Maddie Pickens

Geographic Analysis of Immigration Discourse in the Media

Abstract

Newspaper media serves as an encapsulation of discourse around major world issues. When analyzing media discourse, two potential dimensions are immediately apparent: time and geography. For media discourse about immigration, the geographic component is particularly relevant; immigration is likely to be discussed differently depending on the source or destination region of the immigration in question. This could be because mentalities towards certain immigrant populations change over time, or because other immigrant populations are more "relevant" to current events in the news cycle. In this analysis, I use a corpus of immigration-themed news articles to analyze this geographic component of immigration discourse in the media. I use 11 logistic regression models to predict binary location variables corresponding to each news article. I then use cosine similarity to compare the top-weighted terms from the model results for each geographical category. I find that the top-weighted terms from models that overlap geographically are more likely to be similar, which is intuitive. I additionally find that media discourse is different around regions with a large immigration corridor, such as the US and Mexico, which is unexpected and may provide a venue for further research.

Corpus Overview

My corpus consists of news articles covering immigration topics from the period January 1, 2000 through February 4, 2020. To collect the corpus, I used ProQuest to search for English-language newspaper articles with keywords "immigration", "immigrant", "migrant", "migration", and "border." I did not limit the search to exclusively US-based news sources, because I was interested in global-level discourse, though the English-language restriction kept the sources to primarily the US and the UK.

ProQuest allows for search results to be downloaded in batches, though unfortunately the format is a bit difficult to parse. I ended up with 22 text files that each contained slightly less than 500 news articles, as well as accompanying information about the news articles (i.e. author). After splitting the text files into individual documents, I wrote a custom parser to extract the article text and appropriate metadata. Because the structure of the metadata varied by news article (or was sometimes not included at all), there were some missing values, though I was able to get the text for all but 13 articles. In addition to the text, I extracted covariates for the article author, ProQuest document ID, date published, publisher, and the location tag of the article.

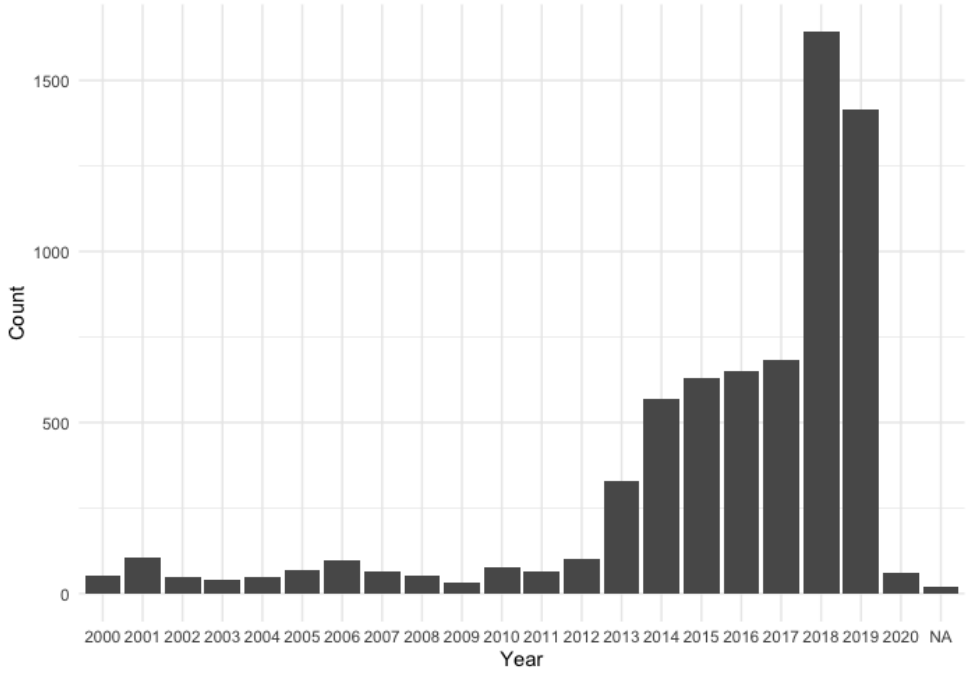
Covariate Creation and Corpus Exploration

Since this analysis was so geographically focused, I wanted to limit my corpus to only articles for which a location tag was available. This meant that my final corpus contains 6,843 documents. In this corpus, there are 2,449 unique authors (with 5,213 observations of the "Author" variable) and 596 unique publishers (with 6,843 observations of the "Publisher" variable).

