Department of Justice: What Are Your Priorities?

Rachael Lam
December 19, 2021

1. Intro

The Department of Justice (DOJ) delivers press releases that are publicly available and contain a wealth of information about the investigations and prosecutions carried out by the DOJ. Using a number of topic modelling techniques, such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF) and Biterm Topic Model (BTM), I attempt to discover if the press releases reflect the priorities and agendas of the active president and political party, or if there is a standard type of case or cases that are continually prosecuted.

In my analysis, I determine that while there is some overlap in case type, there is a clear distinction of language dependent on the President and corresponding Attorney General in power. Even in cases where there are similar topics, there are still dissimilarities in the overall topics, highlighting the nuances between topics based on presidencies. These findings are important in understanding what the true priorities of the President and Attorney General are and how they can directly affect people and their communities.

2. Background

Conversations surrounding crime have always been racialized, genderized and politicized. Historically, crime has unfairly affected and victimized marginalized communities. These roots in America date back to colonization and the mass plundering of African countries of its people and resources. From the transatlantic slave trade to the reconstruction era, Jim Crow, Civil Rights Movement, and the Black Lives Matter protests today, crime rhetoric has mutated in multiple ways to consistently persecute communities of colour. The black codes of the Reconstruction Era established criminality as unemployment and the War on Drugs in the Reagan Era criminalized economic insecurity[1], both demonstrating the redefinition of crime based on political motivations.

More recently yet still historically, immigrant communities have been blamed for decades for increasing crime rates in America. This position has fuelled anti-immigration agenda despite no concrete evidence of rising crime.[2] In fact, crime rates appear to be dropping, especially violent crime, rather than rising.[3] Additionally, it is difficult to interpret crime rates and criminologists prefer to wait several years before declaring a trend in crime.[4] Despite this documentation, the US-Mexican boarder has been heavily policed and human rights violations have occurred as children are separated from their parents.[5]

There are many hands that play a role in dictating the response to crime, one of which is the Department of Justice (DOJ). The DOJ is led by the Attorney General, who is nominated by the

---

[1] Michelle Alexander and Cornel West, *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*, (New York: New York Press, 2012), 48.

[2] Richard Stansfield, "Safer Cities: A Macro-Level Analysis of Recent Immigration, Hispanic-Owned Businesses, and Crime Rates in the United States," *Journal of Urban Affairs* 36, no. 1 (September 2013): 505, https://doi.org/10.1111/juaf.12051.

[3] German Lopez, "After 2 years of increases, the US murder rate officially fell in 2017," *Vox*, September 24 2018, https://www.vox.com/2018/9/24/17895572/murder-violent-crime-rate-fbi-2017.

[4] Ibid.

[5] Julie Hirschfeld Davis and Michael D. Shear, "How Trump Came to Enforce a Practice of Separating Migrant Families," *The New York Times*, June 18, 2018, https://www.nytimes.com/2018/06/16/us/politics/family-separation-trump.html

active president and appointed with the approval of the United States Senate.[6] The DOJ "investigates and prosecutes cases under federal antitrust, civil-rights, criminal, tax, and environmental laws."[7] It also controls several organizations including the Federal Bureau of Investigation (FBI) and the Drug Enforcement Administration (DEA).[8] Over the years, the DOJ has changed tactics and priorities based on the preferences of the president. This direction can change significantly due to the party in office, either the Democrats or Republicans.

During the Obama administration, Barak Obama appointed Eric Holder as the Attorney General. Although criticisms can still be made about the actions during his presidency, Obama and Holder made significant contributions to criminal justice reform. During his incumbency, Holder made efforts to decrease the federal imprisonment rate by 9.5%, as well as deprioritizing prosecuting and reducing sentences of nonviolent drug cases.[9] Additionally, Holder investigated police misconduct and enforced decrees that motivated police reform and better policing practices.[10] The DOJ also prioritized cases against financial institutions related to the financial crisis.[11]

Conversely, during his presidency, Donald Trump actively revoked and deprioritized many of the successes of Holder and Obama. Reverting to a "tough on crime" policy, Trump and Attorney General Jeff Sessions pivoted to harsh sentencing and mandatory minimums for marijuana prosecutions, disproportionately affecting marginalized communities.[12] Sessions also ordered "all U.S. Attorneys to prioritize immigration cases and threaten[ed] to strip funding from cities that [did] not cooperate with federal immigration authorities."[13] During this time, the DOJ also ended the tactic of suing police departments for violating the civil rights of people of colour and instead promoted more "effective policing."[14]

3. Research

Armed with this background of the DOJ's role in prosecuting various crime based on presidencies, I am interested in the DOJ's press releases and the language used. In particular, I want to investigate:

1. Do presidencies influence the language used in the DOJ's press releases?
2. Do the DOJ's press releases reflect the policies during each presidency?

4. Methods

In this analysis, I utilized DOJ press release data from Kaggle[15] that spanned from 2009 to 2018. I then personally scraped the DOJ's website for the press releases from 2018-2021. In

---

[6] "attorney general," Britannica, accessed December 16, 2021, https://www.britannica.com/topic/attorney-general

[7] "U.S. Department of Justice," Britannica, accessed December 16, 2021, https://www.britannica.com/topic/US-Department-of-Justice.

[8] Ibid.

[9] Ames Grawert and Natasha Camhi, "Criminal Justice in President Trump's First 100 Days," *Brennan Center*, April 20, 2017, https://www.brennancenter.org/our-work/research-reports/criminal-justice-president-trumps-first-100-days.

[10] Ibid.

[11] Barak Obama, "Statement by the President and Attorney General Eric Holder" (speech, State Dining Room, September 14, 2014), The White House, https://obamawhitehouse.archives.gov/the-press-office/2014/09/25/statement-president-and-attorney-general-eric-holder.

