan, as well as organizing
suicide attacks against Coalition and ISAF forces. During the operation, armed militants
engaged the force. Coalition forces responded with air strikes and small-arms fire
killing the militants. For any query with respect to this article or any other content
requirement, please contact Editor"
## [2] " The U.S. Department of Defense issued the following news release: The Department
of Defense announced today the death of a soldier who was supporting Operation Enduring
Freedom. Pvt. Tan Q. Ngo, 20, of Beaverton, Ore., died Aug. 27 in Kandahar, Afghanistan,
of wounds suffered in Zabul Province, Afghanistan, when his mounted patrol received small
arms and rocket-propelled grenade fire. He was assigned to the 1st Battalion, 4th
Infantry Regiment, Hohenfels, Germany. For more information media may contact the U.S.
Army, Europe, public affairs office at 011-49-6221-57-5816 or 8694, or email: For any
query with respect to this article or any other content requirement, please contact
Editor at Load-Date: August 29, 2008 "
```

In 2009, There is a clear uptick in positive tone in the documents right after Barack Obama took over as the president of the United States. This positivity can perhaps be explained by one or two of the following reasons: 1) optimism stemming from the fact that he deployed 33,000 additional troops to Afghanistan[5], and 2) positive undertones associated with a focus on institution building, governance, and development as opposed to military intervention. With the deployment of additional forces in 2009, President Obama set a clear timeline for the war in Afghanistan, asserting that the U.S. would pull out its troops from the country by 2014. In line with his promise, in 2014, President Obama started the process of military withdrawal, ordering half of the troops to leave the country[6]. 2014 also marks the most positive year for documents in the data set. This might be due to the fact that the official documents by the U.S. administration tried to portray the withdrawal as the end of the war and the ability of the Afghan government to hold down the fort on its own against the insurgents. The following are two randomly generated excerpts from the documents released in 2009 and 2014, respectively, which support this hypothesis.

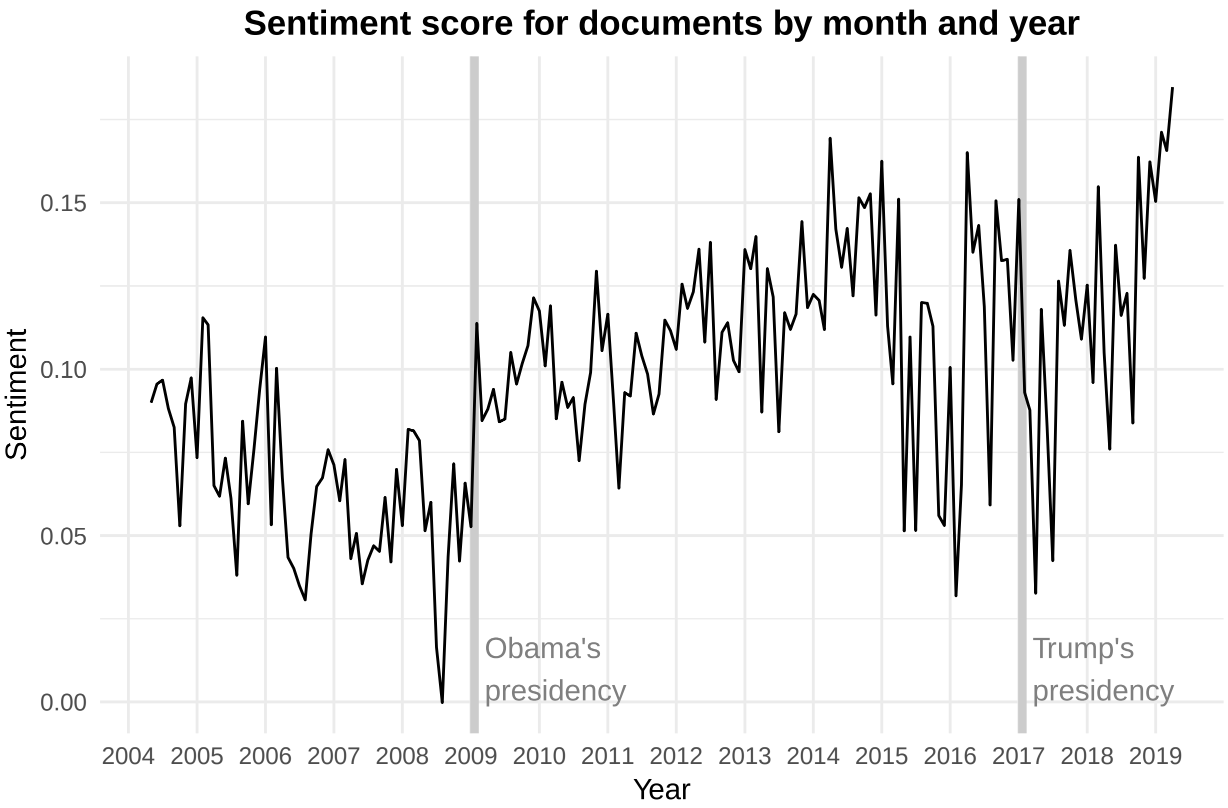
```
## [1] " The U.S. Navy issued the following press release: By Utilitiesman 3rd Class
Katelyn Knowles Det Kandahar Afghanistan Public Affairs While Naval Mobile Construction
Battalion (NMCB) 4's deployment is almost over, they recently received tasking to create
a contingency construction plan preparing sites for the arrival of NMCB 7 and NMCB 11. In
Kandahar, Afghanistan Seabees are renovating buildings and facilities for extra troops
and incoming battalions that will be arriving in the near future. One building being
renovated currently houses the battalion, but will be converted into the base exchange.