The "Date" variable required additional cleaning in order to actually use it; since the parser read it as a string, I needed to transform it so that it could be read in datetime format. After cleaning and manipulation, Figure I (below) shows the distribution of documents over time, specifically the year that they were published. The corpus is much denser for later years, particularly beginning in 2014. This might mean that immigration began to occupy a more prominent place in media discourse around this time, which makes sense given events like Brexit, the policy decisions of the Trump administration, and the refugee crisis in Europe. However, given that the years are distributed so unevenly, any sort of time-based estimates would be prone to more instability in

earlier years, which is another reason I chose to focus on geography as the main dimension for analysis.

Figure I: Distribution of the Year Variable



The "Location" variable, while the heart of the analysis, was the hardest to transform into something interpretable. Each article in the corpus was pre-tagged with a "location" (which I used to create my initial Location variable) based on its content, but those locations varied widely in scale and length. For example, one such article tag is "Mexico Ireland United States--US Canada Quebec Canada Maine Kurdistan Europe." This tag includes locations on the continent level (Europe), the country level (Canada, Kurdistan, US, Mexico, Ireland), and the state or province level (Quebec, Maine). Additionally, the locations are really diverse, so it wouldn't be possible to rename this tag as one single location "category" – this article could be discussing immigration from Mexico to the US, or from the Kurdistan region to Europe, or global immigration patterns more broadly.

To avoid incorrectly categorizing a document, I decided to take an approach with dichotomous variables instead. Based on the locations represented in the "Location" tag, I created the following 11 categories based on regions and countries that I found most relevant to immigration and the broader corpus:

- United States (US)
- United Kingdom (UK)
- Mexico
- Canada
- Latin America and the Caribbean (LAC)
- Central America
- Middle East and North Africa (MENA)
- Europe
- Asia