[12] Grawert and Camhi, "Criminal Justice in President Trump's First 100 Days."

[13] Ibid.

[14] Pete Williams, "AG Sessions Says DOJ to 'Pull Back' on Police Department Civil Rights Suits," *ABC News*, February 28, 2017, https://www.nbcnews.com/news/us-news/ag-sessions-says-trump-administration-pull-back-police-department-civil-n726826.

[15]

the end, the dataset used contains 18290 press releases before cleaning from January 2009 to October 2021. Additionally, the dataset contains the press release ID, title, contents, date, topics and components. The ID is the press release number created by the DOJ; topic is the subject of the press release, such as “Environment,” “Civil Rights,” or “Foreign Corruption;” and component is the division that investigates and prosecutes the case, such as “The Federal Bureau of Investigation (FBI)” or “Antitrust Division.” In both the topic and component features, multiple tags can be given as each case may be investigated by a number of organizations and be prosecuting a number of convictions.

There were several observations that had errors due to the scraping. Each of these observations were problematic because they were of a different format than the others, mainly speeches or award presentations. After removing these observations by only selecting observations that included a year between 2009 and 2021, there were 17733 rows remaining. This tactic was used because most incorrect observations did not have a corresponding date. Additionally, I removed the features “topics” and “components” as they were not necessary to this analysis that was without a prediction element.

For a successful natural processing language (NLP) model, it was necessary to clean the documents. To do so, I converted all text to lower case and removed special characters and

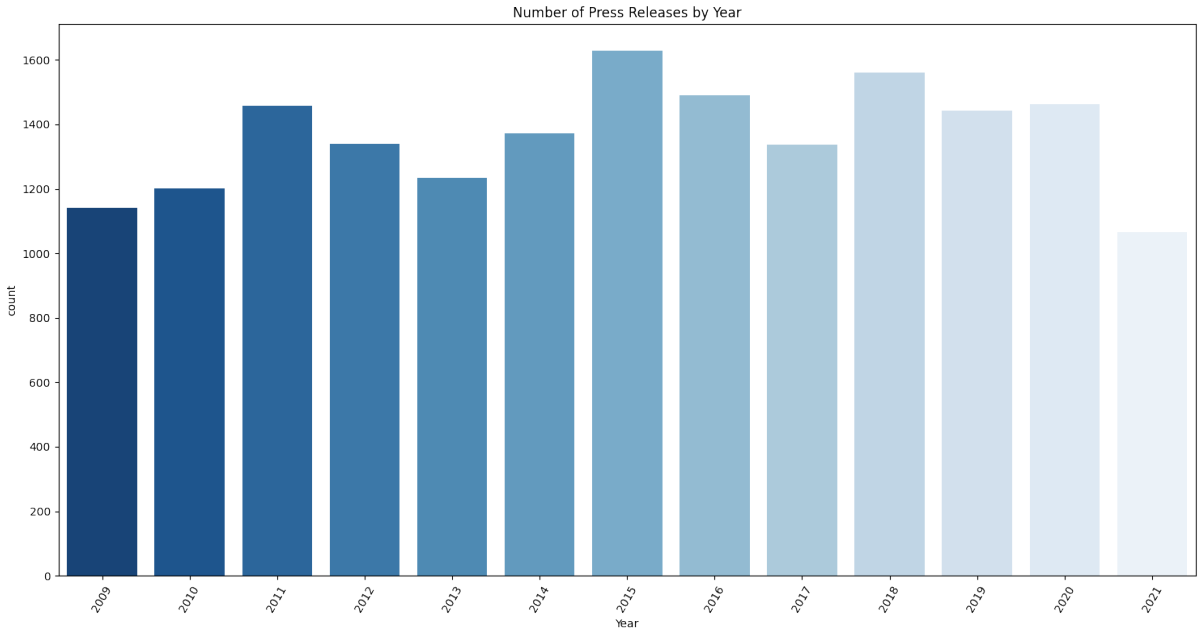

Figure 1: Number of Press Releases by Year

punctuation using the package `Re`. The documents then were lemmatized using `spacy` and tokenized using `gensim`. Another step of cleaning that was experimented with was performing a TF-IDF (term frequency-inverse document frequency) vectorization using `sklearn` to select words that only appear in more than 5% and documents and less than 90%. I then produced a list of words from the results of the vectorization and filtered out words that did not appear in the list. In doing so, I removed words that were meaningless to the analysis and words that were overused, such as attorney, department and law. These words would not have added any substance to the analysis and were easily filtered using TF-IDF vectorization.

5. Analysis Results

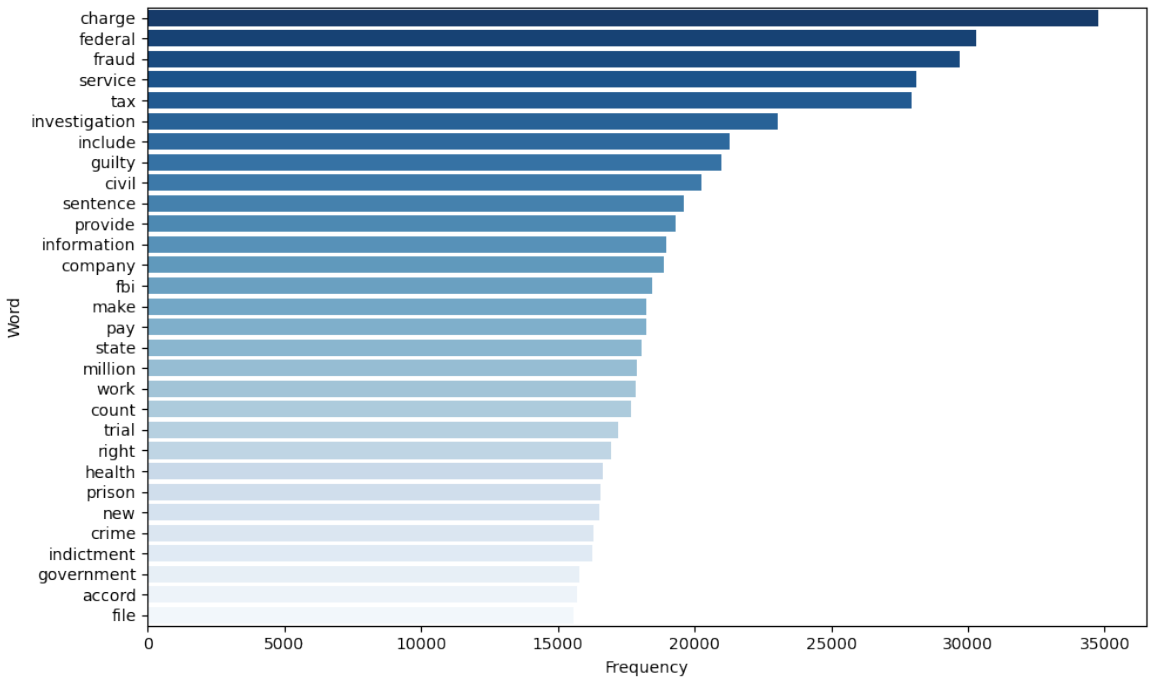

Figure 2: Word Frequency in Documents

In order to better understand the dataset at hand, I began a preliminary exploratory analysis by creating several visualizations of the data. First, I wanted to understand if there were any fluctuations of the number of press releases throughout the years (Figure 1). We can see that the fluctuations are not dramatic, with the largest gap being around 500 press releases. There are fewer press releases in 2021, but that is partly due to it being an incomplete year so the data collection is not whole. Between 2009 and 2014, there are on average, fewer press releases than from 2015 to 2020. This could be interesting as Obama's presidency was from 2008 to 2016 and Trump's was from 2017-2020.

Before navigating the topic modelling, I also wanted to better understand the distribution of words from a frequency perspective. To begin, I calculated the frequency of words in all the documents after lemmatizing and removing stop words (Figure 2). I then focused on the top 30 words in all documents and found a few words that stood out, potentially representing some of the most investigated or prosecuted cases. These words included "tax," "health" and "crime." Additionally, these frequencies uncovered that the FBI may be involved in many of these cases, which is understandable as the DOJ typically investigates federal cases. We can also see that "civil" appears often as the DOJ also commonly works closely with state prosecutors in civil cases that have federal elements. This word could also be tied to the word "right," seen further down the graph, to indicate cases of "civil rights."

I then compared this overall graph with two graphs comparing the word frequencies between the presidencies of Obama and Trump to uncover any potential differences in press releases during the two time periods (Figure 3). Again, I chose the top 30 words to explore. Although many of the words are similar across the two periods, such as "charge," "federal," "tax," and "fraud," there were a few words that either differed fairly substantially between the two graphs, or were absent for one of the two, shown in red. For instance, in the years of Obama's