\"A lot of troops on this crew are new to battalion, and this project is great for on the
job training. They are really holding their own and excited to be a part of something so
important,\" said Builder 2nd Class Eric Pimentel, project supervisor."
```

```
## [1] "In this photo, Boehner points to a Hartzell propeller on a military plane in
Afghanistan. Hartzell maintains 75 percent of the world's market in airplane propellers
and is headquartered in Piqua, Ohio. The company is one of the largest employers in Miami
County. This week, in the wake of the pivotal first round of voting and in the midst of
the country's first-ever democratic, peaceful transfer of political authority,
Congressman John Boehner (R-West Chester) and a group of senior House Members visited
Afghanistan to assess the political progress in the country, the security situation, and
the transition of American forces after more than a decade. \"Since shortly after 9/11,\"
Boehner said, \"our troops have fought to bring peace and security to Afghanistan and to
ensure it can never again be used as a safe haven for terrorists to attack the United
States. Many Americans have sacrificed to secure these goals, and far too many have made
the ultimate sacrifice or suffered life-changing wounds in the past twelve years of
fighting. Now, the Afghans are poised to elect a new government for the first time in
their history."
```

A look at the distribution of sentiment scores by month and year corroborates the findings described above. There is a clear spike in optimism right after President Obama comes to office. The figure also suggests that the documents take on a more pessimistic tone right before and during the time of elections in the United States.

[5] Baker, Peter. “How Obama Came to Plan for ‘Surge’ in Afghanistan.” The New York Times, December 5, 2009, sec. Asia Pacific. https://www.nytimes.com/2009/12/06/world/asia/06reconstruct.html.

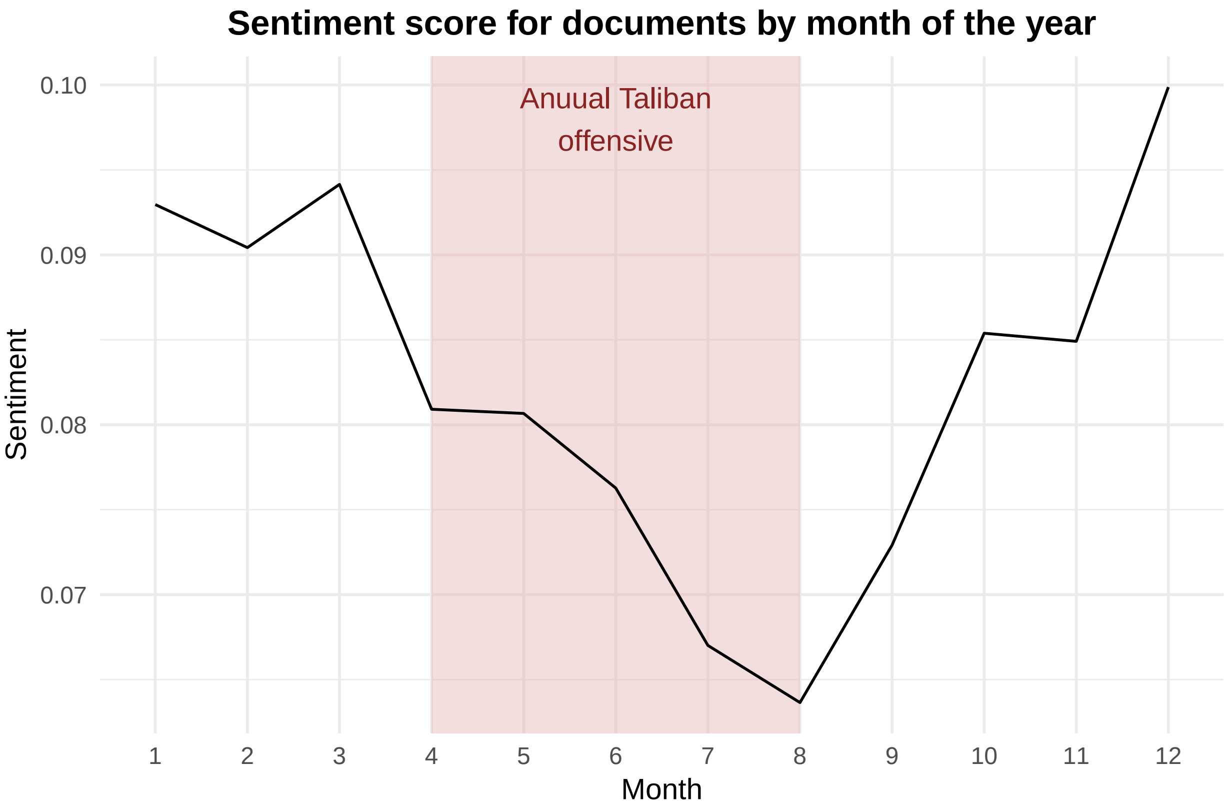
[6] Associated Press. “A Timeline of U.S. Troop Levels in Afghanistan since 2001.” Military Times, August 8, 2017. https://www.military times.com/news/your-military/2016/07/06/a-timeline-of-u-s-troop-levels-in-afghanistan-since-2001/.



Source: Documents compiled from Nexis Uni

The data also lends itself to the hypothesis that there is seasonality to the document sentiments (Figure below). During the annual Taliban Spring Offensive that falls between the months of April and August[7], the documents have a more negative tone than the other months. On the other hand, the sentiments are more positive during the winter season when there is less fighting between the two sides. The negative sentiments during the fighting season, however, are not determinant of which side is winning the war and merely indicate that the documents contain more negative words.