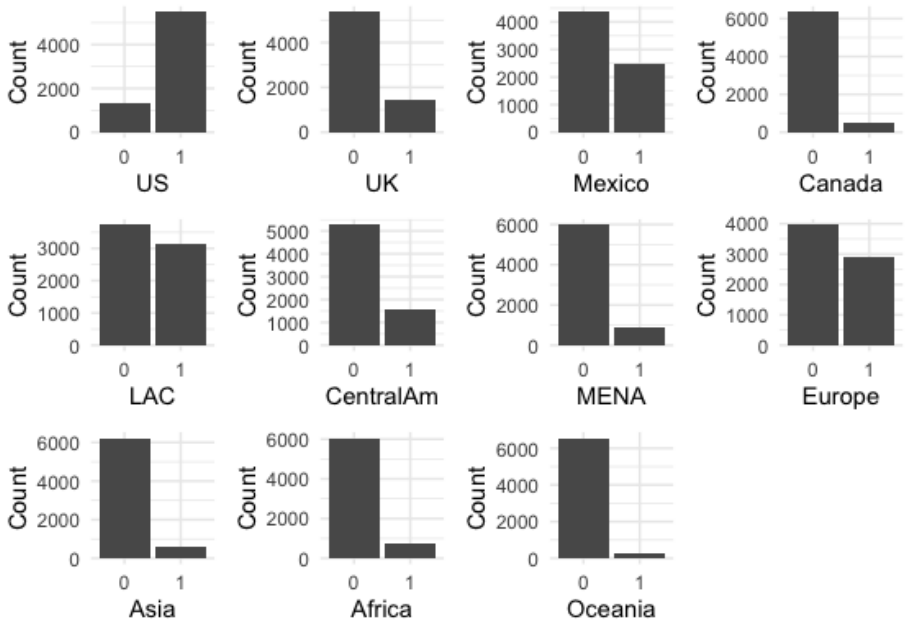
- Africa
- Oceania

Using the location tags of each document in the corpus, I also created a list of every keyword in the tags, ending up with about 650 location keywords. The example above would create 8 keywords: "Mexico", "Ireland," "United States", "Canada", "Quebec", "Maine", "Kurdistan", "Europe." Finally, I manually categorized these locations as belonging to each of these above categories. There was some overlap between the categories; for example, the "Mexico" keyword would belong to not only the "Mexico" category, but also the "Central America" and "Latin America and the Caribbean" category. I based the category designations mostly off of World Bank designations ("World Bank Country and Lending Groups") but also grouped these categories as they were likely to matter within the immigration domain; for example, Middle East and North Africa as one category rather than two.

Finally, I created 11 dummy variables for each of the location categories above and searched for the location keywords in the location tag of every document. If a keyword appeared in the location tag of a document, the corresponding dummy variable was assigned a value of 1; otherwise, the variable was assigned to be 0. In a document with the location tag given above, the "Mexico", "United States", "LAC", "Central America", "Europe", "UK", "Canada", and "MENA" variables would all have a value of 1, while other variables would have a value of 0. This way, I didn't accidentally double-count keywords, and I knew that all the regions represented in the article would be accounted for in at least one of the variables. My eventual goal was to use each of these variables as target variables in a logistic regression, to attempt to predict whether an article was tagged with a location keyword in a given category.

These dummy variables were distributed as shown below in Figure II:

Figure II: Distribution of Location Dichotomous Variables



The US is the only location that is tagged in the corpus more than it is not (5,493 documents are tagged with US locations); however, the LAC variable is also fairly balanced. Oceania is tagged in

the corpus the least, with only 284 documents representing the region, followed by Asia, represented in 610 documents. After the US, the most common category was LAC (3,117 documents) followed by Europe (2,884 documents) and Mexico (2,481 documents). Since most of the variables demonstrate some form of imbalance, this should be a consideration going forward, though I was not expecting the locations to be evenly distributed. The locations that are represented are consistent with the immigration domain, particularly in the media and particularly for an English-language corpus.

Preprocessing

In addition to the decisions about how to create location variables, there were a few preprocessing decisions that were important for making my logistic regression models as accurate as possible for predicting location categories. I decided to remove the punctuation in my corpus, since the punctuation doesn't contribute any additional information that could be used to predict location. I used similar logic for removing symbols. I also decided to remove numbers; in this case, I was actually concerned that if a document included a year, this year would contribute too much towards predicting a given location. For example, there was a lot of news coverage in 2017 and 2018 of the migrant caravans traveling from the Northern Triangle in Central America to the United States (Lind 2018). If a corpus document included "2018", it might actually be a predictor of the fact that an article was about US immigration. However, since I am interested in not only model accuracy but the words that are weighted most strongly in the model, inclusion of years as tokens could distort my results. I also removed stopwords, not only to reduce the dimensionality of the document term matrix that I was working with, but because stopwords were unlikely to contribute additional information towards predicting location since they are probably distributed fairly evenly across locations. I briefly considered using n-grams (either bigrams or trigrams), but decided that I could more directly compare the similarity of different geographical predictors if they were only one word. As such, the only tokens generated by the corpus are unigrams.

Since I did not want my model labels (location variables) to be derived from input variables (the document tokens), I also had to create a custom dictionary to remove location keywords. Since I already had a list of location keywords from generating the location dummy variables, I just built off of this list and included a few other common forms of the country or location keyword. For example, I added "Mexican" to the list in addition to "Mexico", "Iraqi" in addition to "Iraq", and "Californian" in addition to "California." Finally, I added the original search terms from the ProQuest search that I ran to gather the news articles – "immigrant", "immigration", "migrant", "migration", and "border." My hope was that after removing these terms, the remaining terms used to fit the model and predict location would accurately reflect differences across the documents.

Analysis – Supervised Learning and Cosine Similarity

After cleaning and preprocessing my data, as well as generating the necessary covariates, I was prepared to use supervised learning to predict the document location category or categories. Since my location variable is not a single categorical variable, but rather 11 dichotomous variables, I ran 11 separate supervised learning models to attempt to predict each location category.