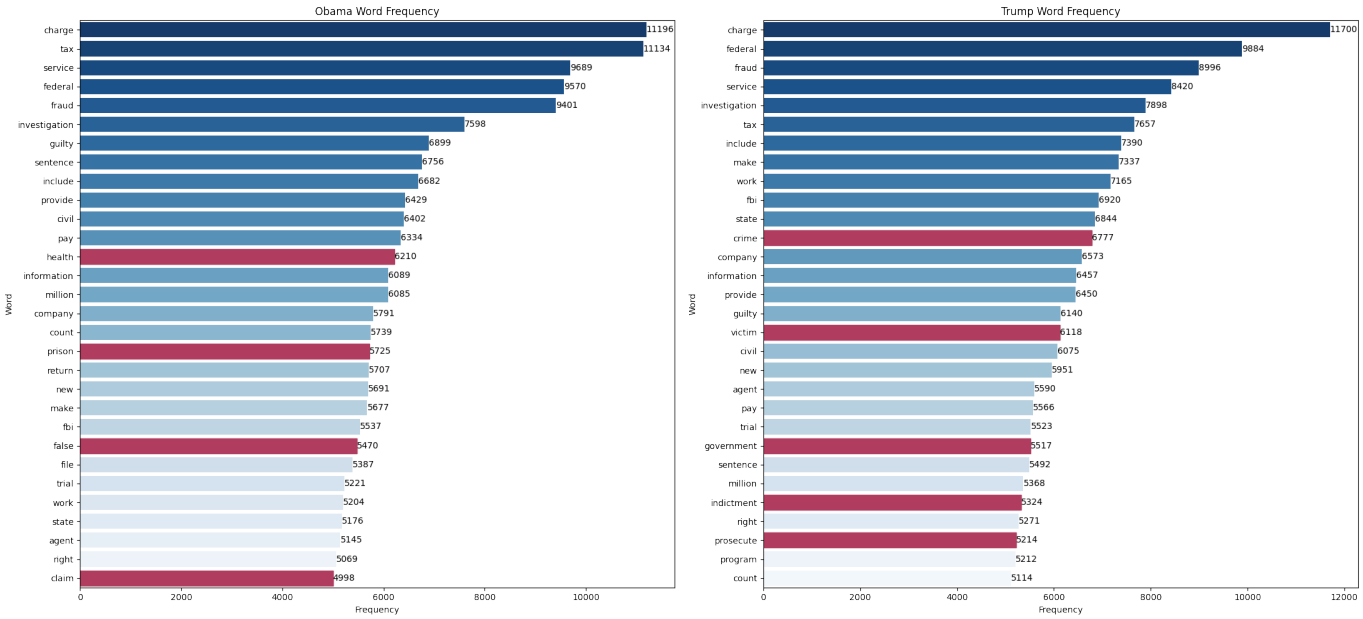

Figure 3: Word Frequency between Obama and Trump

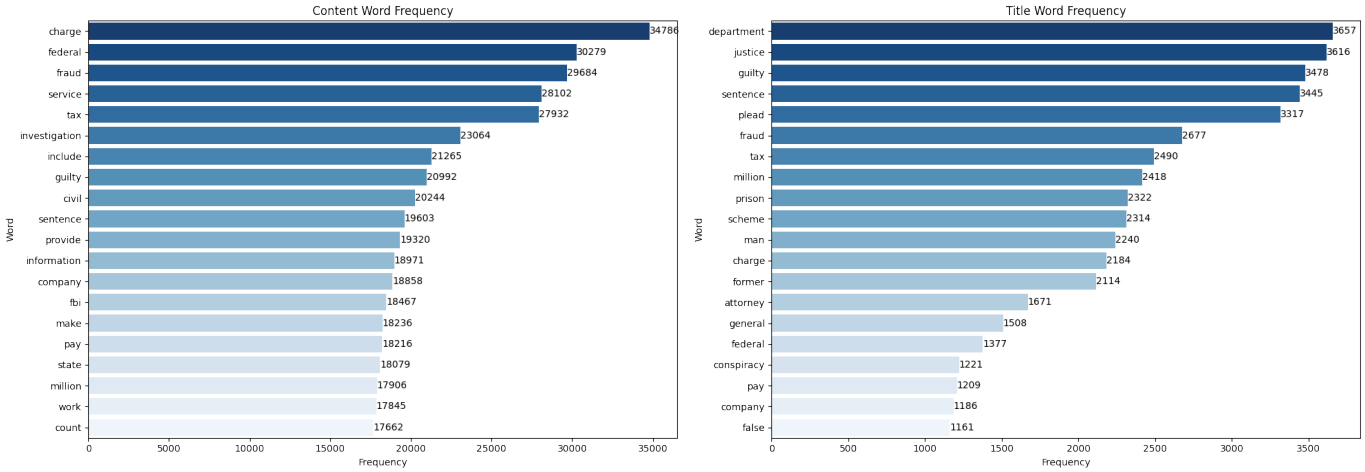

Figure 4: Word Frequency between Content and Titles

presidency, “health” shows up quite frequently, whereas it is not mentioned in the top 30 words during Trumps presidency. Similarly, “crime” is mentioned at a high frequency during Trump’s presidency, but is not in the top 30 words during Obama’s presidency. This is not to say that these words do not show up at all, as I am only examining the top 30, but it does highlight the difference in topics between the two presidencies. This could be an initial representation of the contrast between the agendas and priorities of the DOJ in each time period.

Finally, I compared the top 20 word frequencies between the titles and the content of the DOJ press releases to see the similarities and differences (Figure 4). There were are few differences between the frequencies but more importantly, it demonstrated that the content and the titles

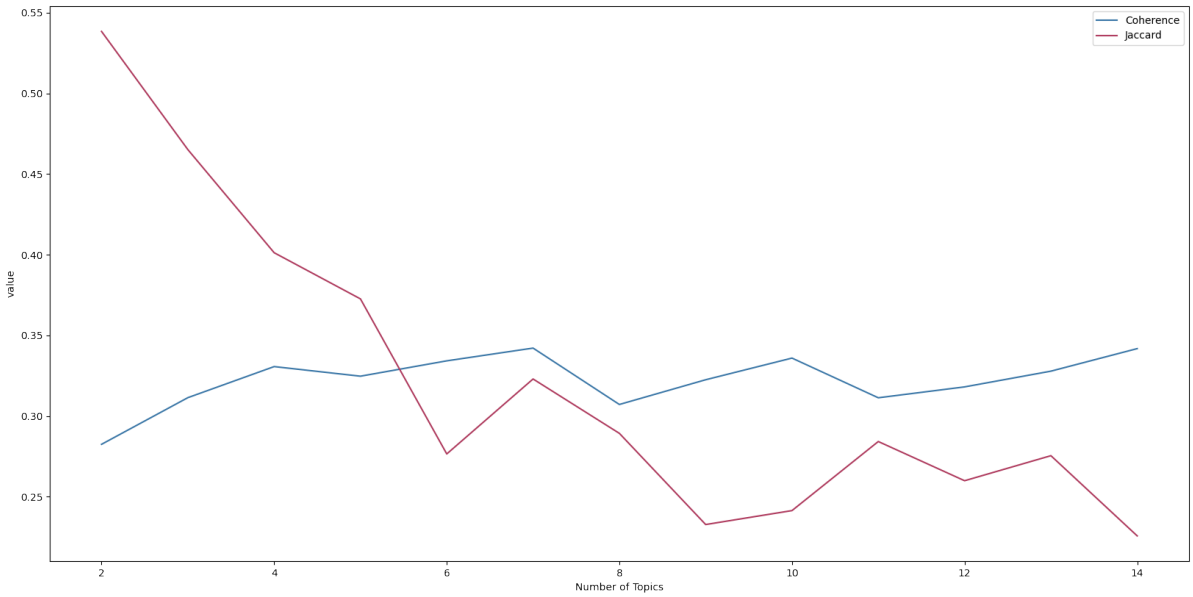

Figure 5: Gensim NMF Coherence vs. Jaccard

Topic #	Topics
1	0.020 * "tax" + 0.011 * "federal" + 0.010 * "charge" + 0.009 * "return" + 0.008 * "service" + 0.008 * "investigation" + 0.007 * "fbi" + 0.006 * "fraud" + 0.006 * "include" + 0.006 * “sentence"
2	0.010 * “fraud” + 0.009 * “charge” + 0.009 * “federal” + 0.007 * “fbi” + 0.007 * “investigation + 0.007 * “provide” + 0.006 * “new” + 0.006 * “drug” + 0.006 * “include” + 0.006 * “crime”
3	0.010 * "service" + 0.010 * "right" + 0.009 * "federal" + 0.009 * "charge" + 0.009 * "fraud" + 0.008 * "company" + 0.008 * "civil" + 0.007 * "guilty" + 0.007 * "program" + 0.007 * “health"
4	0.008 * "charge" + 0.008 * "service" + 0.008 * "federal" + 0.007 * "fraud" + 0.007 * "investigation" + 0.007 * "information" + 0.006 * "pay" + 0.006 * "provide" + 0.006 * "state" + 0.006 * “civil"
5	0.009 * "civil" + 0.009 * "company" + 0.009 * "service" + 0.008 * "federal" + 0.008 * "health" + 0.007 * "charge" + 0.007 * "fraud" + 0.007 * "medicare" + 0.006 * "million" + 0.006 * "provide"
6	0.014 * "charge" + 0.011 * "fraud" + 0.007 * "service" + 0.007 * "include" + 0.007 * "guilty" + 0.007 * "sentence" + 0.006 * "tax" + 0.006 * "work" + 0.006 * "health" + 0.006 * “investigation"

Table 1: Gensim LDA Topics

may be closely related, and we can understand the information in the content by reading the titles.

With this information from the exploratory analysis in mind, I began the topic modelling analysis by performing Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF) and Biterm Topic Model (BTM). In order to avoid clouding the topic model with infrequently used words that added no value, such as single use nouns or numbers, I used the list of tokenized words run through a count vectorizer that had been created earlier. The NMF model was applied in a loop to experiment from 2 to 15 topics. I then calculated the Jaccard and Coherence metrics on each topic model to see which had the optimal results. Figure 5 shows the best topic model appears to be 5 topics. I then extracted the topics from the model to further understand what words were in each topic (Table 1).

Topic 0 seems to be related to fraud, Topic 1 to drugs, Topic 2 to civil rights fraud or healthcare, Topic 3 to fraud again, Topic 4 with medicare and health and Topic 5 with fraud,

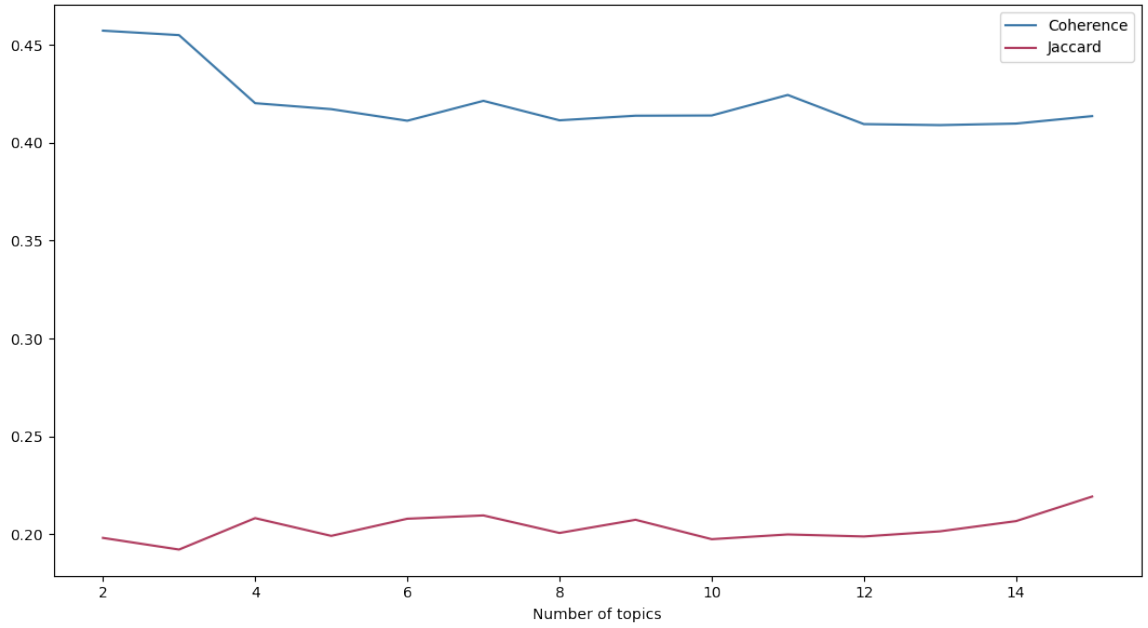

Figure 6: Sklearn NMF Coherence and Jaccard

tax and health. From here, we can see that the topics are all very similar, with some nuances in health or civil rights. While there is some valuable information, the topics seems to overlap in cases relating to fraud and the potential outcome or charges in those cases.

NMF using `sklearn` produced some more interesting results. Similarly to the LDA model, I iterated from 2 to 15 topics and calculated the Jaccard and Coherence for each topic model. I used a TF-IDF vectorizer to remove words that were found in less than 5% of documents. Unfortunately, I was unable to set a maximum document frequency as it is not an available feature in `gensim Word2Vec`. Despite only using a minimum document frequency, this model produced clearer, more informative topics based on the domain knowledge. Figure 6 shows the

Topic #	Topics
1	charge, fraud, criminal, indictment, conspiracy, company, attorney, count, district, guilty
2	tax, return, irs, income, refund, false, prepare, preparer, file, business
3	medicare, health, care, fraud, hhs, patient, oig, claim, medical, strike
4	settlement, department, civil, discrimination, right, housing, act, agreement, justice, disability
5	right, officer, attorney, police, law, victim, crime, civil, department, assault
6	child, sexual, project, safe, abuse, victim, attorney, internet, exploit, district

Table 2: Sklearn NMF Topics

Coherence and Jaccard metrics and although there is no clear point of overlap, Table 2 highlights the more interesting topics found. Based on these results, the best topic model appears to be with six topics or 14, but following the Principle of Parsimony, six topics was selected for further evaluation.

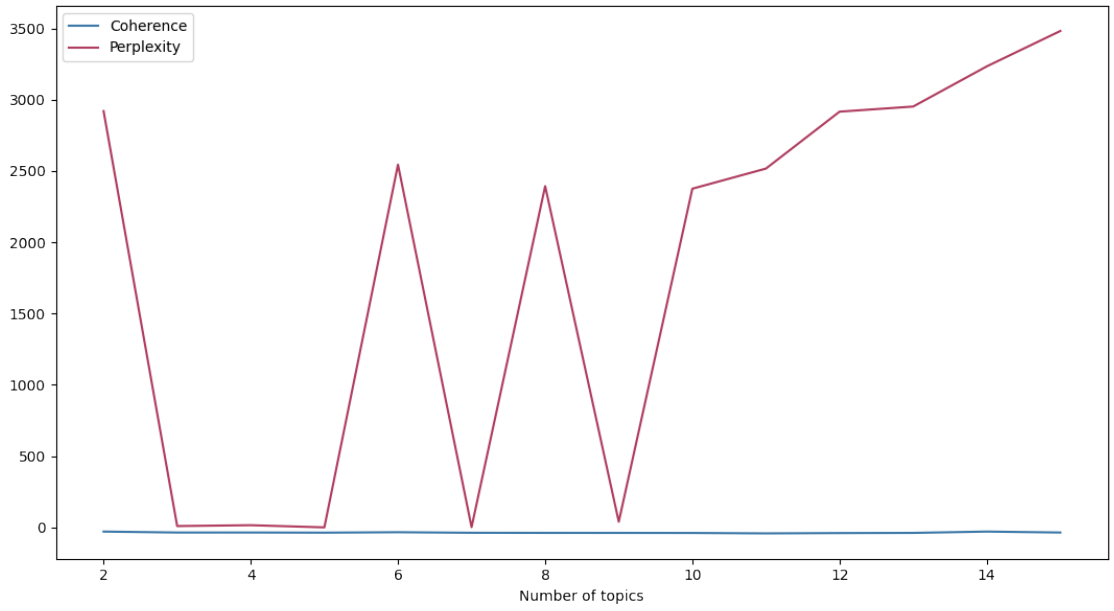

Figure 7: Bitermplus BTM Coherence and Perplexity

Based on the results displayed in Table 2, we can see that Topic 1 is about fraud, Topic 2 is about taxes and IRS investigations, Topic 3 is about medicare and medical services, Topic 4 is about civil rights in housing and disability, Topic 5 is about civil rights and potentially police assault or misconduct, and Topic 6 is about child sexual exploitation with regards the the

Topic #	Topics
1	charge, fraud, investigation, fbi, trail, special, agent, prosecute, investigate, section
2	work, crime, state, right, federal, community, victim, include, civil, continue
3	count, sentence, charge, prison, conspiracy, guilty, plead, indictment, accord, maximum
4	tax, information, return, civil, file, federal, complaint, right, discrimination, individual
5	company, pay, million, account, claim, bank, service, false, provide, business
6	fraud, health, medicare, service, care, force, federal, program, settlement, hhs

Table 3: Gensim BTM Topics

internet. These topics are extremely interesting and provide a wide range of topics that the DOJ may handle in federal cases, especially surrounding tax fraud, civil rights, and child abuse. Finally, I conducted a BTM analysis on the content of DOJ press releases. Although BTM is not the best model for this text due to the high volume of words, I wanted to see if it would produce any varying results. Following the same methods as the LDA and NMF analysis, I iterated from 2 to 15 topics, discovering the best topic model at 6 topics (Figure 7). Unfortunately, the topics produced were not as insightful as the NMF model (Table 3), although

	LDA	NMF		BTM
Coherence	0.5210	0.4112		-33.990706
Jaccard	0.2765	0.2080	Perplexity	2544

Table 4: LDA, NMF and BTM Metrics

they were still more informative than the LDA model. Topics seem to be less defined: civil rights and discrimination overlap with taxes in Topic 4, and in Topic 2, civil rights overlap with crime. Some topics are a bit clearer, such as Topic 6 which discusses fraud in healthcare and medicare.

Table 4 highlights the metrics found in the LDA, NMF and BTM models, and allowed me to determine the best model was the NMF model using sklearn. This model had a lower coherence score than the LDA model, and although it had a higher Jaccard score, based on my domain knowledge, the NMF provided compelling results. The BTM model produced some unclear metrics, which could be due to the fact that `bitermplus` uses a different coherence calculation, rather than 'c_v.'

