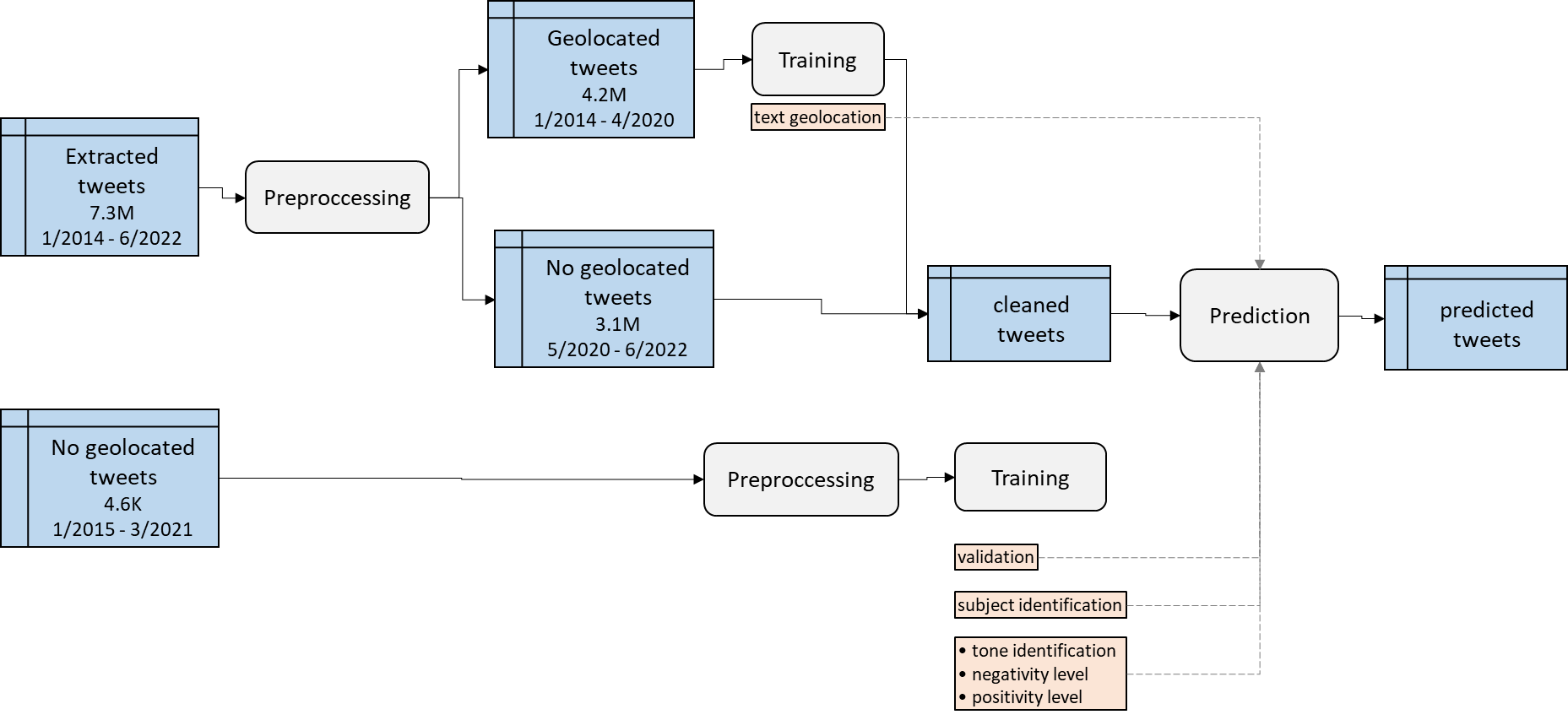
**Exploratory analysis of Venezuelan migration Twitter data through manual and automated approaches**

# Methodology

During this approach, the method used was divided into the following steps.



## Data collection and preprocessing

We used the Twitter Streaming API to collect 7.3M tweets using the query in Figure 1 in the range of January 2014 and June 2022.

(((venezolano OR venezolana OR venezolanos OR venezolanas OR venezuela OR vzla OR vnzla) (migrante OR migrantes OR migracion)) OR veneco OR veneca OR venecos OR venecas)

Figure 1. Tweets extraction query.

Those tweets were not geolocated via the Twitter API functionality because running different tests and found that the number of tweets that matched was too low; they will be filtered using an algorithm about it.