I used a lasso logistic regression model with glmnet to make these predictions. I wanted to use logistic regression for two reasons; first of all, it works well on binary outcomes, and more importantly, it is easily interpretable. In addition to trying to predict whether a document is tagged with a location category, I was interested in *why* the document was tagged that way, so that I could examine differences and similarities in the media discourse about each location. Since the features of

the model are terms used in the documents, they contain information about which terms are strong predictors of certain locations and can be compared to predictors of other regions.

To test the model, I trimmed the corpus so that the minimum term frequency (the minimum number of times a term had to appear in the corpus to be included) was 50, and the minimum doc frequency (the minimum number of documents a term had to appear in to be in cluded) was 10. I split the data into a training and a test set, with 80 percent of the documents in the training set. I specified the prediction cutoff as 0.5, so that if the predicted value was greater than 0.5, the document was classified as a "1", otherwise it was classified as a 0. I chose 0.5 because I didn't feel I knew the domain well enough to choose another model and I wanted to keep the cutoff consistent.

To compare the top features for each of the 11 location categories, I selected the top 100 features (terms) that were weighted most positively and the top 100 features (terms) that were weighted most negatively for each location category and combined them into a "document." I then used cosine similarity to compute how different each of sets of terms actually were, giving me information about the differences in discourse across geographical categories.

While I considered using topic modeling rather than supervised learning to investigate geographical differences in discourse, I felt that investigating the features of a supervised learning model would ultimately be more insightful. Since I collected the corpus using immigration keywords, in a sense I already had a good idea of what topics would be included, which was confirmed when I ran a basic LDA model during the semester. These topics modeled current events rather than discourse and showing that the topics varied by location was therefore not very interesting. Finally, it was easier to use cosine similarity to compare lists of top-weighted features than to manually compare words in each topic to look for differences across locations.

Results – Model Accuracy

Using the model specification defined above, I obtained the following confusion matrices and accuracy scores for each model (in Table I, below):

Table I: Confusion Matrices and Accuracy Scores

US			**UK**			**Mexico**		
	Test Label			**Test Label**			**Test Label**	
Predicted Values	0	1	**Predicted Values**	0	1	**Predicted Values**	0	1
0	172	52	0	1052	54	0	806	84
1	98	1046	1	21	241	1	66	412
Accuracy	0.8904		**Accuracy**	0.9452		**Accuracy**	0.8904	
Balanced Accuracy	0.7948		**Balanced Accuracy**	0.8987		**Balanced Accuracy**	0.8775	
Canada			**Latin America and Caribbean**			**Central America**		
	Test Label			**Test Label**			**Test Label**	
Predicted Values	0	1	**Predicted Values**	0	1	**Predicted Values**	0	1
0	1268	31	0	669	78	0	995	84
1	3	66	1	76	545	1	58	232

Accuracy	0.9751	**Accuracy**	0.8874	**Accuracy**	0.8969
Balanced Accuracy	0.83903	**Balanced Accuracy**	0.8864	**Balanced Accuracy**	0.8400

Table I: Confusion Matrices and Accuracy Scores (cont'd)

Middle East/North Africa			**Europe**			**Asia**		
	Test Label			**Test Label**			**Test Label**	
Predicted Values	0	1	**Predicted Values**	0	1	**Predicted Values**	0	1
0	1156	81	0	760	64	0	1717	52
1	39	92	1	31	512	1	29	70
Accuracy	0.9123		**Accuracy**	0.9305		**Accuracy**	0.9452	
Balanced Accuracy	0.74958		**Balanced Accuracy**	0.9248		**Balanced Accuracy**	0.8987	
Africa			**Oceania**					
	Test Label		**Test Label**		**Test Label**			
Predicted Values	0	1	**Predicted Values**	0	1			
0	1181	65	0	1303	10			
1	30	91	1	8	46			
Accuracy	0.9305		**Accuracy**	0.9868				
Balanced Accuracy	0.77928		**Balanced Accuracy**	0.90766				

These resulting models are fairly good at predicting location. The accuracy scores are all above 0.85, higher than 0.90 for seven of the models, and higher than 0.95 for two of the models. Of course, balance in the variables is an issue, and the balanced accuracy scores are all lower. Of all the models, the balanced accuracy in the LAC model is closest to the actual accuracy, which makes sense because this variable was the most balanced to begin with.

