Examining Xenophobia Using Twitter

November, 2022

Kevin Motia

INTRODUCTION

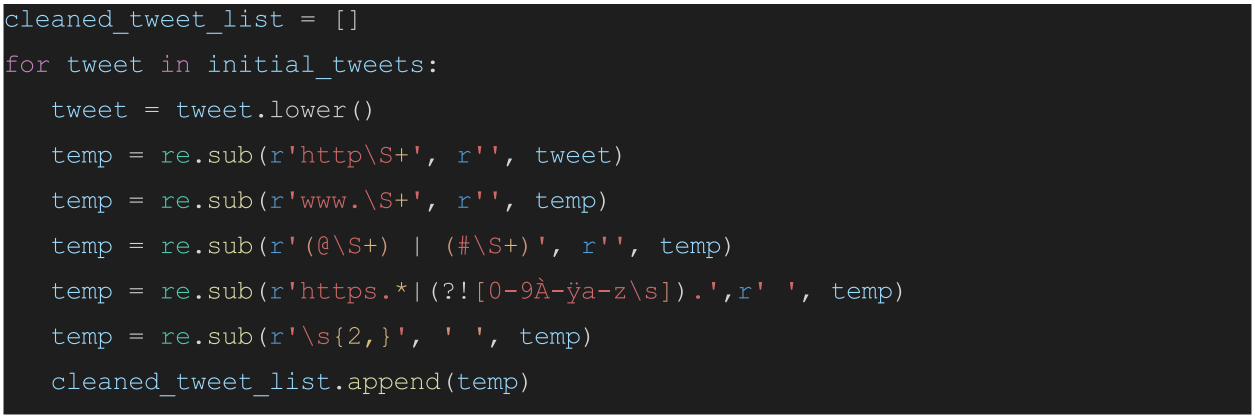
In the wake of Russia's invasion of Ukraine on February 24, 2022, eurocentric sentiment appeared within media outlets stationed in multiple countries [2]. During media coverage of the invasion, statements were made by members of multiple media organizations, both based in the U.S. and abroad, to express why they did not expect the outbreak of war within a European country [2]. The reporters noted the appearance of Ukrainians, and made xenophobic comparisons to MENA (Middle East and North Africa) countries in order to express their surprise at the invasion occurring in Europe [2]. Despite media outlets based in multiple countries demonstrating xenophobic sentiment, the scope of this project is limited to media originating within the U.S. The purpose of this project was to explore the differences in how people in the U.S. perceive people of various regions, and in particular understand if some regions are targeted with xenophobic tweets more than others. In order to do so, data was gathered from Twitter using a particular search query to obtain a set of xenophobic tweets from March 6, 2012 to December 5, 2022. The tweets were filtered by those that contained mention of specific countries, nationalities, or regions, and were categorized by their geographic regions. 500 tweets from the region-categorized data set were used to train a naive Bayes classifier to identify xenophobic tweets. Finally, the number of xenophobic tweets that mentioned each region were counted by region based on the day that the tweet was created. The resulting counts were used to generate visualizations which might allow for the inference upon how people in the U.S. perceive the different regions for which data was gathered. The report is structured as follows. In the Methods section, the methodology for gathering and filtering tweets are described along with how the visualizations were created. In part 1 of the Results section, we consider the time-series plot of counts of region-mentioned xenophobic tweets, and describe what it can be used for. In part 2 of the Results section, we examine the findings of a pie chart of the proportions of xenophobic tweets mentioning each region, and a table of the average count of xenophobic tweets attributable to each region. The report concludes with a discussion of the results, elaborating on sources of inaccuracy in the data sets, and how the methodology might cause the visualizations to be misleading. We end the paper by suggesting ideas for further development of the scope and methodology.

METHODS

A data set was scraped from Twitter by using Tweepy, a library for accessing the Twitter API [10]. A query was used to download tweets from March 6, 2012 to December 5, 2022. The query intended to included words and phrases that might yield tweets of xenophobic sentiment from the search.

```
query = ' ("send them back to" OR "send him back to" OR "send her back to"
OR "illegals from" OR "illegal aliens from" OR "illegal alien from" OR "illegal immigrants from"
OR "illegal immigrant from" OR "illegal criminals from" OR "illegal criminal from"
OR "foreign criminals from" OR "foreign criminal from" OR "illegal terrorist from"
OR "sending us their criminals" OR "ban people from" OR "deport them"                    (1)
OR "deport people from" OR "deport all these"
OR "immigrants from" OR "invading our country"
OR "being invaded" OR "invaders" OR "immigrants")
place_country:US -is:retweet'
```

The query returns tweets conditionally by using the OR operator, so tweets only containing one of the terms are returned at a time, and phrases stated earlier in the query are searched for first [16]. Additionally, the query only returns tweets that originate from the U.S., and are not retweets. The scraped tweets were cleaned by removing URLs, twitter mentions, hashtags, and special characters such as emojis. Stop words were removed at a later time in order to reduce the amount of irrelevant information given to a naive Bayes classifier. They were not removed at this time to make manual evaluation of tweet sentiment easier.



```
cleaned_tweet_list = []
for tweet in initial_tweets:
   tweet = tweet.lower()
   temp = re.sub(r'http\S+', r'', tweet)
   temp = re.sub(r'www.\S+', r'', temp)
   temp = re.sub(r'(@\S+) | (#\S+)', r'', temp)
   temp = re.sub(r'https.*|(?![0-9À-ÿa-z\s]).',r' ', temp)
   temp = re.sub(r'\s{2,}', ' ', temp)
   cleaned_tweet_list.append(temp)
```

Figure 1: Tweet cleaning procedure.

The cleaned data set was filtered by region using Geotext, a text classifier that returns information about a country by identifying country names and nationalities that are contained within a string of text. Geotext’s dependent documents were modified in order to return the region name instead of country abbreviation when it identified a country name, nationality, or specific region. Tweets analyzed by Geotext were appended to a data set. Countries that correspond to each region are color coded in the map seen in **Figure 2**, where the U.S. is omitted for the purpose of evaluating xenophobia of people in the U.S. towards *other* countries. Tweets that mentioned more than one region were duplicated for each mentioned region. The cleaned and region-filtered data set contained a total of 67,448 tweets.

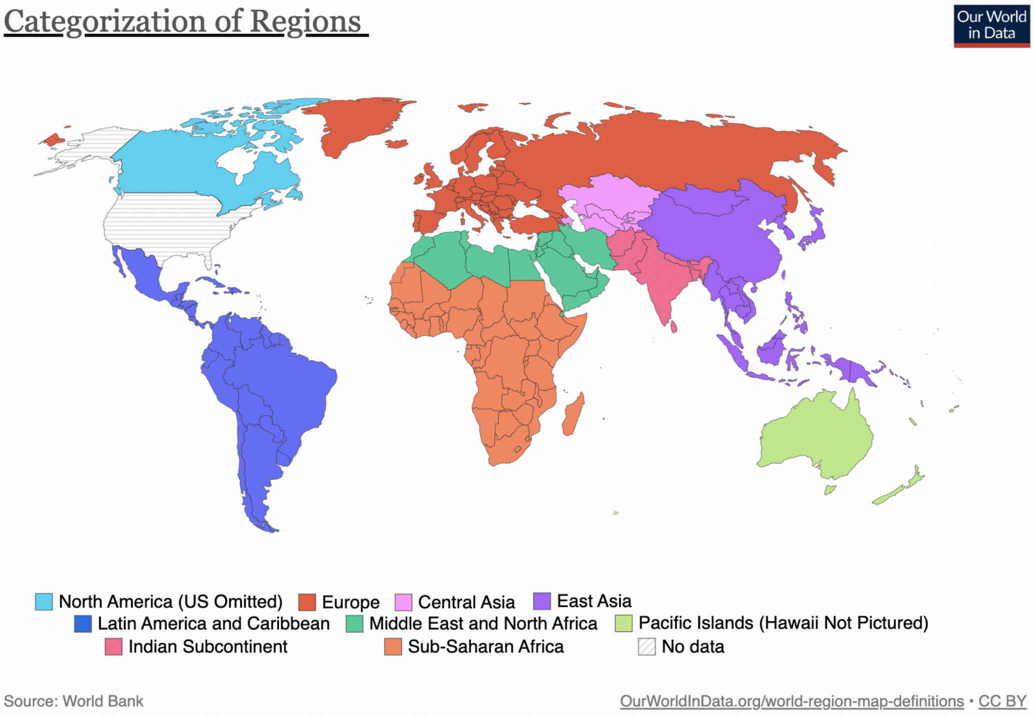
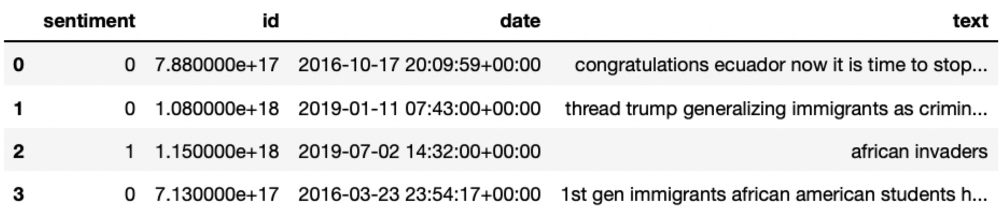

Categorization of Regions
Our World in Data
North America (US Omitted)
Europe
Central Asia
East Asia
Latin America and Caribbean
Middle East and North Africa
Pacific Islands (Hawaii Not Pictured)
Indian Subcontinent
Sub-Saharan Africa
No data
Source: World Bank
OurWorldInData.org/world-region-map-definitions • CC BY

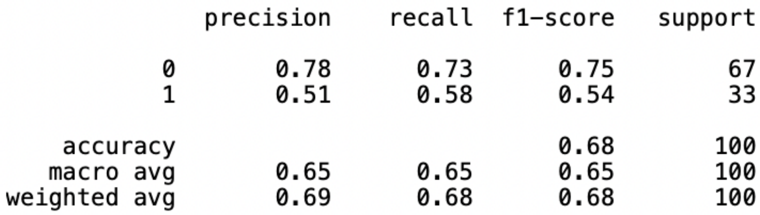
Figure 2: Visual categorization of geographic regions. Files for categorizing each region using Geotext contain countries, nationalities, and regions, and are linked from [15].

A sentiment column was added to the cleaned and region-filtered data set to prepare it to use the classifier for training, testing. This data set was then shuffled, and 500 tweets were separated to train and test the classifier. The remaining 66,948 tweets were then evaluated by the classifier. An example of the structure of the resulting data set can be seen in **Figure 3**.



	sentiment	id	date	text
0	0	7.880000e+17	2016-10-17 20:09:59+00:00	congratulations ecuador now it is time to stop...
1	0	1.080000e+18	2019-01-11 07:43:00+00:00	thread trump generalizing immigrants as crimin...
2	1	1.150000e+18	2019-07-02 14:32:00+00:00	african invaders
3	0	7.130000e+17	2016-03-23 23:54:17+00:00	1st gen immigrants african american students h...

Figure 3: First 4 rows of training and testing data set. This figure provides an example of how the data set was structured in order to correspond tweet sentiment with tweet text and other information. The same data frame format is used for the training/testing data set, and the data set for which sentiment was predicted using the classifier.

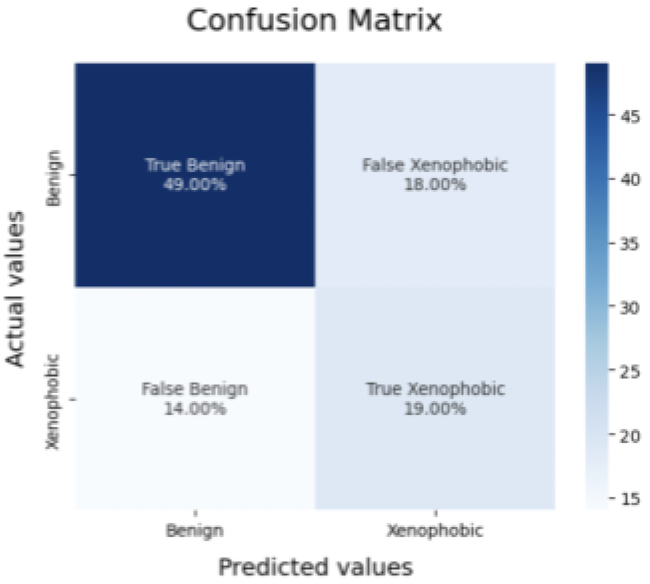


```
              precision    recall  f1-score   support

           0       0.78      0.73      0.75        67
           1       0.51      0.58      0.54        33

    accuracy                           0.68       100
   macro avg       0.65      0.65      0.65       100
weighted avg       0.69      0.68      0.68       100
```

Figure 4: Naive Bayes Classifier Report. The precision is intuitively the ability of the classifier not to label a benign sample as xenophobic. The recall is intuitively the ability of the classifier to find all the positive samples [6]. The F1 score can be interpreted as a harmonic mean of precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal [5]. The macro average is an unweighted average of the f1-scores of the benign (0) and xenophobic (1) categories, while the weighted average biases the f1-score of these categories based on the proportion of their support relative to the total support. The support is the number of tweets within a particular category.



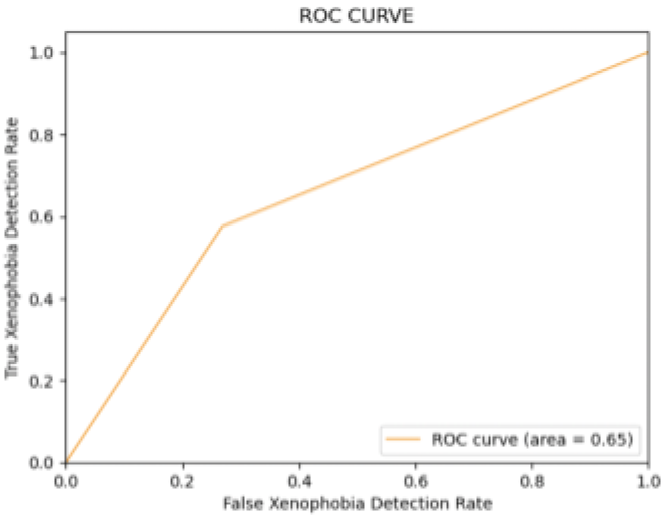
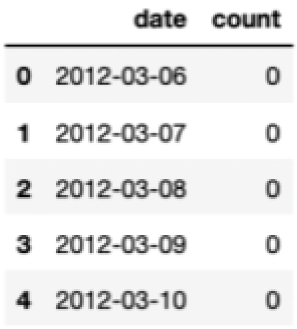


Figure 5: Confusion Matrix. The matrix shows the classifier's four possible outcomes when attempting to predict tweets as benign or xenophobic. The possible outcomes of prediction are shown, and are "true benign", "false xenophobic", "false benign", and "true xenophobic".

Figure 6: ROC Curve. The ROC curve features the rate of accurately detecting xenophobic tweets on the Y axis, and the rate of inaccurately detecting xenophobic tweets on the X axis [11]. ROC curve area is an indication of the classifier's ability to distinguish between the two possible classes of tweet [9].

The 500 tweets that were separated for training and testing were manually evaluated as benign (0), or xenophobic (1). 80% of these tweets were used for training, and 20% of these tweets were used for testing. A classifier report was generated for the 100 tested tweets as seen in **Figure 4**, and a confusion matrix and ROC curve were also generated in order to visualize the performance of the classifier as seen in **Figures 5** and **6.**

After these performance indicators were produced for the training/testing set, the naive Bayes classifier was used in order to predict the sentiment of 66,948 tweets. The resulting predicted tweets were categorized by their corresponding region by Geotext. Tweets that belonged to more than one region were duplicated again for each region they mentioned, which might have caused visualizations to be misleading, and is discussed later in the report. For each of these region-corresponding data sets, the number of xenophobic tweets were counted for each date. Dates between the start and end dates that did not appear within the data sets were added, and their counts were filled with a value of 0. The structure of these data sets is exemplified by **Figure 7**.



	date	count
0	2012-03-06	0
1	2012-03-07	0
2	2012-03-08	0
3	2012-03-09	0
4	2012-03-10	0

Figure 7: Count of detected xenophobic tweets by date for Latin America & Caribbean

Once the counts for each date were calculated, the resulting data sets were visualized on a single time-series plot in **Figure 8**. The time-series plot has labels for some political events that made news headlines on the same dates that correlate with spikes in detected xenophobic tweets in an attempt to provide an explanation for them. To accompany this time-series plot, and provide a more direct comparison between the region-specific data sets, **Figure 9** visualizes the number of xenophobic tweets by proportion attributed to each region.

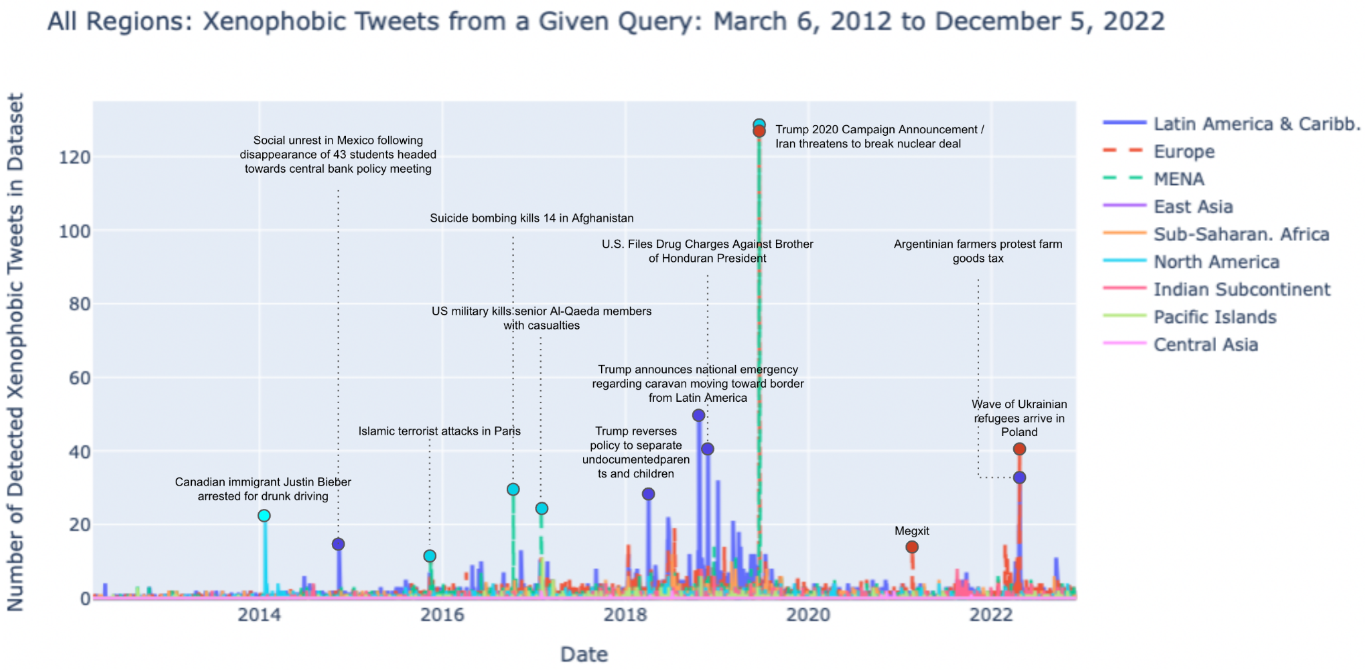


Figure 8: Time-series plot of counts of xenophobic tweets mentioning a specific region from March 6, 2012 to December 5, 2022. Events do not necessarily occur on the same date they are paired with, but they do appear in news headlines on those dates.

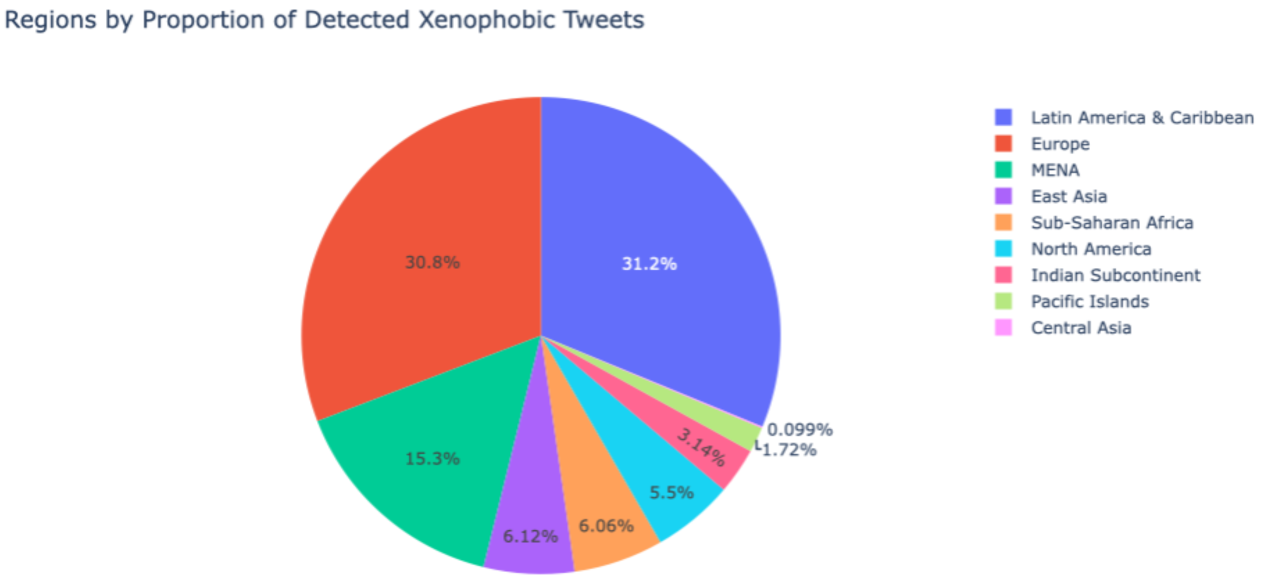
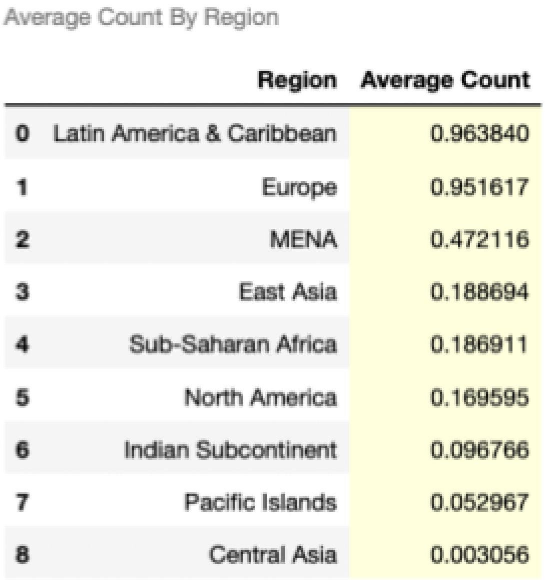


Figure 9: A pie chart of the proportions of xenophobic tweets attributable to each region from March 6, 2012 to December 5, 2022.



Average Count By Region

	Region	Average Count
0	Latin America & Caribbean	0.963840
1	Europe	0.951617
2	MENA	0.472116
3	East Asia	0.188694
4	Sub-Saharan Africa	0.186911
5	North America	0.169595
6	Indian Subcontinent	0.096766
7	Pacific Islands	0.052967
8	Central Asia	0.003056

Figure 10: Average count in xenophobic tweets mentioning a certain region from March 6, 2012 to December 5, 2022.

RESULTS

We first consider the purpose of the time-series plot, **Figure 8**. The plot was meant as an attempt at visualizing how different regions are reacted to with different counts of xenophobic tweets. It was also the purpose of this plot to correlate spikes in xenophobia with dates during which relevant events were in the news in an attempt to explain why tweets with xenophobic sentiment spiked.

Examining the plot, it appears that there are relevant political events that correlate with spikes in xenophobic tweets pertaining to each region. We observe that spikes larger than 20 counts were only attributable to the regions of Europe, Latin America & Caribbean, MENA, and North America.

For an example of an attempt at explaining a spike in **Figure 8,** we identify the largest one, which occurred for the regions of MENA and Europe. This spike correlates with news headlines about Donald Trump's 2020 U.S. presidential campaign announcement on June 18, 2019 [9]. It has been reported that Trump uses anti-immigrant rhetoric during his rallies [8], which might explain the sudden spike in xenophobic tweets directed toward the two regions if twitter users producing those tweets were inspired by Trump's xenophobic rhetoric. The spike in tweets mentioning these regions might also have been in part attributable to headlines about Iran's announcement that it planned to breach a multinational nuclear deal [7], although the event directly involves the acts of a country in the MENA region, indicating that it might be more likely that it correlates with its spike than Europe's.

To provide another attempt at explaining a spike in **Figure 8,** we identify the second largest spike, which occurred for the region of Latin America & Caribbean. This spike occurred on October 22, 2018, and correlates with a headline about Donald Trump declaring a national emergency with regard to a large number of people who were moving through Mexico and reportedly intended to cross the U.S.-Mexico border [17]. Xenophobic tweets may have been produced by twitter users in order to attack the group of traveling people in Latin America. Several other spikes in detected xenophobic tweets that mentioned particular regions correlated with region-relevant political events, but while **Figure 8** may provide an intuition of how the counts of xenophobic tweets attributable to each region differ, they do not provide a direct comparison.

For more straightforward ways to compare the counts of xenophobic tweets attributable to each region, we examine **Figure 9**, and **Figure 10** which respectively look at the proportions of xenophobic tweets attributable to each region, and the average count of xenophobic tweets attributable to each region. From **Figure 9**, we find that the regions of Latin America & Caribbean and Europe had the highest number of xenophobic tweets. The remaining regions in descending order of proportion were MENA, East Asia, Sub-Saharan Africa, North America, the Indian Subcontinent, Pacific Islands, and Central Asia. The same order of regions was found by **Figure 10**, which for each region, shows the average count in xenophobic tweets mentioning the region per day in descending order. Consequently, it was found that the regions in descending order of proportion of total xenophobic tweets attributable to a region correlated with the regions in descending order of average daily count.

DISCUSSION

During this report, we have touched on the uses for the data sets and visualizations. To elaborate on them further, we first consider a potential takeaway of **Figure 8**. It might be inferred from the time-series plot that the regions of Latin America & Caribbean, Europe, and MENA were targeted by tweets with xenophobic sentiment more due to more labeled events involving them relative to other regions during the time frame of March 6, 2012 to December 5. The data from this project does not provide evidence for this idea. Events correlating with spikes in the figure were sought out for the visualization, and it is likely that there are many other events involving the regions that did not correlate with spikes in xenophobic tweets. In order to further explore this idea, it might be beneficial to plot more events for all regions in order to provide more information to compare how the plots for each region generally behave through events pertaining to them.

It might be inferred from **Figures 9** and **10** that xenophobic tweets targeted at Latin America & Caribbean, and Europe were more prevalent during the time frame of March 6, 2012 to December 5, 2022. In order to provide support for this idea, improve these plots, and address the inference that tweets that mentioned particular regions targeted them, we turn our attention to the classification report, confusion matrix, and ROC curve seen in **Figures 4, 5**, and **6** to understand limitations of the data, and their implications on the figures.

From the classification report, it is apparent that the training/testing set is unbalanced. From the 100 tweets being used for testing, 67 were benign, and 33 were xenophobic. This means that the classifier was more likely to predict a given tweet as benign rather than xenophobic, causing the precision and recall of the benign tweets to be high relative to the precision and recall of the xenophobic tweets. The lack of balance in turn affected the f1-scores. The f1-score for the benign class would have been lower, and the f1-score for the xenophobic class would have been higher if the training/testing set was more balanced.

The information depicted by the confusion matrix appears to agree with that of the classification report in that xenophobic tweets were classified less accurately than benign tweets. The ROC curve presents an area under the curve of .65, where the best possible area is 1. The area under the curve is indicative of the classifier's ability to distinguish between benign and xenophobic categories, where a value of .7 is considered acceptable [12], indicating that our classifier did not produce accurate enough predictions of xenophobic and benign sentiment to be considered acceptable.

To elaborate further on reasons for the ineptitude of the classifier, it is possible that the amount of data used to train the classifier was insufficient to train the model. It is also possible that the manually evaluated data inconsistently categorized tweets as xenophobic and benign, leading to a classifier that could not classify sentiment with higher precision. The evaluator's potential ignorance and personal biases might have also impacted the evaluation of the tweets in the training/testing set, and in future iterations of this project it may be beneficial to have an expert manually evaluate the tweets for training and testing. Any incompetence of the classifier will have impacted all visualizations and results from the data it evaluated.

The visualizations and results could have also been impacted by the methodology. During the process of filtering tweets by region, tweets were duplicated for each region mentioned within it. Each of these tweets were duplicated again when they were being assigned to their region-specific data sets. This might have caused regions that had a propensity to be mentioned along with other regions to be duplicated more often than other regions, causing them to have a larger number of xenophobic tweets.

Another potential source of inaccuracy in this report pertains to the data collected from the regions of East Asia, Central Asia, and the Indian Subcontinent. Words that directly mentioned regions were among those that were used to categorize tweets into the data sets for the regions they mentioned. For East Asia, these words were "east asia" and "east asian". Similarly, for Central Asia they were "Central Asia" and "Central Asian". Some people may not draw a distinction between Central Asia and East Asia, and might have referred to these regions without a cardinal direction by simply referring to them by "asia" or "asian". Xenophobic tweets that contained these words were not picked up by Geotext, and thus did not make it into any visualized data set, potentially resulting in misleading visualizations for **Figure 8**, **9**, and **10**. Likewise, data for the Indian Subcontinent may have been similarly impacted because the documents that our modified Geotext uses to categorize tweets fail to include the term "south asia", "south asian" , or "asian" as identifiers for the Indian Subcontinent.

Another fault of Geotext to be particularly mindful of when reading the figures is that the tweets are categorized by regions mentioned within them, not regions that are targeted by them.

when will germany realize tourist dont want to come to your country on vacations holidays anymore these immigrants are hurting your tourism *(2)*

(2) shows a tweet that is generally xenophobic, but does not target any particular region. This tweet was placed in the data set corresponding to the Europe region. While it was our intention to leave tweets like this out of the data sets, Geotext was unable to make the distinction between a targeted region and a mentioned region. If it is assumed that the tweets are all region-targeting when reading the visualizations, **Figures 8, 9,** and **10** will be misleading.

To conclude the findings of this report, the analysis of the classifier indicates that it was not accurate, so the visualizations created from data evaluated by it may be misleading. In future iterations of this project, it would be beneficial to refine the methodology to produce a more accurate tweet classifier, and categorize xenophobic tweets to the regions that they target instead of mention. The project scope might also be widened to include the analysis of racist tweets to understand any correlation that xenophobia towards particular regions might have with the majority race of those regions. We plan to revisit this project to make such changes in a later course.

REFERENCES

1. Abdullah. (2022, July 22). Twitter sentiment analysis using Sklearn and NLTK. Retrieved December 16, 2022, from https://medium.com/red-buffer/twitter-sentiment-analysis-using-sklearn-and-nltk-3bd79521141a
2. Bayoumi, M. (2022, March 02). They are 'civilised', 'European' and 'look like us': The racist coverage of Ukraine. Retrieved December 01, 2022, from https://www.theguardian.com/commentisfree/2022/mar/02/civilised-european-look-like-us-racist-coverage-ukraine
3. Countries and economies. (n.d.). Retrieved December 12, 2022, from https://data.worldbank.org/country
4. Davies, C. (2021, February 19). Harry and Meghan will not return as Working Royals, says palace. Retrieved December 16, 2022, from https://www.theguardian.com/uk-news/2021/feb/19/harry-meghan-tell-queen-not-return-working-royals
5. Developers, S. (n.d.). Sklearn.metrics.f1_score. Retrieved December 11, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
6. Developers, S. (n.d.). Sklearn.metrics.precision_recall_fscore_support. Retrieved November 9, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html
7. Erlanger, S. (2019, June 18). As U.S. and Iran face off, Europe is stuck in the Middle. Retrieved December 12, 2022, from https://www.nytimes.com/2019/06/18/world/europe/iran-us-nuclear-europe.html
8. Fritze, J. (2019, August 21). Trump used words like 'invasion' and 'killer' to discuss immigrants at rallies 500 times: USA today analysis. Retrieved December 12, 2022, from https://www.usatoday.com/story/news/politics/elections/2019/08/08/trump-immigrants-rhetoric-criticized-el-paso-dayton-shootings/1936742001/
9. Haberman, M., Karni, A., & Michael. (2019, June 18). Trump, at Rally in Florida, kicks off his 2020 re-election bid. Retrieved December 13, 2022, from https://www.nytimes.com/2019/06/18/us/politics/donald-trump-rally-orlando.html
10. Harmon, Roesslein, J., & other contributors. Tweepy [Computer software]. https://doi.org/10.5281/zenodo.7259945
11. Hernandez, S. (2022, August 29). Language use in the United States: 2019. Retrieved December 14, 2022, from https://www.census.gov/library/publications/2022/acs/acs-50.html
12. Mandrekar, J. N. (2015, November 20). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology, 5*(9), 1315-1316. doi:10.1097/jto.0b013e3181ec173d
13. Palenzuela, Yaserand M., & other contributors. Geotext [Computer software].https://github.com/elyase/geotext
14. Scikit-Learn, D. (n.d.). Multiclass receiver operating characteristic (ROC). Retrieved December 15, 2022, from https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#:~:text=This%20example%20describes%20the%20use,FPR)%20on%20the%20X%20axis
15. Supplementary material for this article is available online at https://github.com/kdm004/pocsFinalProject, https://github.com/kdm004/pocsProject, https://github.com/kdm004/modified_tubular
16. Search tweets - how to build a query | docs | twitter developer platform. (n.d.). Retrieved December 14, 2022, from https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query
17. Trump says he has alerted military to 'national emergency' of migrant caravan. (2018, October 22). Retrieved December 10, 2022, from https://www.theguardian.com/us-news/2018/oct/22/trump-migrant-caravan-central-america-military