The first approach to geolocation was using the Nominatim API that relies on OpenStreetMap. We obtained 4.2M users’ locations using the API and the public information in his profile. However, this process was restricted to the API rate limits establishing the maximum of one call per request. In this large dataset, the total amount of time would be around 90 days in a row. So, we did the geolocation for the first 4.2M tweets for 51 days straight. With this, we build a document geolocation model that uses the text of the tweet to predict whether it is from Colombia or not and then use it to geolocate the remaining 3.1M tweets left.

## Data labeling

A sample dataset of 4K tweets between 2015 and 2021 was selected for manual labeling.

There were trained 6 NLP classification models in total. Listed here:

|  |  |  |
| --- | --- | --- |
| **Model name** | **Purpose** | **Levels of the classification variable** |
| Geolocation\* | Identify whether a tweet is from Colombia or not using the tweet text as input | • Colombia  • another country |
| Validation | Identify whether a tweet is talking about the Venezuelan migration topic using the tweet text as input | • Valid  • Not valid |
| Subject | Identify who or what the tweet is talking about | • Migrants  • Migration  • Government  • Geopolitics |
| Tone | Identify the tone of the tweet based on its text. | • Positive  • Negative  • Neutral |
| Positivity | Identify the level of positiveness of the tweet based on its text (when applicable). | • 1 (low)  • 2 (medium)  • 3 (high) |
| Negativity | Identify the level of negativeness of the tweet based on its text (when applicable). | • -1 (low)  • -2 (medium)  • -3 (high) |

\*The geolocation model was not trained with the 4K dataset sample but using the 4.2M tweets geolocated previously with the Nominatim API.

We used this criterion for the classification of the tone:

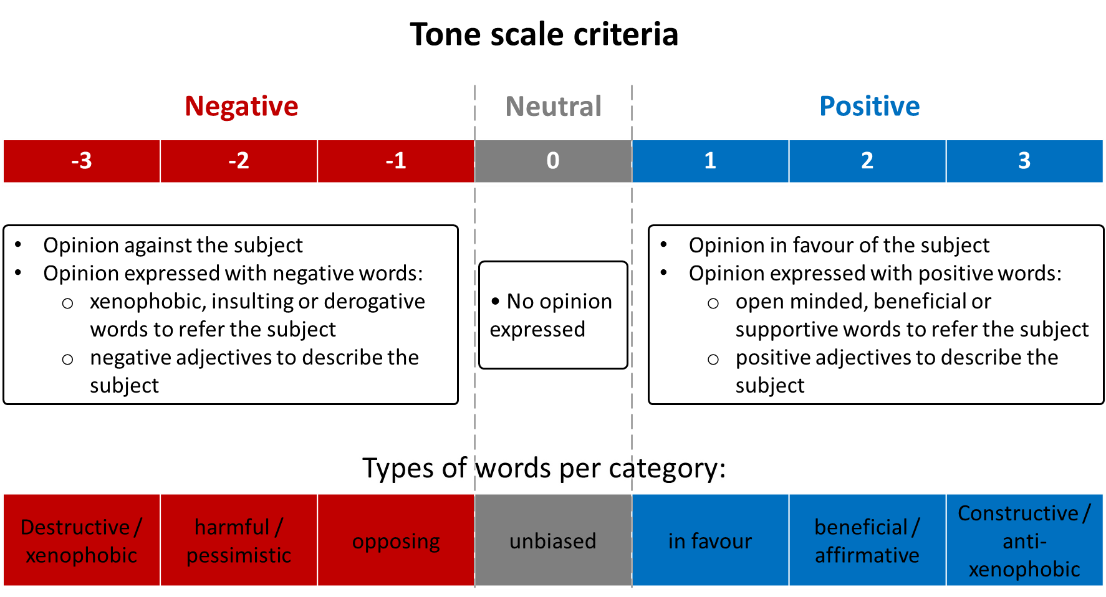


Figure 2. Criteria used for the tone scale labeling.

## Models training

After the filtering, this table shows how many tweets were used for every model training and testing. We used 70% of them for training and 30% for testing.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Training size** | **Testing size** | **Total size** |
| Geolocation | 189700 | 81300 | 271000 |
| Valid | 2170 | 939 | 3109 |
| Subject | 824 | 356 | 1180 |
| Tone | 1344 | 576 | 1920 |
| Negativeness | 441 | 189 | 630 |
| Positiveness | 90 | 42 | 132 |

[metrics paper]: METRICS FOR MULTI-CLASS CLASSIFICATION: AN OVERVIEW

# of epochs will vary for HuggingFace models to reduce the training time

These were the models used and the performances:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model 0: geolocation** | | | | | |
| **Source** | **Architecture** | **Training size** | **Testing size** | **epochs** | **Training time** (h:m:s) |
| SpaCy | es\_core\_news\_sm | 189700 | 81300 | 4\* | 9:12:47 |
| bert-base-spanish-wwm-cased | 1 | 13:02:00 |
| Hugging Face | bert-base-multilingual-uncased-sentiment | ❌ | ❌ |
| bert-base-multilingual-uncased-sentiment with hyp. optimization | ❌ | ❌ |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model 0: geolocation** | | | | | |
| **Source** | **Architecture** | **Accuracy** | **Precision** | **Recall** | **F1** |
| SpaCy | es\_core\_news\_sm | 0.794 | 0.794 | 0.794 | 0.794 |
| bert-base-spanish-wwm-cased | 0.5 | 0.25 | 0.5 | 0.333 |
| Hugging Face | bert-base-multilingual-uncased-sentiment | ❌ | ❌ | ❌ | ❌ |
| bert-base-multilingual-uncased-sentiment with hyp. optimization | ❌ | ❌ | ❌ | ❌ |

## Prediction

# Results

## Second dataset 7.4M

### Exploratory Analysis

From all the 7.4 million tweets, supported by the geolocation and valid tweets model, we filtered them to have a better sample. The first step was geolocating the remaining 3.2 M tweets that were not geolocated with the Nominatim API. The second step was location filtering; this led to the highest reduction because from the 7.4 M tweets, only 1.8M were from Colombia. The last step was filtering only to utilize the tweets about the Venezuelan migration topic. After the previous filter, the total valid tweets were 1.5M. This workflow is represented in Figure 1.

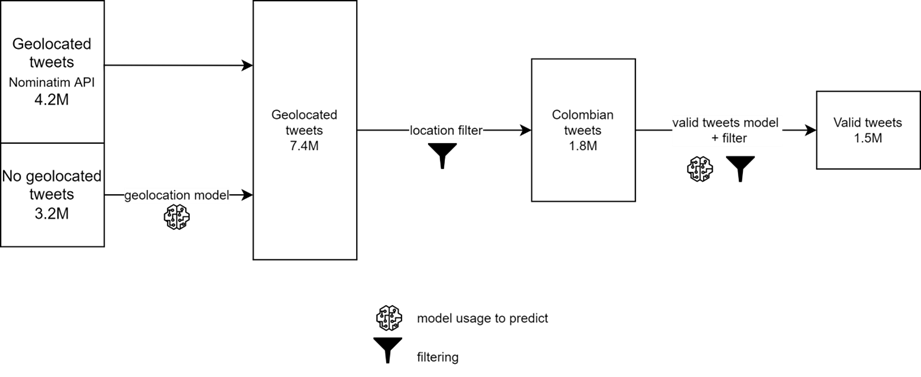


Figure 3. Tweets filtering process.

In Figure 2, we can see how those 1.5M tweets were distributed over the years. It shows that from 2014 to 2017, the number of tweets was lower than 10,000 every and an increasing number for the next years coming. This trend is clearer to follow in Figure 3.

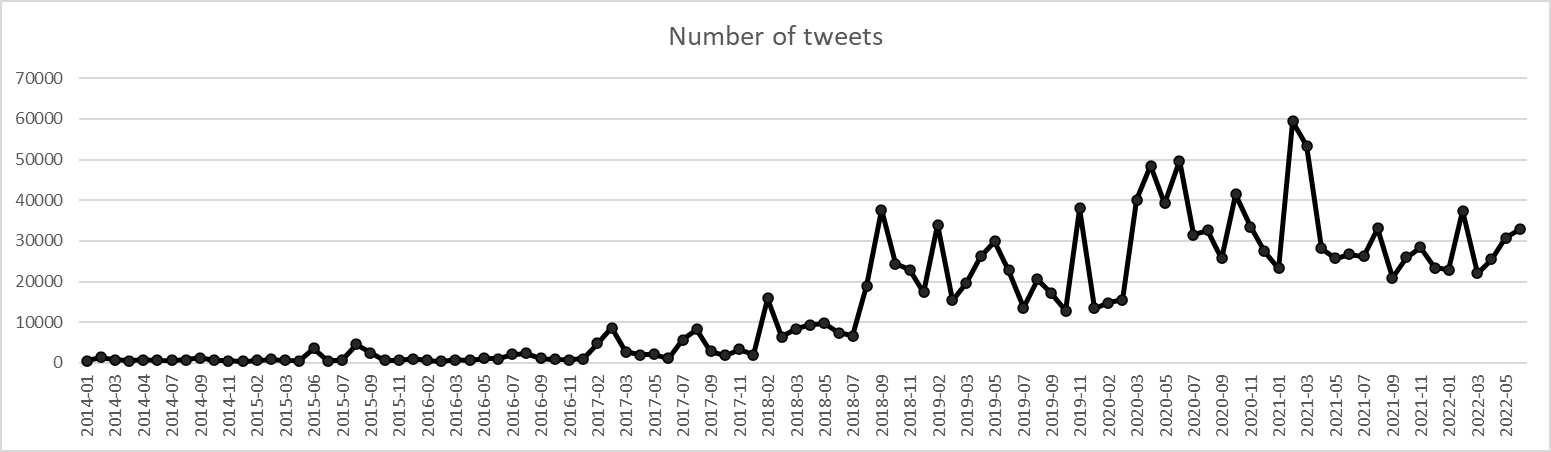


Figure 4. Monthly number of valid tweets

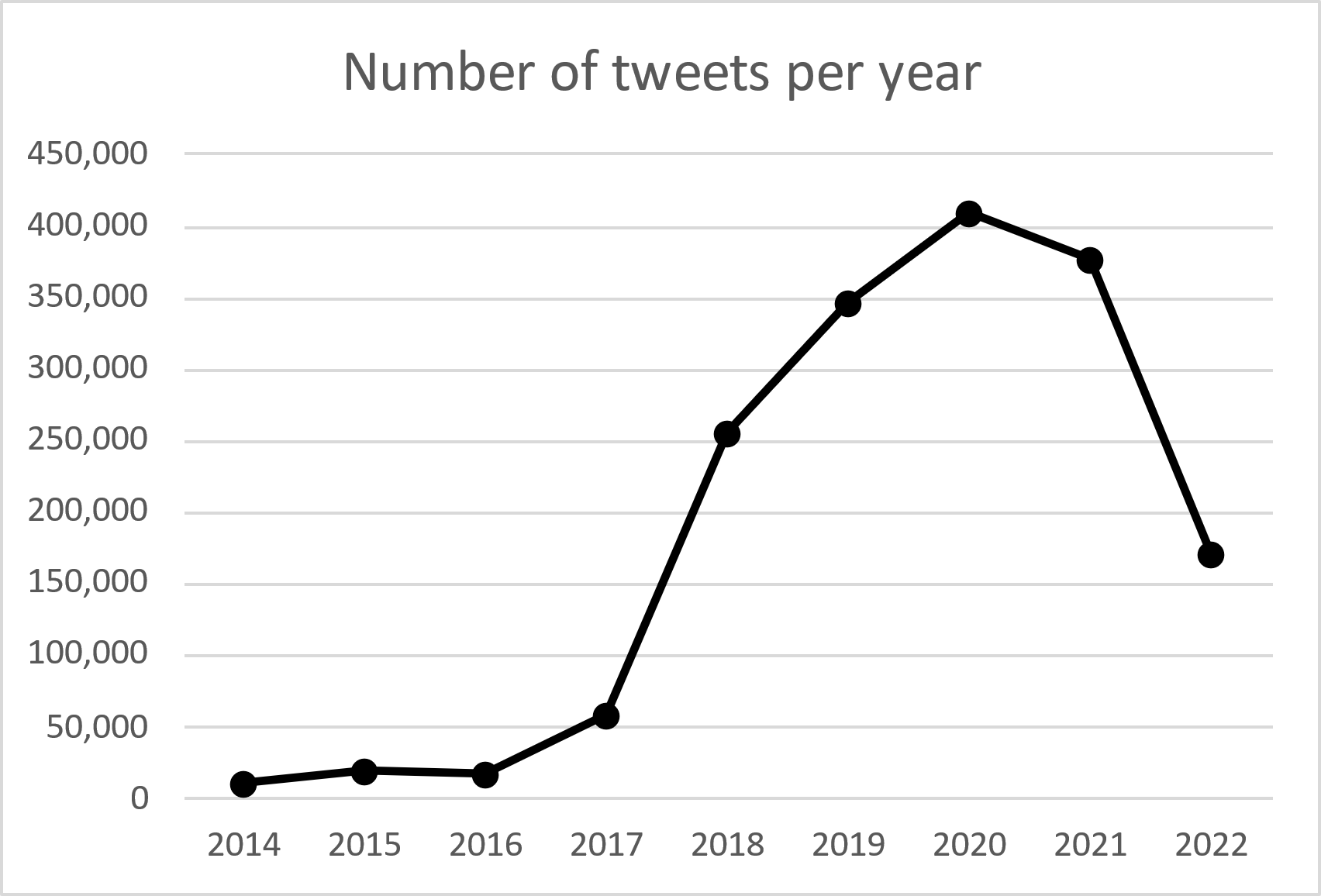


Figure 5. The yearly number of valid tweets.

With this large dataset, it is normal to have a higher number of retweets since all of them are captured. It is shown in Figure 6 the proportion of retweets over the years. This leads us to an increasing proportion of retweets from 23% up to 74% in 2018, and then a slight decrease until 52% in 2022.

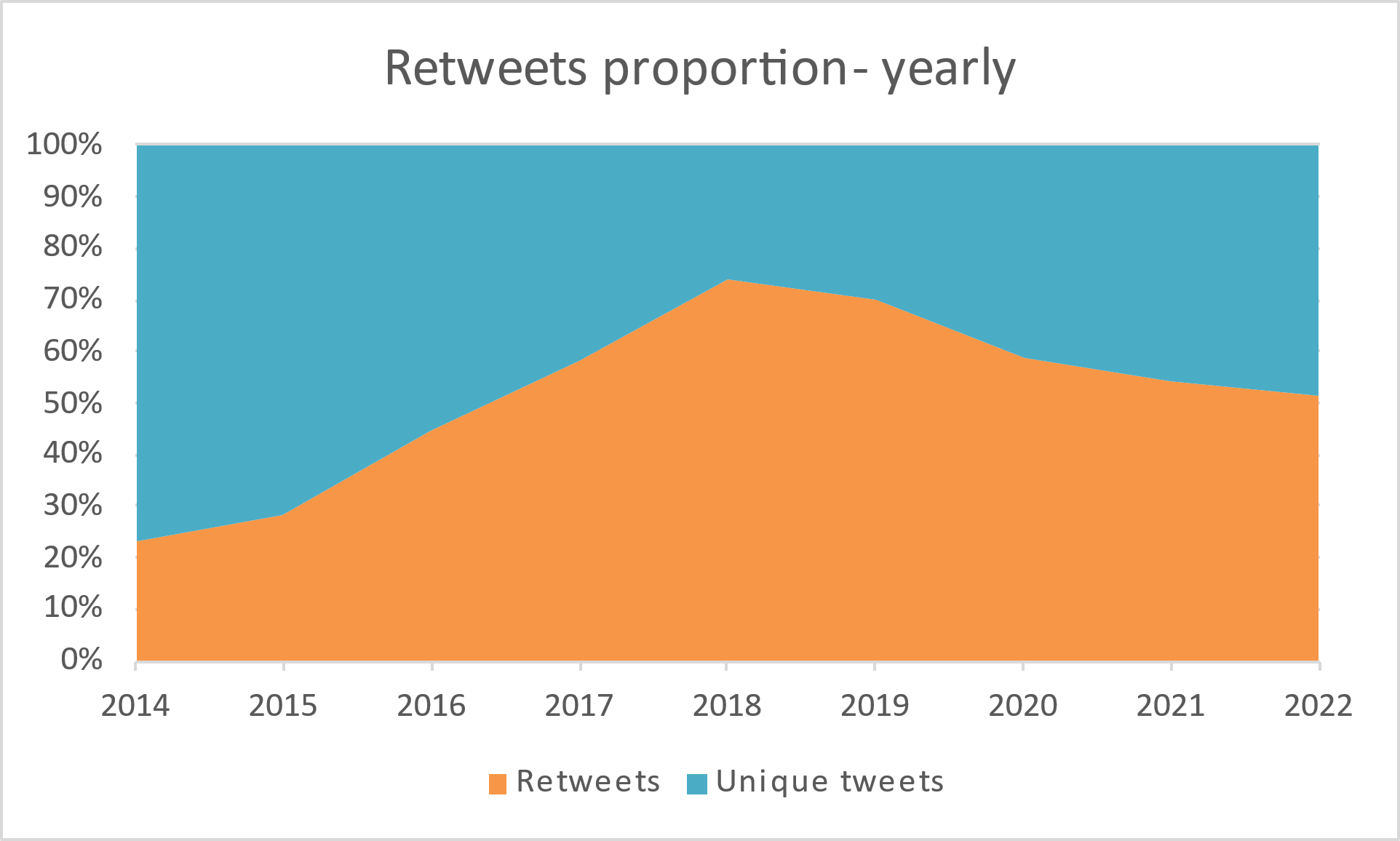


Figure 6. Yearly distribution of retweets

The impact of this huge number of retweets will be discussed in the following analysis.

### Sentiment Analysis

Sentiment analysis determines whether data is positive, negative, or neutral. The proportion of tweets with every type of tone was graphed in Figures 7 and 8. Figure 7 shows this proportion only for the unique tweets, omitting the retweets. The mean proportion of negative tweets is 55%, neutrals are 25% and positives are 20%. Figure 8 shows the proportion among all the tweets. The impact of retweets in the tone proportion can be noticeable given the diminishment to 44% the portion of negative tweets and was driven by an increasing in the portion of neutral tweets to 35%.

|  |  |
| --- | --- |
| Figure 7. Tone yearly proportion without retweets. | Figure 8. Tone yearly proportion. |