[7] Mashal, Mujib. "Taliban Announce Spring Offensive, Even as Peace Talks Gain Momentum - The New York Times." The New York Times, April 12, 2019. https://www.nytimes.com/2019/04/12/world/asia/taliban-spring-offensive-afghanistan.html.


Sentiment score for documents by month of the year

Source: Documents compiled from Nexis Uni

In order to ensure that monthly data is not influenced by the number of document published in a year, I ran a year fixed effects regression on the months. the following table summarize the results of the regression and confirm the seasonality in sentiments.

```
##
## Call:
## felm(formula = overall_score ~ month | year, data = document_scores %>% mutate(year =
as.factor(year(date_time)), month = as.factor(month(date_time))))
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.34790 -0.04928 0.00824 0.05691 0.27629
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## month2 0.001058 0.004791 0.221 0.82528
## month3 0.003997 0.004577 0.873 0.38258
## month4 -0.009327 0.004889 -1.908 0.05644 .
## month5 -0.011279 0.004948 -2.279 0.02267 *
## month6 -0.012109 0.004657 -2.600 0.00933 **
## month7 -0.023147 0.004733 -4.891 1.02e-06 ***
## month8 -0.027468 0.004691 -5.856 4.93e-09 ***
## month9 -0.015393 0.004684 -3.286 0.00102 **
## month10 -0.003791 0.004660 -0.813 0.41597
## month11 -0.005226 0.005009 -1.043 0.29683
## month12 0.006994 0.004650 1.504 0.13263
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08655 on 8254 degrees of freedom
## Multiple R-squared(full model): 0.09288 Adjusted R-squared: 0.09002
## Multiple R-squared(proj model): 0.01386 Adjusted R-squared: 0.01075
## F-statistic(full model): 32.5 on 26 and 8254 DF, p-value: < 2.2e-16
## F-statistic(proj model): 10.54 on 11 and 8254 DF, p-value: < 2.2e-16
```

The coefficients for the months of July and August, which are significant at 1% level, back the finding that the overall document sentiments become more negative during these months compared to the month of Jaunary.

Future Research

This research was designed as an initial exploratory exercise to inductively learn from the vast corpus of documents available on Afghanistan. Further research is necessary to delve deeper into each of the topics explored in this paper. Future studies, for instance, can look at whether the same sentiment patterns hold when using other dictionaries. In particular, using a dictionary that is more suited to military interventions can perhaps reveal which side has the upper hand on the battlefield. Moreover, there is a need for further analysis to assess how sentiment and topics change based on office and document type. These research projects could perhaps reveal important insights into the dynamics of the interminable conflict in Afghanistan.