The models all have a balanced accuracy above 0.75, with the LAC, Oceania, Europe, UK, and Mexico models all above 0.85. The model with the highest balanced accuracy was the Europe model, which also had a fairly high "general" accuracy (0.92 and 0.93, respectively). Interestingly, the US model performed the worst in terms of balanced accuracy. I suspect this is because documents that were tagged "US" were likely to overlap with many other regions (especially given that US was the most common tag), meaning that there were likely fewer unique features among those documents to help the model distinguish between US and non-US articles. This would also explain why location categories that may have had little overlap with other categories, such as Oceania and Asia, had good accuracy, though does not explain the strong Europe, UK, LAC, and Mexico models.

Results – Top Weighted Terms

In order to see what may have distinguished these models, it is helpful to view to top 10 most positively weighted terms and top 10 most negatively weighted terms for each model. Tables II and III, below, show the most positively weighted terms and most negatively weighted terms, respectively:

Table II: Top 10 Most Positively Weighted Terms

US	UK	Mexico	Canada	LAC	Central America	MENA	Europe	Asia	Africa	Oceania
comparing	rejecting	us-mexico	rcmp	us-mexico	Union-tribune	war-torn	hillary	teaches	reasoning	destructive
manhattan	lorry	rank	rust	pledging	preliminary	woes	backlogged	hardening	malcolm	tropical
relieved	observatory	militarized	relieved	vicente	prevents	rejects	labelled	classified	explore	greens
ralph	sajid	regulated	comparable	inspect	integral	benjamin	vary	historian	loophole	depart
abolish	woolfe	fever	varied	rank	decried	condemnation	relocate	bangladeshi	aspect	morrison
minimal	office	sinaloa	vows	sprawling	heartbreaking	guarantees	angela	observed	incorrect	occupation
slight	britain	futures	gwynne	construct	diplomacy	whoever	britain	depart	mat	teaching
separately	lands	construct	examining	continually	emphasize	applies	differ	continents	watchdog	oath
tendency	un	protocols	predictable	castro	escaping	turks	morocco	essex	battles	Cracking

Table III: Top 10 Most Negatively Weighted Terms

US	UK	Mexico	Canada	LAC	Central America	MENA	Europe	Asia	Africa	Oceania
better-paying	bureaucrats	foothold	google	syndication	stretching	brutally	bearing	jobless	washing	earns
Risked	downturn	skyrocketing	dual	fast-track	melania	reputation	naturally	count	instructions	predecessor
Distribute	gesture	amongst	banned	efficiency	enshrined	perfectly	educate	continental	facilitating	foremost
Stuff	marched	conclusion	vetting	waving	boasted	systematically	repair	classic	il	circles
Indicating	misguided	imagination	leading	skyrocketing	defends	responding	one-third	remind	confident	relying
Intensified	in-work	inspectors	landed	reid	hostage	angered	stealing	draws	transform	david
Damaged	scenario	syndication	cast	negotiate	fast-track	destructive	occupied	pervasive	systematically	vision
Cram	intensified	disregard	cheap	imagination	d-n.j	nail	vicente	continent	defining	prefer
Posts	punish	worsen	propose	ron	barbed	calm	oval	grappling	joseph	gove
Katie	copy	suggestion	favors	namely	blocks	publishing	Militarization	illustration	profound	controlled

The most positively weighted terms display variation across categories, which is useful for comparison. They also present some opportunities for further data cleaning; for example, one of the top-weighted UK terms is "Britain", which should have been included in the dictionary for removal from the beginning. LAC and Mexico also have "US-Mexico" as a top term, which also should be removed. One could make the same argument for "Manhattan" in the US list, though that is a bit

more ambiguous as it wasn't technically one of the words used to decide the location category, so the label isn't directly derived from it as a feature. Despite the fact that the results are not perfect, some of the adjectives are interesting; for example, "destructive" in Oceania and "militarized" in Mexico. They might begin to hint at differences across the location tags, though manually comparing the terms is time consuming.

