Making sense of text with topic models: Using computational methods to reveal latent document groupings in a corpus

Simon Rodier

Concordia University

Abstract

This paper examines the use of topic models - a computational method for uncovering latent groupings of documents in a corpus - as a tool for a human analyst to make sense of texts in a large corpus. The analytic methodology is presented and then applied to a reasonably large corpus; large enough to have the computational method applied, yet restrained enough for a human analyst to read and interpret. The results of the analysis are presented and then compared to a human interpretation following a read-through of the text. Significant overlap is found between the two interpretations, suggesting that topic models can be used to infer interesting information about a corpus. Further avenues to extend this method are presented, which will form the basis of an upcoming project.

Introduction

This project is motivated by a need to bring additional computational tools into qualitative analyses of social media texts. While traditional qualitative methods may bring great insight, the sheer quantity of text produced on social media platforms is overwhelming. This project utilizes topic modeling, a computational method which takes a corpus as input, divides it into a set number of topics, and outputs two particularly interesting distributions. First, the algorithm determines a probability that a word from the corpus is representative of each topic, and second, for each document in the corpus, it determines the likely composition of that document from the topics available in the corpus.

This method may be particularly useful in helping to inform qualitative analyses. These analyses often rely on analysts reading texts thoroughly to determine recurring, relevant themes. While the methodologies may be effective, they require significant investments of time, and may be difficult to apply to large corpora. As we will see, an automated method like topic modeling can help to quickly uncover regular statistical patterns within the documents that comprise a corpus.

Background

Topic Modeling. A topic model is a statistical model that describes abstract concepts ("topics") that are latent in a corpus of documents (Blei & Lafferty, 2009). Topic modeling is the process of uncovering the latent topics. This process takes as input a collection of document and generates a list of probabilities that each document was generated from each of a list of potential topics uncovered in the corpus. Furthermore, the *topics* predicted by topic modeling are each described by the most common words found in that topic in the corpus.

The Latent Dirichlet Allocation (LDA) algorithm by Blei, Ng, and Jordan (2003) is one of the standard ways of automatically finding topic models for large corpora. Fundamentally, the algorithm assumes that a corpus is generated by a set of k topics, where each of the topics is associated with some set of words. The algorithm then assumes that each document in the corpus is a mixture of these k topics, and thus contains words from each of the topics which generated it. We can illustrate this with an example: if one of the categories in a corpus is *dessert*, the words associated with the category might be *cake*, *bake*, *pastry* and *fruit.* Another category, *pets*, could be associated with the words *dog*, *cat*, *play* and *companion.* A document might then be generated from one or both of these categories using the words that define the category. If we see more words like *bake* and *cake* in the text, we might assign a greater probability to the document of being generated by the *dessert* category. If the document contains few words from the *pet* category, we might assign a low probability that the document was generated primarily by the *pet* category.

With these assumptions, Blei et al.'s (2003) algorithm not only generates a list of probabilities that each document is generated from each topic, but also determines the list of words associated with each of the k topics, which can be of help to a human analyst to determine the semantic significance of each of these topics.[1]

While the above assumptions may seem like oversimplifications of the processes by which documents in a corpus are *actually* generated, we can look at other applications of topic modeling to convince ourselves that these simplifying assumptions still lead to usable results.

Wallace (2012) used an LDA-based approach to parse out the major narrative threads of David Foster Wallace's novel, *Infinite Jest.* Paul and Dredze (2012) found that applying topic modeling to posts on a drug discussion forum could help to automatically infer information about the drug discussed in a post, the associated delivery method (e.g. ingestion, vaccine) and some aspect of drug use (i.e. health, culture or effects). Gerber (2014) uses topic models, augmented with geolocation data of tweets

[1] In this project, we will be working specifically with the implementation of the LDA algorithm provided by the Sci-Kit Learn library in Python (Buitinck et al., 2013).

from Chicago and historical crime maps from the city's available data, to assist in predicting future crimes. Althoff, Danescu-Niculescu-Mizil, and Jurafsky (2014) used topic modeling on posts from the sub-reddit *Random Acts of Pizza* to determine that the narrative behind a request for a gift of free pizza was a significant predictor of the success of that request. Of particular interest to this project, Baumer, Mimno, Guha, Quan, and Gay (2017) found that topic models can converge with grounded theory qualitative analyses; although the models may require human interpretation, contextualization and combination to match the themes uncovered by grounded theory analyses.

Methodology

This study uses a subset of the data collected by Project Someone for the Words in Context (Venkatesh et al., 2019) project. Words in Context used Corpus-Assisted Critical Discourse Analysis to uncover and analyse discursive patterns in conversations happening across popular social media platforms - namely Reddit, YouTube, Facebook and Twitter.

This study uses data captured from the social media website Reddit [2]. Reddit is divided into 'sub-reddits', which each essentially act as social media platforms catering to a specific community or interest. Reddit users can post content to a subreddit, and other Reddit users can interact with that content, voting the content up or down on the site (which influences its visibility), as well as comment on it. These comments are displayed as conversation trees, allowing readers to see what each post on the site is responding to. Also, Reddit's organization into particular communities and areas of interest helps us find publicly available and relevant conversations.

As this study is intended as a first experiment into using topic modeling to

[2] https://www.reddit.com

analyse social media texts, it uses a much restrained subset of the Words in Context dataset. Whereas the Words in Context Lebanese corpus included Reddit data spanning 27 different threads, this study samples a single thread consisting of 2,399 individual documents (or posts). This ensures that we have a substantial amount of text, while also keeping the corpus at a small enough size that a meaningful comparison can be made between the algorithm's results and a human reader's interpretation. Reddit is a particularly useful site for this first foray into topic modeling, as its users post publicly, and the data can be collected programmatically through Reddit's publicly available Application Programming Interface (API) [3].

Analysis

In this section, we will describe how we go from raw data to interpretation of the data using the topic modeling algorithm. We examine the preprocessing stage, which transforms the data from its raw state to a machine-interpretable form, and then the analysis stage, where we use the topic modeling algorithm to glean insights into the text.

Preprocessing

A key step in any computational analysis of text is *preprocessing*, where we take the raw text and standardize it as input for the algorithm. We begin by *tokenizing* the text — that is, separating each document into a list of tokens (individual words) — taking care to remove *stop words*, words like *the* and *and* that commonly occur in all texts and would not be informative to our analysis. We also transform all words into lowercase so that we do not consider capitalized versions of words as different from a lowercase variation. We also only collect tokens consisting of letters.

[3] `https://www.reddit.com/dev/api/`

Analysing corpora

It should be noted that the LDA topic modeling algorithm requires the user to specify the number of topics present in the text. Clearly, this is counter to our aims: our goal is rather to uncover an unknown number of latent topics within the text. Therefore, we will perform a series of analyses, running the topic modeling algorithm for differing numbers of topics, and then examine the best results for each. For each run of the algorithm, we also experiment with removing common and rare words beyond the generic stopword list discussed in the preprocessing section. On each run, we experiment with removing common words that occur in at least 40%, 50%, 75% and 90% of individual posts. We also experiment with removing rare words that occur in less than 15, 20, 25, 50 and 100 individual posts. We select the topic model with the best *perplexity* score (a measure of how well the model predicts a sample of data).

We run the LDA topic modeling algorithm on the corpus a total of 380 (19 different numbers of topics ranging from 2 to 20 × 4 values for common word removal × 5 values for rare word removal) times, on each run assuming a different number of topics. For each run of the algorithm, we examine the words best representing each proposed topic, as well as the documents (thread posts) deemed most representative of each topic. From these, we form an initial interpretation of the topics and select the topic models with the most interpretive salience. Because the corpus has been restrained to a size that can be reasonably read by a human, we then read over the thread. This read-through is not meant to be an extensive qualitative analysis, but merely to provide context to our previous interpretation. From this, we will not only have created an interpretation of the text using automated means, but will also have a short, subjective evaluation of the output in context.

Discussion

In this section, we examine the findings of the analytic methodology on our dataset. We begin by discussing the findings of the method applied to a restrained, human readable corpus, and then discuss the findings in light of having read the thread post-analysis.

Topics Found by the Algorithm

The runs of the algorithm that found the most interpretable topics were with the numbers of topics k set to 2, 13, 14. The topics found with k set to 13 and 14 are rather similar, so in this analysis, we will restrain ourselves to examining the results with $k = 2$ and $k = 14$.

With the number of topics set to two, the algorithm performs an effective broad delineation of documents into two categories. In order to illustrate the found themes, we can examine the algorithm's outputs: first, the words found in texts that are particularly common for the topic, and second, the individual documents in the corpus that best represent the topic. Table 1 displays the most representative words for topics 1 and 2. While a theme for topic 1 does not immediately emerge by scanning its list of words, the list of words for topic 2 does immediately suggest a militaristic, Middle Eastern theme. We can try to deepen our understanding of these themes by more closely examining the most characteristic posts for each topic.

Let us examine the top 5 posts for each topic. (A quick note to the reader: the > symbol at the beginning of a line represents a line quoted from another source, be it another post in the thread, or from outside reddit completely).

The most representative documents for topic 1 include:

Rank	Topic 1	Topic 2
1	just	saudi
2	like	lebanon
3	oil	israel
4	people	war
5	don['t]	iran
6	think	arabia
7	know	hezbollah
8	country	saudis
9	really	iraq
10	world	military

Table 1
The ten most significant words for topics 1 and 2 when the algorithm is run with number of topics set to 2.

1. > I don't think you can.

You certainly can. Actually, if you look closely, you'll realize it'd be stupid to forbid people from office simply because they are citizens of another country. The problem is this: every country has its own laws that our own country gets no say over, and those foreign laws often say that many of our own citizens are also citizens over there, even though they may have never chosen to be so or even know about it.

One practical example is that many Korean-Americans have traveled to South Korea under their American passport, only to learn that they were also South Korean citizens and are then required to enroll in obligatory military service. [This story is typical](http://narrative.ly/how-one-american-citizen-was-forcibly-drafted-into-the-south-korean-army/):

> Young Chun looked at the letter dumbly. It was all in highly formal Korean, a language he barely understood. But there was no mistaking the second sheet of paper contained in the envelope that had arrived

at his apartment that morning. It was a notice from the Department of Justice, written in both English and Korean. The young American, offspring of naturalized Korean immigrants, was barred from leaving South Korea.

>

> Overcome with dread, Chun now knew what the first letter meant. It was his draft notice for South Korea’s mandatory military service.

>

> "It’s scary, and the same time it’s like, there’s no way this is true," Chun says twelve years later, remembering the moment in 2003 that would seal his fate as an American citizen forcibly drafted into the South Korean military.

>

> It wasn’t supposed to be this way. Chun had only come to the country with the plan of teaching English for a year, a seemingly easy way of making a dent in his mounting credit card debt.

>

> Unbeknownst to him, Chun was a dual Korean citizen. More than two decades previously, a family member — who, he’s still not sure — added his name to the family register then used in South Korea to determine citizenship, incorrectly listing his place of birth as Seoul. Like all able-bodied Korean men, Chun would be obligated to serve two years in the military.

Some countries (like the USA!) have laws that if you are born abroad to citizens of that country, you are automatically a citizen as well. So many natural-born Americans of immigrant parents are automatically dual citizens—and if they've lived their whole life in the USA they may not even know it.

Here's another quirk: some countries do not allow parents to renounce their childrens' citizenship on their behalf. In fact, USA law is like that. If you're a minor your parents can go to a foreign country and follow all the procedures in that country's laws to make all of you a citizen of that country, even up to the point of renouncing their own US citizenship. The USA will recognize the parents' renunciation but not their minor children's, who are now dual citizens, like in the previous cases through no choice of theirs.

So any law that disqualifies persons from office simply because *another country's laws* say that person is a citizen of that country is a bad law. Any such prohibitions really should be based on *voluntary* actions by the person who is a dual citizen. (Doc. 291)

2. To be fair if Smith and Wesson produces fire arms and supplies the gang member are they responsible? It's sort of a similar situation, previously there were nefarious reasons for US arming them when Obamas admin was running the show, but it isn't always the case.

http://amp.washingtontimes.com/news/2017/jan/10/obama-hoped-to-use-isis-as-leverage-against-assad-/

http://www.telesurtv.net/english/amp/english/news/Kerry-Leak-Shows-US-Let-Islamic-State-Grow-to-Leverage-Assad–20170107-0012.html

The secretary of his admin was all up in it too http://amp.nationalreview.com/article/438605/hillary-clinton-benghazi-scandal-arming-syrian-rebels

Hell the hero of the whole Benghazi embassy incident (What the movie 13hours was based on)Kris Paronto, who's lucky to be alive after Obama admin shenanigans calls Hillary out on that non sense regularly but media will never give it any real attention

http://www.tactical-life.com/news/kris-paronto-hillary-clinton-twitter/

https://www.mediaite.com/online/benghazi-survivor-responds-to-hillarys-tweet-on-hurricane-damage-wish-you-had-sense-of-urgency-in-libya/amp/

http://www.thegatewaypundit.com/2017/09/army-ranger-kris-paronto-destroys-hillary-clinton-viral-tweet-anniversary-benghazi-attack/

I digress though; point is I wouldn't necessarily blame the US as a whole but more so the former administration, and seeing as the past few admins going back to jimmy carter have all had relatively the same agenda at hand I can see how people would like to blame US as opposed to the criminal elements whom highjacked it. (Doc. 794)

3. The people who ask for sources can't critically think or don't have enough understanding of the world to talk international politics. Even if you gave them sources, they wouldn't read them and continue to dispute you. Even most sources are complete garbage.. /r/askhistorians can't even have threads anymore because they can't find any sources.

   It's a way for them to look like they're adding to the discussion without adding anything to the discussion. Like they belong in the class without doing their homework. It's a self-gratifying ego trip, "look at me, I'm a truth seeker and a good one at that, I'm gonna need a source, buddy or you got nothin'!"

A person's analysis of a situation can be verified or discredited if you do your own research or have been paying attention the last twenty years. You can entertain their analysis without fully committing to it being true. You don't have to accept anything as truth but hearing all possible options and opinions, you have yourself a solid foundation of the situation.

I.e. Fuck pretentious redditors (Doc. 713)

4. The only power anyone has is that recognized by those around them. If the US is the only power that matters, then it alone is the player that determines what happens in the world, and what it wants, is decided for the world. If it's "Western Culture" or the UN, then so be it. It's not a unanimous decision made by every person on earth. Either the powers that exist allow something or challenge something.

When I say 'the world has decided' I don't mean that every person likes that decision. But every person with the power to matter in the decision has been summed up. Power is disequal among nations, as it has been since time immemorial. In the time of Alexander it was nigh his will alone that made decisions. In the time of Genghis Khan it was that of his influence. The Romans ruled how they wished. Britain controlled most of the world with colonialism for it's time.

The way the world decides comes down to who holds the power to decide. I don't believe the US alone makes that decision, but it and it's allies certainly seem to hold all the power they need to make near unilateral decisions, within certain bounds, on behalf of the world. That's the realistic decision matrix that matters.

If you don't like that, lead the opposition, and see who the world decides to be the best option between the two.

Yes, this is bleak, but it is realistic. Social change is difficult if those with the power don't care. Something to keep in mind. (Doc. 159)

5. you're failing to grasp the point. I'm sincerely racking my brain to think of how to be more succinct. let's try this:

ok, so, let's say I grab the finest, most titled, doctor they have at the hospital (per your example). however, he's actually a serially negligent Dr. & performs malpractice. hmm, what did his title do for me that day?

You are using titles as a "filter" in your example. I'm not asserting titles are worthless, but they are no guarantor for success. or truth. or fact. or whatever applies in the scenario you're in.

I read the clown's post. it wasn't trash, but it was flawed; I have background in this issue that allowed me to know why his conclusion was false. I pointed out specifically how & why. he had no response & just fell back to elitism & condescension. he failed. then you jump in. I'm taking for granted you have some prior relationship with him, but I really don't care.

at the end of the day, his/your analysis fails to account for USA debt holdings. it's ideas that matter. not our titles. do some research build on what you have. but your conclusion is false due to an arbitrarily narrow lens. (Doc. 1524)

Even reading through each post individually, it is difficult to a pinpoint a specific "topic" that arises naturally. We can however, see a common thread linking these posts together. They are all arguments meant to persuade a reader. Let us look, in contrast, at the top five representative posts of the other topic.

1. So Israel is going to let Saudi Arabia, a rising military power, send its frigates, corvettes and missile boats into Israeli waters. It will then watch as this rising military power lands troops in and occupies Lebanon? That completely flies in the face of Israeli military doctrine, cedes sovereignty over its security sphere, enables another military power to establish land based military assets on its opposing border. There is no way that would ever, ever happen. I doubt even Turkey would allow it, as they are also vying for influence in the Muslim world. Remember, Israel's military doctrine is based on having the superior army over all Arab nations AND being so strong that no Arab country would ever think of invading them. Allowing Saudi Arabia to exert military control over Lebanon would be a death knell for any Israeli politician and it would be a major security concern for Israel. This is a fantastical scenario my friend. (Doc. 551)

2. show me the historical precedent of Israel allowing an Arab army bombard a country near its borders? Israel allowing Saudi Arabia to bomb Lebanon would concede that Saudi Arabia is a military power in Lebanon, something that is ridiculously foolish for the Israeli regime. If the war goes badly and Saudi Arabia wants to put troops

in Lebanon? Israel will never, ever let an Arab power launch a war near its borders, even if it is against Israel's enemies. (Doc. 520)

3. Why on Earth do you think Saudi needs to go through Iraq to get to Lebanon? Iraq is to the East of Saudi and Lebanon is to West.

   Saudi has Jordan, Israel and Egypt on its side. It doesn't need Iraq and it doesn't need Syria.

   Jordan will allow Saudi to do whatever. Jordan depends entirely on Saudi, Israel, and US for its security. And Israel will jump on any chance to damage Hezbollah. And Egyptians HATE Iran. They fought them during the Iraq-Iran war and they're eager to fight them again. (Doc. 584)

4. why would Israel let an Arab power encroach upon its military hegemony in the region? why would Israel signal that Saudi Arabia is able to extert military power in Israel's security sphere? Is there a historical precedent for that? It is one thing to cooperate with Saudi Arabia diplomatically and covertly, it is another to allow Saudi fighter jets bomb another country at Israel' borders, gather intelligence, etc. It is profoundly stupid for Israel to ever allow such a thing. (Doc. 993)

5. No he was not an Iran proxy. He is Sunni and has Saudi citizenship FFS. He's a Saudi puppet and they ordered him to resign. Saudis and Israel want to start a war in Lebanon, having failed to oust Assad in Syria. They think Hariri resigning will throw Lebanon into chaos, giving Israel its "legitimate security concerns!!!!omg!!!" and its pretext

> for another one of its defensive preemptive invasions of Lebanon.
>
> Assad is allies with Iran. Hezbollah in Lebanon has support from Iran as well. So far, Lebanon hasn't taken the bait. (Doc. 1829)

These posts are clearly all argumentative as well, although here we may note one important difference. A topical theme does begin to emerge: posts that clearly and specifically address Middle Eastern politics and the relationships between different states.

For both topics, these differences persist as we continue drilling down through the list of representative posts. We may therefore begin to conclude that the topic modeling algorithm has sorted the posts into two groups based on the language used within them: those posts specifically discussing Middle Eastern politics and conflict (topic 2), and those posts arguing other issues (topic 1).

We can then shift our attention to the algorithm's results when the number of topics is set to 14 (Table 2). While not all of the topics seem immediately discernible from their word lists, some do present clues. Topic 2 seems to hint at citizenship and history; topic 4 contains several country names (and hints mostly at the Middle East); topic 6 hints at war and inter-state relationships; topic 9 seems to discuss countries in the Middle East; topic 10 appears to be very clearly discuss a range of Middle Eastern countries and groups; topic 11 may be hard to discern but words like "oil", "american", "control" and "security" certainly hint at American involvement in oil operations; topic 12 is intriguing insofar as it presents the elements of web addresses ("https", "org"), as well as mentions to wikis ("wiki", "wikipedia"); topic 13 also contains several references to web addresses ("com", "www", "https", "http", "html"); and finally topic 14 seems centered around certain Middle Eastern countries and their inhabitants. Clearly, the word lists allow us to begin to form certain initial interpretation (moreso if we continue past the top ten words), but these can be amplified or deepened by reading through the most characteristic posts within each topic. For brevity, we will not examine the top five in each topic, but rather, a small sample of representative posts to capture the general flavour of the topic.

Taken together with the most common words, the following posts lead to the interpretation that topic 1 is primarily composed of personal opinion posts: individuals either stating their own beliefs/opinions, or challenging those of others.

1. Why do you think that had to do something to warrant it. Are there not people raped, murdered, and robbed without provocation? The Hawaiians and indigenous people had their lands taken because it was possible to do so and no one was stopping them, that's all it takes - have something people want with no way to defend yourself and no support and there you go. (Doc. 29)

2. This is something that I think about a lot. They haven't been eradicated, obviously they cannot be, and I believe that they will be looking to attach themselves to a movement that has a "future" at this stage.

   I wonder how many have joined the ranks of the FSA or SAA or are shopping around for the next up-and-comer in the region. (Doc. 179)

3. I want to go, but I want to see the actual, physical country, and not the spectacle of how the DPRK runs. I'm happy waiting until that problem is solved, because I don't want to support the regime. (Doc. 2258)

Rank	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	years	need	iran	right
2	think	president	power	iraq	said
3	don['t]	citizenship	mean	country	let
4	want	conflict	actually	war	fuck
5	know	family	just	reddit	fucking
6	wouldn['t]	ago	doesn['t]	iranian	won['t]
7	lot	history	does	syria	bomb
8	region	basically	true	russia	happen
9	aren['t]	citizen	look	hope	literally
10	believe	point	don['t]	used	invasion
Rank	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	war	really	good	countries	israel
2	military	involved	like	middle	lebanon
3	government	america	did	east	hezbollah
4	world	didn['t]	going	shit	iran
5	isn['t]	know	lebanese	military	saudi
6	israeli	got	thing	money	syria
7	arab	source	just	way	saudis
8	want	probably	conspiracy	prince	war
9	trump	yes	make	country	sunni
10	civil	makes	political	korea	lebanese
Rank	Topic 11	Topic 12	Topic 13	Topic 14	
1	oil	https	like	saudi	
2	deleted	org	com	arabia	
3	yeah	trying	www	lebanon	
4	things	wiki	https	time	
5	american	wikipedia	http	saudis	
6	bad	lol	news	pretty	
7	control	exactly	say	yemen	
8	comment	wars	just	citizens	
9	make	know	html	maybe	
10	security	politics	current	country	

Table 2

The ten most significant words for topics 1-14 when the algorithm is run with number of topics set to 14.

Topic 2 fairly clearly groups a number of posts discussing the issue of American citizenship (as well as the citizenship requirements to be president of the U.S.A.), as illustrated by the following representative posts:

1. All three of them primarily resided in the US, had dual citizenship, were born to citizen parents *and* ended up relinquishing their foreign citizenship explicitly because of complaints about them.

   All three also lost primaries and never made it "anywhere near the general election for a major party." (Doc. 36)

2. Yes, as long as you are a natural born citizen of the U.S., 35 years or older, and have lived in the U.S. for at least 14 years, you can have a dual citizenship and still be President.

   Realistically, I doubt Americans would elect a dual citizenship holder unless they renounced their foreign citizenship. (Doc. 258)

3. My sister is a natural born American and has both Russian and American citizen ship. Theoretically she can be president (but imagine having to defend yourself on charges of collusion) (Doc. 1745)

Topic 3 is not as neatly defined as the two previous ones. While there is definitely a common thread of discussions of "power" through some of the most representative posts (examples 1-3 below), there is also a fair bit of noise among them, posts that don't necessarily fit thematically at first glance (examples 4-5 below):

1. The only power anyone has is that recognized by those around them. If the US is the only power that matters, then it alone is the player that determines what happens in the world, and what it wants, is

decided for the world. If it's "Western Culture" or the UN, then so be it. It's not a unanimous decision made by every person on earth. Either the powers that exist allow something or challenge something.

When I say 'the world has decided' I don't mean that every person likes that decision. But every person with the power to matter in the decision has been summed up. Power is disequal among nations, as it has been since time immemorial. In the time of Alexander it was nigh his will alone that made decisions. In the time of Genghis Khan it was that of his influence. The Romans ruled how they wished. Britain controlled most of the world with colonialism for it's time.

The way the world decides comes down to who holds the power to decide. I don't believe the US alone makes that decision, but it and it's allies certainly seem to hold all the power they need to make near unilateral decisions, within certain bounds, on behalf of the world. That's the realistic decision matrix that matters.

If you don't like that, lead the opposition, and see who the world decides to be the best option between the two.

Yes, this is bleak, but it is realistic. Social change is difficult if those with the power don't care. Something to keep in mind. (Doc. 159)

2. My theory is that there is an erratic baboon in the most important post in the world. So now all the other erratic baboons in power around the world sees this time of instability as their time to act. (Doc. 1820)

3. Power... unlimited power! (Doc. 491)

4. This is stupid, even if your intentions are best. Look what happened to poor Otto. (Doc. 63)

5. No need to apologize! I was just curious. (Doc. 95)

Topic 4 deals primarily with international relations and conflicts (focused on the Middle East), yet also makes several mentions of the reddit site itself. While this topic is majoritarily about middle eastern relationships and conflict, there is also some noise, as evidenced in examples 5 and 6 below. Example 6 is also interesting however insofar as it mentions a specific country (Australia), the word "country" and discusses weapons, even though it appears to be written in jest. In terms of language, it therefore "fits" the theme, even though it might not semantically be in line with the other examples in 1 through 4.

1. Been here for five wars and reddit talked about all of them except S sudan which was ignored (other four Syria, Liberia, Yemen, Nigeria/Boko Haram–wait Iraq Civil War vs iSIS, a little corner of the Philippines had an insurrection and held territory, many more minor conflicts) (Doc. 275)

2. You're thinking of the Levant and present day Iraq. The Arabian peninsula contributed little, if anything, to the Islamic Golden Age. (Doc. 1957)

3. The Iraqis (who outnumbered the ISIS invaders 10:1) fled Mosul and left the weapons and vehicles the US gave them behind. (Doc. 766)

4. The damn mess is entirely Obama's fault for cutting and running from Iraq, and for appeasing Iran. (Doc. 870)

5. Spa day, are you saying spaghetti day? (Doc. 57)

6. Given this was started by an Australian. The weapons will be frozen kangaroo tails at 3 paces. And I'm serious when I say that kangaroo tails are used as weapons in certain parts of our country. Suckers can do some real damage. (Doc. 2288)

Topic 5 isn't quite thematically-defined, but seems to largely group posts that contain the words "fuck" and "right". Many of these posts are very short, and likely difficult to classify; they therefore end up here together, as the examples below illustrate:

1. Fuck, it was right there, and I forgot him in the list. (Doc. 125)

2. Fuck me, for a second I thought that said London. (Doc. 1455)

3. And literally anyone can be appointed to Supreme Court if I remember the Constitution right (Doc. 213)

4. Saudis try to fuck up our winter tourism with threats, israelis try to fuck up our summer tourism (Doc. 367)

War is the prevailing theme for topic 6. While war and military are the primary topic of conversation throughout these posts, there is a strong undercurrent in the discussion about who stands to benefit and/or profit from military action:

1. Well the Saudis have the third highest military budget in the world, and spend billions every year purchasing US military equipment. (Doc. 970)

2. This argument would have been nice to throw around prior to the 2016 election when everyone was deriding Clinton for accepting donations from Saudis, when it turns out Trump is no different in that regard. His supporters were fooled during the campaign on that issue. (Doc. 897)

3. Just as the Syrian civil war seems to be winding up, another potential war is being planned. (Doc. 1287)

4. They will definitely want to sell arms and profit from the war! (Doc. 954)

Topic 7 groups together posts that respond to other statements regarding their accuracy or inaccuracy. Posts range from those upholding previous statements (example 1 below) to those expressing understanding after reading a post (example 2), while others seek further clarification (example 3) or challenge previous posts (example 4):

1. Anyone who does not acknowledge this actual, provable fact is doing a disservice to themselves. Hold all your politicians' feet to the fire. (Doc. 888)

2. Oh, ok that makes sense. (Doc. 1261)

3. Saudis are more involved in syria? Damn I didnt know this, whats your source? (Doc. 1363)

4. Find me ONE reputable source that says we didn't arm rebels who later joined ISIS. ...just one. I'll wait (Doc. 773)

Topic 8 is another topic that is somewhat difficult to nail down. It has a few related strands, notably: discussions about which actors know what in a given situation (examples 1 and 2 below), as well as discussion about conspiracy theories (examples 3 and 4):

1. well, NK, obviously but there's all kinds of places on the planet that might still be dangerous but not as well known. What if I want to go visit some natural sinkhole somewhere but it happens to be in an area that's not widely known to be dangerous, but is. Like the Boko Haram area was before they came about. If the State Department had up to date information about some local dangerous activity, I'd sure love a call from my airline saying "hey, we're just been notified that you might be flying into a dangerous situation and we'd like to make you aware of it." (Doc. 2282)

2. Except that Kushner doesn't know a thing about US-Saudi relations. Hopefully there are State and CIA folks actually running the meetings. (Doc. 917)

3. That it doesn't validate all other conspiracy theories is just another conspiracy theory. (Doc. 1621)

4. Love a good conspiracy theory. (Doc. 1758)

Topic 9 brings fairly well-defined themes together. Several posts describe American involvement in Middle Eastern affairs (examples 1-3 below), the Saudi Crown Prince (examples 3-5), as well as general discussion about Middle Eastern countries (examples 6-7).

1. The US is such a major player in the Middle East that non-involvement is involvement. (Doc. 799)

2. US interfering in the Middle East : Teasing the cat with a laser pointer (Doc. 972)

3. Jared met with the crown prince and king days before the purges started, no way the trump administration didn't prescribe this (Doc. 806)

4. What you perceived as left might not be the same what the crown prince perceived as left. (Doc. 1965)

5. You can use MBS to refer to the prince. Much easier than remembering his name (Doc. 242)

6. It's one of the most westernised countries in the Middle East. The people are friendly and t has a vibrant culture. The other is the most easternised country in the Middle East and the people have a certain arrogance. (Doc. 1108)

7. Im not so sure about that.

   If you knew what you were doing you'd stay way the hell away from the Middle East.

   Even if your a Muslim it's dangerous. (Doc. 126)

Topic 10 is one of the most clearly defined in this collection. Much as its word list suggests, it groups conversation about Hezbollah, Lebanon, Israel, and other Middle Eastern countries, specifically discussing military action between these actors, as exemplified in the posts below:

1. No he was not an Iran proxy. He is Sunni and has Saudi citizenship FFS. He's a Saudi puppet and they ordered him to resign. Saudis

and Israel want to start a war in Lebanon, having failed to oust Assad in Syria. They think Hariri resigning will throw Lebanon into chaos, giving Israel its "legitimate security concerns!!!!omg!!!" and its pretext for another one of its defensive preemptive invasions of Lebanon.

Assad is allies with Iran. Hezbollah in Lebanon has support from Iran as well. So far, Lebanon hasn't taken the bait. (Doc. 1829)

2. Israel fared really badly against Hezbollah in the 2006 war. It was a major failure. The rockets "stopped" but Hezbollah has stockpiled even more missiles, with heavier payloads and has even better training. The next go around will say Tel Aviv heavily hit. I also have heard Hezbollah has upgraded its anti-ship capabilities. (Doc. 1807)

3. Israel has repeatedly bombed Hezbollah targets in Syria, showing that their air defenses are useless. The Saudi air force is made up of modern US equipment, so while it's no match for Israel, it's definitely a match for Syria's/Hezbollah's outdated AA. (Doc. 1171)

4. I'll bet on Saudi airstrikes to soften Hezbollah. The Lebanese National Army takes control, Iran gets kicked out. (Doc. 1044)

Topic 11 is also well-defined. It contains substantial discussion about oil (and its role as a motivator in geopolitical conflicts). The topic also received all of the comments that had been deleted at the time that the texts were collected from Reddit (in future studies using Reddit as a data source, it may be interesting to cull these posts from the corpus for clarity).

1. Oil will always be valuable to control. Energy is one of millions of applications for oil. (Doc. 940)

2. Duhurr!!! We are in Afghanistan for oil!! (Doc. 1225)

3. They're protecting the petrodollar, not the physical oil. Jesus. (Doc. 1677)

4. Invest in defense and oil stocks (Doc. 418)

Topic 12 also appears to be well-defined, although it is well-defined in terms of post content, as opposed to theme. This topic groups posts that link to wikipedia pages, which matches up well with its word list:

1. https://en.m.wikipedia.org/wiki/Uncontacted_peoples (Doc. 2296)

2. https://en.wikipedia.org/wiki/Bhumibol_Adulyadej (Doc. 269)

3. I thought a haboob was one of those giant dust storms? https://en.wikipedia.org/wiki/Haboob (Doc. 290)

4. That's not much different from Europe's religious wars, for example. https://en.wikipedia.org/wiki/European_wars_of_religion (Doc. 222)

Topic 13 is practically an extension of the previous topic. It groups together posts that generally contain links to (non-Wikipedia) websites, also coordinating well with its word list.

1. https://www.nytimes.com/reuters/2017/11/06/world/middleeast/06reuters-lebanon-politics-saudi.html (Doc. 1043)

2. oh dear god... it's like you haven't been reading the news. https://www.washingtonpost.com/posteverything/wp/2014/08/18/the-terrorists-fighting-us-now-we-just-finished-training-them/

https://www.thedailybeast.com/main-us-backed-syrian-rebel-group-disbanding-joining-islamists

https://www.theguardian.com/commentisfree/2015/jun/03/us-isis-syria-iraq

https://www.google.com/search?q=us+supported+rebels+join+isis&tbas=0&source=lnt&
(Doc. 771)

3. [Hmmmmm. What's that?](http://www.aljazeera.com/indepth/features/2011/12/20111228
   (Doc. 1654)

4. Would that be survival of the fittest?

   http://www.dailygalaxy.com/my_weblog/2017/11/destroyer-of-cancer-cells-iridium-from-the-chicxulub-asteroid-66-million-years-ago-that-ended-the-di-1.html

   The potatoes info you'll have to find for yourself. (Doc. 50)

The final topic, topic 14, is largely centered around discussions of Saudi Arabia and its relationship with the United States, Lebanon, and other countries. Like the preceding topics, this is also largely in line with its word list.

1. What else is Saudi Arabia going to do with those weapons we sold them? (Doc. 278)

2. Maybe it's the Saudi version of Simon says. Saudi says. Saudi says leave Lebanon. They all leave Lebanon. Saudi says lets your wives learn to drive. The wives learn to drive. Film your wives driving. One guy films his wife driving. Everyone goes ape! (Doc. 1150)

3. Hi father made his fortune in Saudi Arabia and then became Prime Minister of Lebanon. And then he was assassinated with a car bomb (most likely Syrian intelligence) - and his son Saad assumed the political mantle. But he is not his father. (Doc. 2025)

4. The US has bases in and around Saudi Arabia. An attack on Saudi would be presented as an attack on those US forces. (Doc. 1235)

Discussion of Found Topics

The information we can gleam from the corpus is interesting in several respects. First, while intuitively we might expect to find topics of discussion within the corpus, we have also learned something about discursive patterns within this Reddit conversation. Also, while the algorithm's topics sometimes seem a little noisy, the posts that are grouped together to share qualities that might be worth exploring. Third, the algorithm produces results that link together themes that occur across the corpus, independently of where and when they occur.

The first significant finding when examining the algorithm's results with k set to 2 was not particularly surprising, but somewhat validating: the clear delineation of posts into two topics, where one topic clearly discusses conflict in the middle east while the other does not is a major and effective separation, which allows a human interpreter to glean the topic of the most important discussion in the corpus.

With k set to 14, topics were somewhat noisier, but nonetheless revealed interesting discursive patterns and themes. In order to discuss the relevance of the found topics, we should first examine the thread itself, which was initially posted about a news item describing how Saudi Arabia had just advised its citizens to leave Lebanon. A read-through of the corpus (in order), reveals the following loose topics of conversation within the thread. It should be noted that the topic was read in the order that

Reddit presents discussions by default: sorted by most popular first level comments, with each comment having its own tree of responses & discussions beneath.

1. Citizenship of the Lebanese prime minister & issues of citizenship for country leadership.
2. Dependence (or lack thereof) of the United States and its allies on middle eastern oil.
3. The roles of different countries in creating the current situation (questioning whether certain politicians were democratically elected or possibly installed by other governments), and which countries benefit the most from it. A part of this discussion explores whether the U.S.A. might be orchestrating the situation in order to maintain control of oil supplies (and whether or not this is a conspiracy theory).
4. Which middle eastern countries want to go to war with which other countries and what motivations exist for war.
5. One short subthread consists mainly of jokes about the saying that "even a broken clock can be right twice a day".
6. Which entities are supplying which Middle Eastern entities with weapons.

There is certainly some correspondence between the ‘topics’ found by the algorithm and those found by a simple read-through. Some of the found topics are grouped moreso by structure and specific word choice than by theme, such as topics 5 (short posts containing the words ’fuck’ and ’right’), 12 (Wikipedia links) and 13 (non-Wikipedia links). Other found topics seem to reflect some of the meta-discussion in the thread, such as topics 1 (statements of belief/opinion and challenges thereof)

Algorithm Topic	**Conversation Topic**
2 - citizenship & country leadership	1
9 - American involvement in Middle East, 11 - oil	2
4 - international relations, 8 - conspiracy theories, 9 - American involvement in Middle East, 11 - oil, 14 - Saudi Arabia	3
4 - international relations, 6 - war & profits from war, 10 - Middle Eastern entities, 14 - Saudi Arabia	4
	5
4 - international relations, 6 - war & profits from war, 9 - American involvement in Middle East, 10 - Middle Eastern entities	6

Table 3

A mapping relating the algorithm's found topics to the topics gleaned from a read-through of the Reddit conversation.

and 7 (comments about other posts' accuracy or inaccuracy). The remaining ten categories map more properly into the topics of conversation found in the thread. Topic 2 (citizenship & country leadership) maps well into conversation topic 1. Topics 9 (American involvement in Middle East) and 11 (oil) map well into conversation topic 2. Topics 9 and 11, as well as 8 (conspiracy theories), 4 (international relations) and 14 (Saudi Arabia) generally map into conversation topic 3 above. Topics 4, 6 (war and profits from war), 10 (Middle Eastern entities) and 14 map very well into conversation topic 4. Conversation topic 5, likely due to its brevity, did not figure among the algorithm's found topics. Finally, conversation topic 6 is well-represented by topics 4, 6, 9 and 10. This mapping is summarized in table 3.

While the mapping above at first glance does not offer a tidy 1:1 assignment between the algorithm's found topics and topics gleaned in a single read-through, this is likely a good thing. The algorithm's overlapping thematic categories reveal that it has found posts from throughout the thread that share similarities.

For instance, there is clearly thematic overlap between the sub-conversations in conversation topics 4 and 6, as they both pertain to war and military concerns in the Middle East. The fact that the algorithm's topics can lump together posts from across sub-conversations is helpful - it can aid a human analyst to make thematic connections across distant parts of a large corpus. In this way, the algorithm presents an important advantage: it is not subject to the same contextual interpretive biases that a human reader might be, so any posts across the corpus which include similar enough language are grouped together. The flip-side of this is that the algorithm does not understand the context in which a post was made, and therefore sarcasm, satire and posts with ambiguous meanings (for instance), might not be recognized (although these might be difficult for a human reader to identify as well).

Weaknesses. While the algorithm does present many positive points, interpreting a corpus through its results is not purely positive. First, there can be a substantial amount of noise in the topics. Even the most coherent topics contain some posts that don't seem to fit. Perhaps this is unavoidable - it might be an intractable problem to split a large discussion into a predetermined number of classes in a completely tidy manner. It is possible that this could be helped by running the algorithm for a larger number of iterations, as this *might* help better construct the distributions from which the topics are defined, but for the computational resources available for this experiment, that was unfortunately infeasible to test.

The model of topic distributions also seems to struggle with posts that respond to one another without explicitly referring to the original topic of the discussion. These posts either get lumped into categories that share discursive patterns (as opposed to themes), or are assigned to other topics with a low probability. While learning about discursive patterns can be interesting in its own right, the fact that the algorithm's topics can either be thematic or discursive in nature beacause of the way specific words are used in the corpus needs to be understood by human analysts at the outset. These computationally-defined 'topics' are not topics in the colloquial sense - they are groupings of documents that contain similar vocabulary, and therefore are thought to share some latent, underlying similarities.

Future Work

This study was essentially a proof-of-concept to get a sense whether a topic modeling algorithm could help a human make sense of a large quantity of text. This was however somewhat limited because while the corpus was 'large', a larger corpus may have provided the algorithm greater ability to discriminate between significant topics in the text, leaving less room for 'noisy' posts to carry greater weight in the division. Given that this semi-automated method is meant to help human analysts make sense of ever-increasing corpus sizes, increasing the quantity of text we feed to the algorithm would also continue to test its utility.

An increased corpus size would also make it possible to run the topic modeling algorithm on a set of posts identified as being part of a specific subset of topic(s) in order to find the characteristics that might distinguish different subgroups of those topics from one another. For instance, with the above method with the number of topics set to 2, we could re-run the methodology on each of the two resultant sets of documents to see what sub-topics might emerge in each.

Another potentially interesting option would be to run the algorithm on successively larger version of the corpus over time. For instance, if a corpus has 10,000 documents on day 1, and we collect 10,000 more on day 30, and an additional 10,000 documents every 30 days afterwards, we can potentially run the topic modeling algorithm every month and observe the evolution of topics. Do new topics emerge, new modes of discourse that weren't previously present, or is new vocabularly used? Topic modeling may help to shed insight into these questions. While this method may not provide definitive results (the algorithm may settle on different distinguishing criteria in each iteration), it may still help provide clues about the evolution of a corpus over time.

Conclusion

In this paper, we examined a mixed human and computational approach to investigating corpora using topic models. The methodology was applied to a corpus that, while somewhat large, could still comprehensibly be read by a human analyst or interpreter. Comparison of the method's output to a human reader's revealed that the algorithm found thematic distinctions within the text that are in accordance with what a human reader might find. The method also revealed relevant discursive patterns within the text. Finally, future avenues for research were presented that will form the basis of an upcoming project.

References

Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2014, May). How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. *arXiv:1405.3282 [physics]*. Retrieved 2018-10-21, from http://arxiv.org/abs/1405.3282 (arXiv: 1405.3282)

Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, *68*(6), 1397–1410.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In *Text mining* (pp. 101–124). Chapman and Hall/CRC.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *Ecml pkdd workshop: Languages for data mining and machine learning* (pp. 108–122).

Gerber, M. S. (2014, May). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*, 115–125. Retrieved 2018-10-27, from http://www.sciencedirect.com/science/article/pii/S0167923614000268 doi: 10.1016/j.dss.2014.02.003

Paul, M. J., & Dredze, M. (2012). Experimenting with Drugs (and Topic Models): Multi-Dimensional Exploration of Recreational Drug Discussions. In *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text.*

Venkatesh, V., Urbaniak, K., Narayana, M., Scrivens, R., Harb, R., Cheikh-Ibrahim, R., … Rodier, S. (2019). *Words in Context.* Retrieved from http://www.wordsincontext.ca/databrowser/about (Accessed: 2019-08-27)

Wallace, B. C. (2012). Multiple Narrative Disentanglement: Unraveling Infinite Jest. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1–10). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved 2018-10-23, from http://dl.acm.org/citation.cfm?id=2382029.2382031