After the initial analysis using the entire corpus, I wanted to then explore if there were any differences between the DOJ press releases between the presidencies of Obama and Trump. To do so, I split the dataset by presidency and ran NMF using `sklearn` on the separated datasets to understand if or how the topics had changed over the years. I used the available seven years of Obama's presidency and four years of Trump's presidency. Additionally, I only ran NMF as it was the best performing topic modelling technique, so it was unnecessary to run all techniques again. Figure 8 compares the Coherence and Jaccard metrics of both Obama

and Trump NMF models. Both models indicate that the best topic model is around six topics, although I also produced 11 topic models to better understand the variation in topics.

Topic #	Obama Topics	Trump Topics
1	fraud, us, charge, company, conspiracy, financial, attorney, guilty, bank, district	indictment, charge, us, attorney, district, fraud, fbi, count, guilty, conspiracy
2	tax, return, irs, income, prepare, refund, false, injunction, preparer, customer	tax, return, irs, income, zuckerman, deputy, attorney, principal, file, division
3	medicare, health, care, fraud, hhs, patient, oig, claim, service, medical	medicare, health, fraud, care, hhs, patient, strike, oig, kickback, claim
4	settlement, department, discrimination, civil, housing, right, disability, justice, agreement, employment	settlement, antitrust, civil, department, right, discrimination, complaint, division, agreement, company
5	right, officer, attorney, police, indictment, civil, victim, assault, member, charge	crime, law, drug, enforcement, officer, police, community, violent, attorney, department
6	child, sexual, safe, project, abuse, attorney, internet, us, victim, well	child, sexual, exploitation, project, safe, victim, virginia, attorney, abuse, district

Table 5: NMF Obama and Trump Topics

From these topics, we can see that there are some discrepancies between the two presidencies. While both presidencies tackle cases of fraud, tax and child abuse, there are differences in themes for their response to healthcare, civil rights and crime. Topic 3 indicates both Obama and Trump are addressing healthcare cases, Trump's theme seem to be directed around fraud, strikes and kickbacks, while Obama's theme seem to address the patient aspects of medical services. Additionally, Topic 4 in Obama's presidency focuses on civil rights in regards to disability, employment and housing, while Trump is centred around antitrust and settlements. Finally, Topic 5 in Obama's presidency seems to be addressing civil rights surrounding members of law enforcement, while Trump's presidency seems to be more anchored in drugs and crime.

6. Discussion

The DOJ has a decent amount of authority to dictate what is and is not a priority in federal offences. Overall, we can see that there are some obvious themes that the DOJ focuses on, including issues of tax, fraud, healthcare, crime and child abuse. These are federal offences that the DOJ would most likely have a large part in investigating and prosecuting. What is more interesting, is the nuances between themes during Obama's presidency and Trump's presidency. While they both focus on similar issues that the overall analysis pointed out - healthcare, tax, fraud, child abuse - there are some dissimilarities in the language and words used.

Unfortunately, I could not extract the top documents of each topic in `sklearn's` NMF package, but based on the previous lit review, we can see that some of these topic variances are justifiable. We know that Obama made a concerted effort with Holder to focus on issues of police misconduct and civil rights cases. We can see in the topic modelling that these themes come to light with words such as civil, rights, discrimination, housing and disability in Topic 4

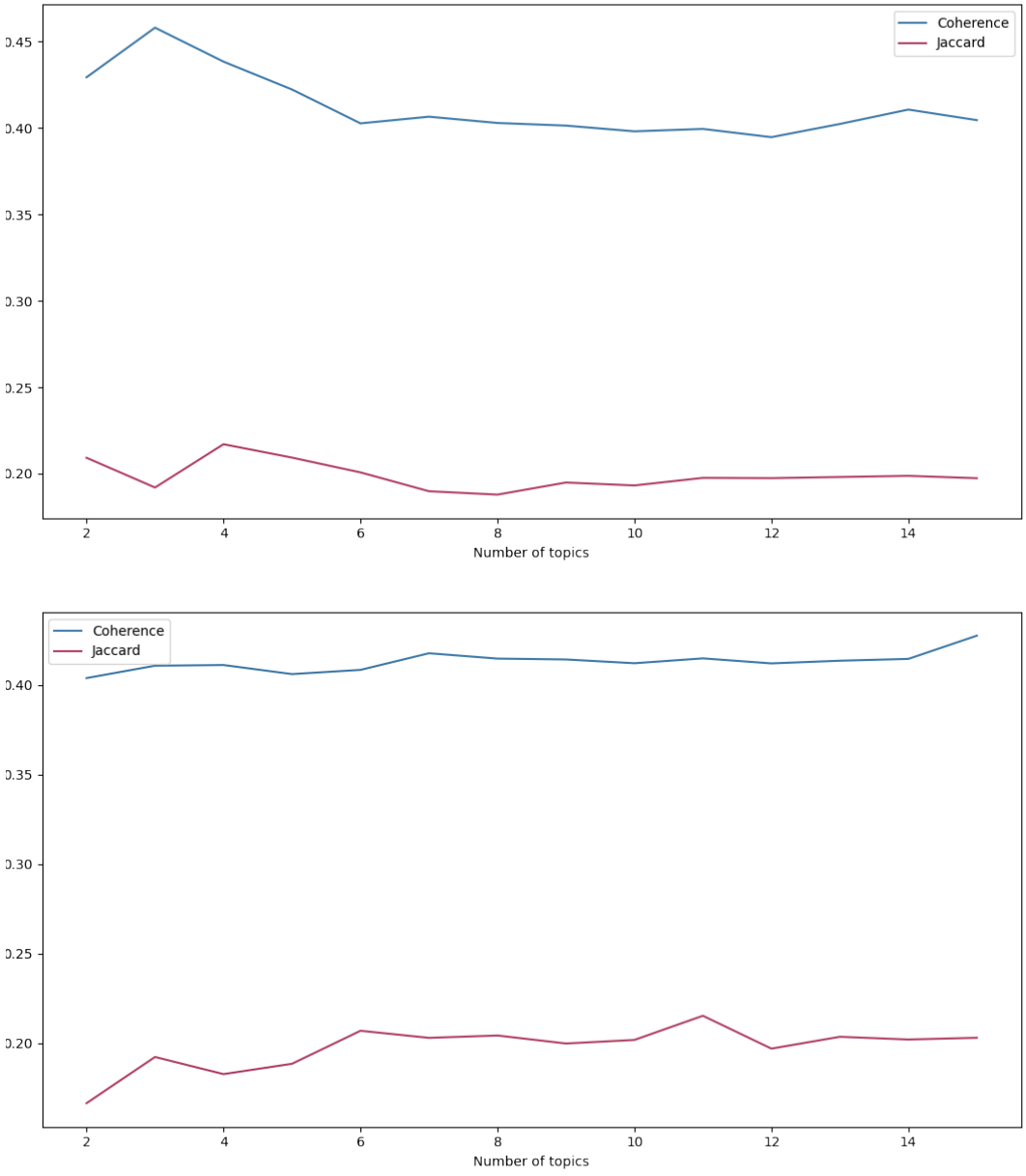

Figure 8: NMF Coherence and Jaccard for Obama (top) and Trump (bottom)

and officer, police, indictment and assault in Topic 5. On the other hand, we know that Trump and Sessions took a tough-on-crime stance, thus producing themes such as enforcement, drug and violent. In topic 3, the term “kickback” occurs, highlighting Tumps institution focused stance on healthcare, prioritizing companies wealth over patients health.

Although no American government entity has full power, this topic modelling analysis shows that the DOJ has enough authority to investigate and prosecute whatever cases fit the political

agenda of the time. If that is a tough-on-crime approach, we will see more DOJ press releases related to crime, thus more cases against people and organizations in the drug industry. Unfortunately, many of those individuals are constrained to a cycle of poverty, and rather than provide the necessary resources to achieve a comfortable standard of living, resources are being piled into the institutions that already marginalize and abuse these communities. Language surrounding these issues can also be extremely detrimental, including words such as “violent” and needing “enforcement” to control and subdue criminal activity, which is usually low level drug offences.

On the other hand, when people are placed at the centre of concern and actions are focused on justice and reform, we can see language change in the DOJ press release to civil rights, housing and discrimination. It is clear from this topic modelling analysis that DOJ press releases correspond with the policies and schemes of the active president. Whether it is Republican or Democratic, each presidency influences the language used, thus actions taken by the DOJ and sitting Attorney General.

7. Conclusion

Through this topic modelling analysis, I attempted to demonstrate how the political agendas of different presidencies can shape and influence the priorities of the DOJ. Because the Attorney General is appointed by the president, this relationship and its outcomes are naturally expected to be closely related. As the NMF modelling has shows, this instinct is correct. Although there are topics that are consistently investigated and prosecuted, irrespective of presidencies, there are still nuances in the themes and language of each case, and thus priorities of the DOJ.

The analysis supported the literature review by finding the press releases of the Obama presidency to be focused on civil rights and police misconduct, while the press releases of the Trump presidency to be focused on drug offences and healthcare corporations. These decisions have had real consequences, especially for marginalized communities in America. Heavy policing, police misconduct and extreme sentencing for minor drug offences have left a trail of exploitation and mistrust. There needs to be greater oversight and restriction to what the DOJ can and cannot do, as the repercussions of poor priorities and prosecution can be detrimental.

Bibliography

NBC News. “AG Sessions Says the Justice Department Will ‘Pull Back’ on Police Department Civil Rights Suits.” Accessed December 20, 2021. https://www.nbcnews.com/news/us-news/ag-sessions-says-trump-administration-pull-back-police-department-civil-n726826.

Alexander, Michelle, and Cornel West. The New Jim Crow: Mass Incarceration in the Age of Colorblindness. Revised edition. New York: New Press, 2012.

“Attorney General | Britannica.” Accessed December 19, 2021. https://www.britannica.com/topic/attorney-general.

“Criminal Justice in President Trump’s First 100 Days | Brennan Center for Justice.” Accessed December 19, 2021. https://www.brennancenter.org/our-work/research-reports/criminal-justice-president-trumps-first-100-days.

“Criminal Justice in President Trump’s First 100 Days | Brennan Center for Justice.” Accessed December 20, 2021. https://www.brennancenter.org/our-work/research-reports/criminal-justice-president-trumps-first-100-days.

Davis, Julie Hirschfeld, and Michael D. Shear. “How Trump Came to Enforce a Practice of Separating Migrant Families.” *The New York Times*, June 16, 2018, sec. U.S. https://www.nytimes.com/2018/06/16/us/politics/family-separation-trump.html.

Lopez, German. “After 2 Years of Increases, the US Murder Rate Officially Fell in 2017.” Vox, September 24, 2018. https://www.vox.com/2018/9/24/17895572/murder-violent-crime-rate-fbi-2017.

Stansfield, Richard. “Safer Cities: A Macro-Level Analysis of Recent Immigration, Hispanic-Owned Businesses, and Crime Rates in the United States.” *Journal of Urban Affairs* 36, no. 3 (August 1, 2014): 503–18. https://doi.org/10.1111/juaf.12051.

whitehouse.gov. “Statement by the President and Attorney General Eric Holder,” September 25, 2014. https://obamawhitehouse.archives.gov/the-press-office/2014/09/25/statement-president-and-attorney-general-eric-holder.

“U.S. Department of Justice | United States Government | Britannica.” Accessed December 20, 2021. https://www.britannica.com/topic/US-Department-of-Justice.