The most negatively weighted terms, meanwhile, do not suffer from the same problem as the positively weighted terms in the sense that there are fewer geographically associated terms. Instead, there are again interesting adjectives ("damaged" decreasing the probability that a document is tagged as US, while "pervasive" decreases the probability that a document is classified as Asia). There is variation across the documents, but also some overlap, with "skyrocketing" appearing in both the LAC and Mexico lists.

Results – Cosine Similarity

Cosine similarity allows for comparison of the categories without needing to manually compare each of the lists. I compared the similarity of the 200 highest-weighted terms (100 most positively weighted and 100 most negatively weighted) from each model; these terms are referred to as "important features" going forward. The results are shown in Table IV below; cosine similarity scores range from 0 to 1, with a score closer to 1 indicating that documents are more similar.

Table IV: Cosine Similarity of High-Weighted Features

	US	UK	MENA	Mexico	LAC	Central Am.	Asia	Africa	Oceania	Canada	Europe
US		0.099	0.074	0.025	0.074	0.050	0.079	0.089	0.045	0.050	0.074
UK	0.099		0.045	0.030	0.045	0.035	0.059	0.069	0.050	0.035	0.153
MENA	0.074	0.045		0.079	0.064	0.050	0.114	0.099	0.045	0.054	0.040
Mexico	0.025	0.030	0.079		0.252	0.069	0.054	0.059	0.040	0.045	0.084
LAC	0.074	0.045	0.064	0.252		0.124	0.050	0.074	0.069	0.045	0.050
Central Am.	0.050	0.035	0.050	0.069	0.124		0.050	0.059	0.069	0.040	0.054
Asia	0.079	0.059	0.114	0.054	0.050	0.050		0.079	0.069	0.045	0.084
Africa	0.089	0.069	0.099	0.059	0.074	0.059	0.079		0.059	0.059	0.084
Oceania	0.045	0.050	0.045	0.040	0.069	0.069	0.069	0.059		0.084	0.069
Canada	0.050	0.035	0.054	0.045	0.045	0.040	0.045	0.059	0.084		0.069
Europe	0.074	0.153	0.040	0.084	0.050	0.054	0.084	0.084	0.069	0.069	

The models with the most similar important features are LAC and Mexico, Europe and the UK, Central America and LAC, Asia and MENA, the UK and the US, and Africa and MENA. All of these are intuitive, since they share regional overlap. This demonstrates one of the limitations of cosine similarity; identifying identical words would require further inspection, so we cannot immediately explain why they might be similar. For example, the Africa-MENA terms could be similar because the articles cover similar topics, or because the discourse used to describe immigration in and out of Africa is similar to the discourse describing immigration in and out of the MENA region.

The models with the least similar features are Mexico and the US, Mexico and the UK, Canada and the UK, Central America and the UK, and Europe and MENA. The fact that these

models have such different features is very interesting to me, particularly for major immigration corridors like Mexico-US and Europe-MENA. Since these regions were likely to be discussed in the same articles or at least deal with the same major events, the fact that their similarity score is so low may indicate that there are actual differences in the language used to describe the two regions. Of course, it is impossible to draw conclusions about what is driving these differences without manual comparison, which is even more difficult and subjective when dealing with words that are different, rather than the same.

Further Research

To further this analysis, I would improve my cleaning of the data by expanding my data dictionary to include more geographical variants of common regions. I would also try performing cosine similarity separately on positive- and negative-weighted terms so that I could see if there are more similarities or differences between the regions if I am only considering model weights that factor heavily in one direction. Finally, I would want to do a bit more domain research, and get into the weeds of term comparison, to understand what might actually be driving the similarities and differences that I noted above.

Bibliography

Lind, D. (2018) The migrant caravan, explained. Vox. https://www.vox.com/2018/10/24/18010340/caravan-trump-border-honduras-mexico

"World Bank Country and Lending Groups." World Bank. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups