

Erica Devine
History and New Media
April 15, 2019

**Topic Modeling *Foreign Relations*:
Implications for U.S. Policy in the Middle East**

The digital age has provided scholars of diplomatic history with immense possibilities to expand their research, including greater access to a diverse range of sources electronically at a cheaper cost. But it has also created significant challenges. The sheer volume of documents must be declassified, processed, and preserved, despite limited funding. Once documents are in the available to the public, there are too many for a research to read in their lifetime, let alone in the limited amount of time they have to complete a project. Finding the right sources is therefore important but more difficult in this expanded archive. A simple keyword search is not enough to mine the dense and complex digital archive.

A solution to these problems involves using digital tools to mine the vast archive. I argue that topic modeling allows historians to generate an index of a vast corpus of digital texts, locating important topics and themes without having to read through all the material. From the results of a topic model, scholars can locate which individual documents contain that topic and can how those topics change over time. Topic modeling is not an analytical tool in and of itself, but when used as an index tool and in conjunction with traditional scholarship, it can open up new avenues of inquiry. This essay will explore the historical context of the period under review, the historiography of topic modeling, my methodology, and finally the results of my analysis.

Historical Context

The early Cold War was a period of intense change in U.S. foreign policy. The transition from an isolationist foreign policy to an intensely interventionist agenda has been explored by many scholars of the Cold War. Some argue that the importance of foreign oil to national security led the U.S. government to actively secure access abroad.¹ Some argue that having fought in Europe and the Pacific, and been attacked at Pearl Harbor, erased the notion that Americans could remain isolated from an increasingly globalized world. Their position as the sole nuclear power, until 1949, combined with the “legacies of an activist state” during the New Deal, contributed to a growing sense of global leadership.² This dramatic shift was especially apparent in U.S. foreign policy towards the Middle East.

Prior to World War II, the Middle East was central to European imperial ambitions. Britain had established virtual colonies in Palestine and Transjordan, while maintaining significant control over Egypt, Sudan, Yemen, Iraq, and southern Iran. France laid claim to the mandates in Syria and Lebanon, while Russia exercised their influence in northern Iran and Afghanistan. During this time, the U.S. had limited interaction with the region outside of private missionary activity and commercial ties, mostly in the budding oil industry. The end of WWII and the beginning of the Cold War signaled a shift towards more government involvement in the region. The U.S. policy of containment, solidified in the Truman Doctrine in 1947, ensured that American foreign policy in the region would be defined in Cold War terms: maintaining access to oil, providing economic and military aid to American-backed regimes, and stifling home-

¹ See David S Painter, *Oil and the American Century: The Political Economy of U.S. Foreign Oil Policy, 1941-1954* (Baltimore: Johns Hopkins University Press, 1986); Irvine H. Anderson, *Aramco, the United States, and Saudi Arabia: A Study of the Dynamics of Foreign Oil Policy, 1933-1950* (Princeton: Princeton University Press, 1987).

² Odd Arne Westad, *The Cold War: A World History* (Basic Books, 2017), 102.

grown nationalist movements, especially ideologically communist or socialist, that threatened American interests.

A more interventionist foreign policy during the early Cold War resulted in the significant expansion of the foreign policy bureaucracy and a subsequent “archival explosion.”³ For historians of American diplomatic history, the sheer volume of sources produced during this time period and the impact of recent historiographical interventions has made archival research more challenging. A scholar studying the history of U.S. foreign policy during the early Cold War must consult more traditional documents from the State Department, the President, Congress, other government bureaucracies such as the CIA as well as international organizations such as the United Nations and the World Bank. Important historiographical interventions in the field have expanded the type of sources consulted to write diplomatic history. The Cultural Turn encouraged the incorporation of less conventional sources, such as travel literature, films, and material culture. The drive towards transnational history encouraged historians to explore sources in other languages and national origins to discuss expressions of U.S. power abroad. The result is that historians can no longer rely on a close reading of all the texts available on their subject of choice.

Historiography

Many scholars across a number of disciplines have explored this dilemma. Macroanalysis or “distant reading” as a methodological approach is a technique that uses digital technology to approach large-scale research in innovative ways. Franco Moretti and Matthew Jockers have applied this concept to their work in literary studies. Moretti, an Italian literary scholar, argues

³ David Allen and Matthew Connelly in David Allen, “Diplomatic History After the Big Bang: Using Computational Methods to Explore the Infinite Archive,” in *Explaining the History of American Foreign Relations*, ed. Matthew Connelly, 3rd ed. (Cambridge: Cambridge University Press, 2016), 78.

that a close reading of texts depends on a small corpus of documents, which limits the scholars' ability to address themes and tropes on a larger scale. He proposes computational textual analysis, which can aggregate and analyze a massive amount of data.⁴ In his book *Macroanalysis*, Jockers details how digital methods can be used to shift "from looking at the individual occurrences of a feature in context to looking at the trends and patterns of that feature aggregated over an entire corpus."⁵ From a methodological standpoint, macroanalysis or distant reading can unearth new forms of digital evidence and uncover new questions and avenues for exploration.⁶

Topic modeling is one method of distant reading that has the potential to inform historical research in new ways. Latent Dirichlet Allocation (LDA) topic modeling is a statistical model that identifies a topic or topics that appear in a single document. When applied to a large corpus of texts, it generates clusters of words that share a topic or theme and determines which documents share themes based on which words are likely to "co-occur in them."⁷ The number of topics generated each time is arbitrary and determined by the scholar. The standard is twenty but larger corpus should warrant a larger number of topics.⁸ Most significantly, topic models generate topics without human input: you do not need to have pre-existing schema to run the program. Once the topics are developed, it is the job of the historian to interpret the topics and

⁴ Franco Moretti, *Distant Reading* (New York: Verso, 2013), 48–49 Moretti is discussing the necessity of distant reading to study world literature, which has a corpus too large to perform a close reading of even half of the texts in this corpus. .

⁵ Matthew L Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana and Chicago: University of Illinois Press, 2013), 24.

⁶ Jockers, 10.

⁷ Sharon Block, "Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources," *Common-Place The Interactive Journal of Early American Life, Inc.* 6, no. 2 (January 2006), <http://www.common-place-archives.org/vol-06/no-02/tales/>.

⁸ Jockers, *Macroanalysis: Digital Methods and Literary History*, 123.

give it a label. Because this program operates without human input, the topics that are generated can reveal themes that were not immediately obvious through a close reading.

Historian Sharon Block, along with computer scientists David Newman at University of California Irvine, used topic modeling to explore the entire eighteenth century corpus of the *Pennsylvania Gazette*. She ran a topic model of a corpus containing approximately eighty-two thousand articles and advertisements from 1728-1800. Using her results, she walks readers through how digital analysis can be used as a tool alongside “the historians’ input and analysis.”⁹ Her research revealed new information about word frequency and word use. More recently, Cameron Blevins has explored the benefits of topic modeling as a historical methodology. In his topic model of Martha Ballard’s diary, a corpus of nearly 10,000 diary entries, he similarly identifies thematic trends and unseen patterns of word usage, which reveal information about the texts that could not be garnered from a close reading.¹⁰ For example, in the topic Blevins labeled death, words like “informd,” “hear,” and “heard” appeared alongside words like “morn” and “death,” indicating that death is often encountered through face-to-face interactions in the form of news. This would be difficult to determine through a close reading of individual diary entries but is revealed clearly with an expanded lens.¹¹

Topic modeling is without its flaws. Jockers discusses the benefits and drawbacks of this tool by examining a topic model of the Stanford Literary Lab’s collection of 3,346 books. Most of the topics developed by the algorithm were easily interpretable. However, some were either too general or incoherent. These types of topics that he calls these topics “black boxes” are often the product of unsupervised models. The machine generates topics with no concern for the

⁹ Block, “Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources.”

¹⁰ Cameron Blevins, “Topic Modeling Martha Ballard’s Diary,” April 1, 2010, <http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/>.

¹¹ Blevins, “Topic Modeling Martha Ballard’s Diary.”

interpretability of the information. What happens when the topic model produces topics that are too general or unclear? He argues that it is legitimate to disregard topics that are incoherent without compromising the integrity of the model.¹²

Similarly, Amanda Regan, a doctoral candidate at George Mason University, details the strengths and limits of this type of research. She argues that “algorithmic reading techniques such as topic modeling” can reveal unexpected trends and patterns or confirm suspicions and hypotheses, thus going beyond simple keyword searches used in most archives.¹³ However, she argues that the use of personalized algorithms to do this type of research pose ethical questions. In particular, personalized algorithms can carry methodological biases that are invisible to the ordinary user. Instead, she advises scholars to “streamline our research methodologies to fit with the landscape of digital sources continuously becoming available” as well as advocating for “openness and transparency in the methodologies and tools we use to navigate the digital environment.”¹⁴ The consensus among scholars is that topic modeling is an innovative methodological tool that can generate new fields of inquiry.

In the field of diplomatic history, computational analysis has generated some attention. In the third edition of *Explaining the History of American Foreign Policy*, David Allen and Matthew Connelly explore the role of computational methods in navigating a dramatically increasing foreign policy archive. These methods, according to the authors, “may offer the only hope of creating order from the chaos and producing anything like a proper finding aid” in

¹² Jockers, *Macroanalysis: Digital Methods and Literary History*, 128–30.

¹³ Amanda Regan, “Mining Mind and Body: Approaches and Considerations for Using Topic Modeling to Identify Discourses in Digitized Publications,” *Journal of Sport History* 44, no. 2 (2017): 174–75, <https://doi.org/10.5406/jsporthistory.44.2.0160>.

¹⁴ Amanda Regan, 174.

archives that are increasingly underfunded and swamped with new digital material.¹⁵ Diplomatic history lends itself well to computational analysis. The state department publications provide machine-readable text with rich metadata to the public. This article provides the readers with a brief example of a topic model of several *Foreign Relations of the United States* (FRUS) volumes related to the Kennedy and Johnson administrations' Vietnam policy. Unfortunately, the author does not explore or analyze the results but rather uses it to exemplify the possibilities behind this method. Topic modeling reveals the "hidden intellectual structures" in a corpus and can demonstrate how those structures change over time.¹⁶

Methodology

Here, I will take up David Allen and Matthew Connelly's call to utilize topic modeling to inform my research into U.S. foreign policy in the Middle East. I used MALLET, "a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text," created by Andrew McCallum at the University of Massachusetts, Amherst.¹⁷ My initial corpus included three volumes of FRUS covering the period 1951-1954 and consisted of of 1,922 State Department documents covering correspondences from Egypt, Iran, Syria, Lebanon, Jordan, Iraq, and Saudi Arabia.¹⁸ Each document within the three volumes was placed in an individual

¹⁵ Allen, "Diplomatic History After the Big Bang: Using Computational Methods to Explore the Infinite Archive," 83.

¹⁶ Allen, 90.

¹⁷ McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." <http://mallet.cs.umass.edu>. 2002.

¹⁸ *Foreign Relations of the United States, 1952–1954, Iran, 1951–1954, Volume X* (Washington, D.C: United States Government Printing Office, 1989); *Foreign Relations of the United States, 1952–1954, The Near and Middle East, Volume IX, Part 1* (Washington, D.C: United States Government Printing Office, 1986); *Foreign Relations of the United States, 1952–1954, The Near and Middle East, Volume IX, Part 2* (Washington, D.C: United States Government Printing Office, 1986). Three documents are related to Kuwait and Yemen.

word document, including the date and title but without footnotes.¹⁹ Once the data was imported into MALLET, I chose to run a forty-topic analysis.²⁰

Results

The results of the topic model were expected in some respects and surprising in others. Figure 1 contains a list of fourteen of the forty topics produced. I excluded some topics that had a high percentage rate but were too general. For example, topics labeled Embassies or Communications were passed over, despite their frequency in the corpus. Other topics were black boxes, like “tilocblob dsdb ocblob bud.” Taking Jockers advice, I passed over these topics in favor of focusing on “the most interpretable.”²¹

Topics such as the Cold War, the Arab-Israel Conflict, Defensive Organization, and Economic Aid were expected. That Cold War factors were a dominant theme during this time period, when the U.S. feared Soviet expansion into the region, was not surprising. Words like “free,” “communist,” and “support” speak to U.S. efforts to repress nationalist movement that the U.S. deemed a threat to regional stability and support militarily and financially those groups most likely to welcome western influence. Similarly, as the U.S. recognized the waning power of Britain in the region, efforts were made for the U.S. to replace them. To secure petroleum-producing regions, prevent the spread of communism, and establish regional stability, the Policy Planning Staff at the State Department proposed the creation of a Middle Eastern Defense Organization (MEDO) similar to NATO. MEDO failed, but forming regional defense

¹⁹ I excluded the footnotes because the footnotes often contain reference to other documents or the editor’s own comments. I wanted the data to contain only the words contained in the messages themselves.

²⁰ As mentioned above, this number is arbitrary. However, all the secondary sources cited in this paper used forty as their default number of topics to generate.

²¹ Jockers, *Macroanalysis: Digital Methods and Literary History*, 129.

organizations continued to be a primary goal of U.S. foreign policy, resulting in the Baghdad Pact in 1955 and Central Treaty Organization (CENTO) in 1958.

The topics I labeled Military Aid, Economic Aid, and Modernization Efforts speak to the implementation of modernization theory in the Middle East. By granting underdeveloped nations economic aid and technical assistance, those countries would soon be able to develop in the same way as the developed Western countries. Countries that were modernized would be more stable and less likely to turn to communism. Similarly, military modernization went hand in hand with economic aid, as both an enticement to remain aligned to the West and to build up the military might of Western allies.

While some topics were expected, the topic model also drew my attention to facets of those topics that I would have overlooked during a close reading of the texts. For example, the concern for British influence throughout the corpus indicates that Britain was still very much a factor in Middle Eastern affairs during these years. The U.S. was competing for influence in Egypt, Iraq, and Iran over oil and the Suez Canal.

One topic that appeared frequently was oil. Eight of the forty topics generated were related in some way to oil: Saudi oil, Iranian oil, the oil industry, oil agreements, oil transportation, etc. In the topic Iranian Oil, words like “British” and “settlement” indicate American concern over Prime Minister Mossadegh’s nationalization of the Iranian oil supply, at the expense of the Anglo-Iranian Oil Company (AIOC). The inclusion of words like “world” and “future” are particularly telling. At first, they do not seem to fit in to the overall category. But the facts that they are frequently found close to the other words in the category speak to how the U.S. viewed maintaining access to Iranian oil. Conversely, discussions about Saudi Oil were centered on trade agreements between the Saudi government and non-American shipping

companies. U.S. efforts during this time were applied towards establishing an American monopoly over the production and transportation of Saudi oil.

The topic model identified two topics related to coups that occurred in the Middle East during this time period. The first took place in Egypt in 1952. A group of army officers led by Mohammed Naguib and Gamal Abdel Nasser overthrew the British-backed monarch, King Farouk, and established a republic predicated on Arab nationalism and anti-imperialism. The second coup was sponsored and funded by the CIA to overthrow the democratically elected Prime Minister Mohammed Mossadegh in Iran after he nationalized the Iranian oil industry.²² What is interesting here is the presence of “tudeh” in the topic labeled Coup in Egypt. The Tudeh Party was an Iranian Marxist-Leninist party that is most known for supporting Mossadegh’s nationalization policies. Their presence in Mossadegh’s coalition was a significant factor in the American decision to stage a coup, fearing that Mossadegh was actually a communist. Associating the Egyptian revolution with a communist party indicates that U.S. diplomats were concerned about communist influences in Egypt as well.

The most striking aspect of this topic model is the absence of religion. Recent scholarship has focused on the role of religion in U.S. foreign policy in the early Cold War.²³ In particular, these scholars have explored the religious rhetoric adopted by figures such as President Eisenhower and John Foster Dulles in relations with Arab leaders, indicating a strong belief that appealing to the Muslim world would combat “godless” Soviet inroads in the region. However, words relating to religion were not statistically significant in this corpus of texts. It would be interesting to see if this changes when the time frame of the corpus is expanded.

²² See Chapter Six; Nikki R Keddie and Yann Richard, *Modern Iran: Roots and Results of Revolution* (New Haven: Yale University Press, 2006).

²³ See Andrew Preston, *Sword of the Spirit, Shield of Faith: Religion in American War and Diplomacy* (New York: Knopf Doubleday Publishing Group, 2012); William Inboden, *Religion and American Foreign Policy, 1945-1960: The Soul of Containment* (Cambridge University Press, 2010).

*I'm struggling to read the document that allows you to determine which document contains which theme. The results section so far contains my first impressions, but I want to demonstrate how I would use this by bringing in the primary source material itself.

*I want to add a mapping component, but realized too late that I didn't label my docs properly and can't use this feature. I will try to add this for the final draft.

Conclusion . . .

Figure 1

% of the Corpus	Words in a Topic	Human Added Topic Label
38%	secretary stated felt ambassador question conversation matter asked replied byroade thought general page expressed subject made pointed problem hoped make	Diplomacy
21.8%	economic soviet political security area forces western military support world policy free measures strength influence communist internal ussr west power	Cold War
21.4%	british foreign office eden negotiations london ambassador proposal kingdom point position united embassy department talks meeting government proposals agreed	British Involvement
12.8%	military aid assistance program defense equipment arms egypt grant department economic mutual security training state list act items supply million	Military Aid
12.1%	united states government kingdom nations east general assistance world respect prepared assist economic great free security terms rights steps direct	Rhetoric
11.4%	israel arab states peace arabs jerusalem eban israeli policy aviv area sharett relations foreign tel ambassador state egypt israelis aid	Arab-Israeli Conflict
11.1%	defense east middle states arab area military organization arrangements powers medo u.s regional planning participation turkey western forces arrangement conference	Defensive Organizations
8.8%	million loan aid economic bank funds program assistance financial year development projects foreign exchange fiscal grant ibrd debt technical request	Economic Aid
7.6%	iran iranian oil shah aid government british henderson kind army make public give problem present future world financial settlement	Iranian Oil

	country	
4.6%	oil compensation company iran mosadeq arbitration aioc court iranian henderson international million contract proposals dmpa terms tehran law claims payment	Coup in Iran
4.5%	naguib army rcc officers tudeh general regime coup maher party august elements control embassy forces staff ali caffery july cabinet	Coup in Egypt
4.3%	jordan syria syrian israel water development project plan johnston waters valley river yarmuk land shishakli zone economic damascus dam moose	Modernization Efforts
4.1%	refugees arab palestine refugee general united lebanon relief assembly nations commission resolution problem resolutions resettlement compensation unrwa program	Palestinian Refugees
2.8%	aramco saudi onassis agreement oil arabia sag department company arabian concession companies duce area king disputed government davies tankers operations	Saudi Arabian Oil