To measure how big was the impact of retweets in the conversation, Figure 9 shows the relative increment in tweets number caused by the insertion of retweets. The neutral tweets are the most impacted by that addition followed by the positive tweets. The negative tweets are less impacted the more negative they are.

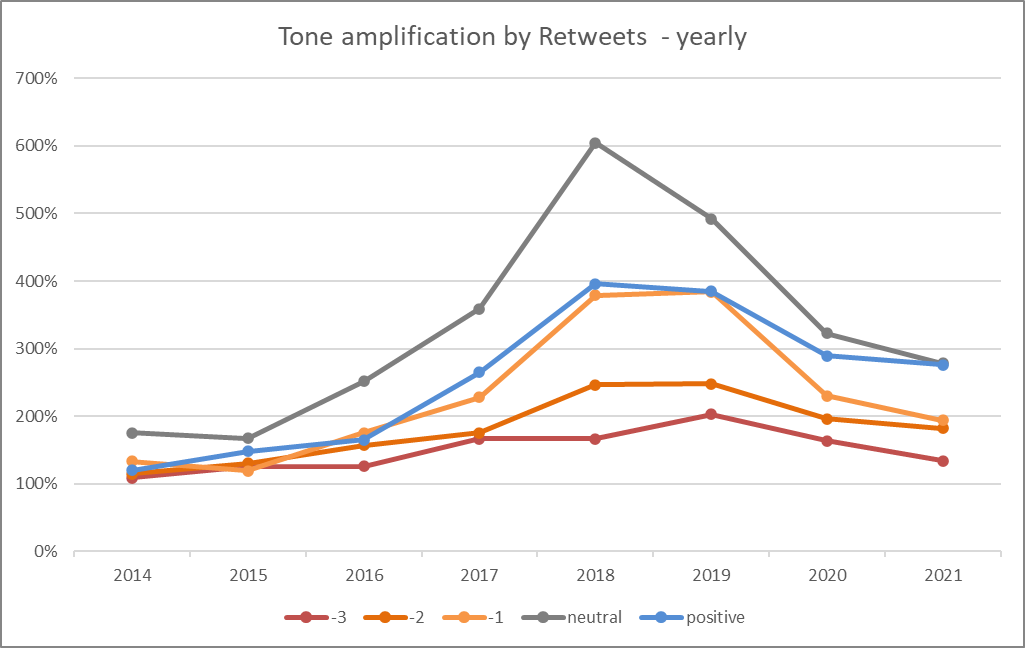


Figure 9. Amplification in tweets tone by retweets by year.

### Xenophobia Identification

The number of hate speech tweets within the total sample was calculated using a rule-based approach. The lexicon utilized is based on hate speech words that the Colombians typically use to insult people; many of them are from Colombiamágica[[1]](#footnote-1). There is a segmentation between insults to women and men, using the grammatical gender of the insulting word. As shown in Figure 11 and Figure 12, this number is always increasing with some negativity peaks in certain months like November 2019 and June 2020.

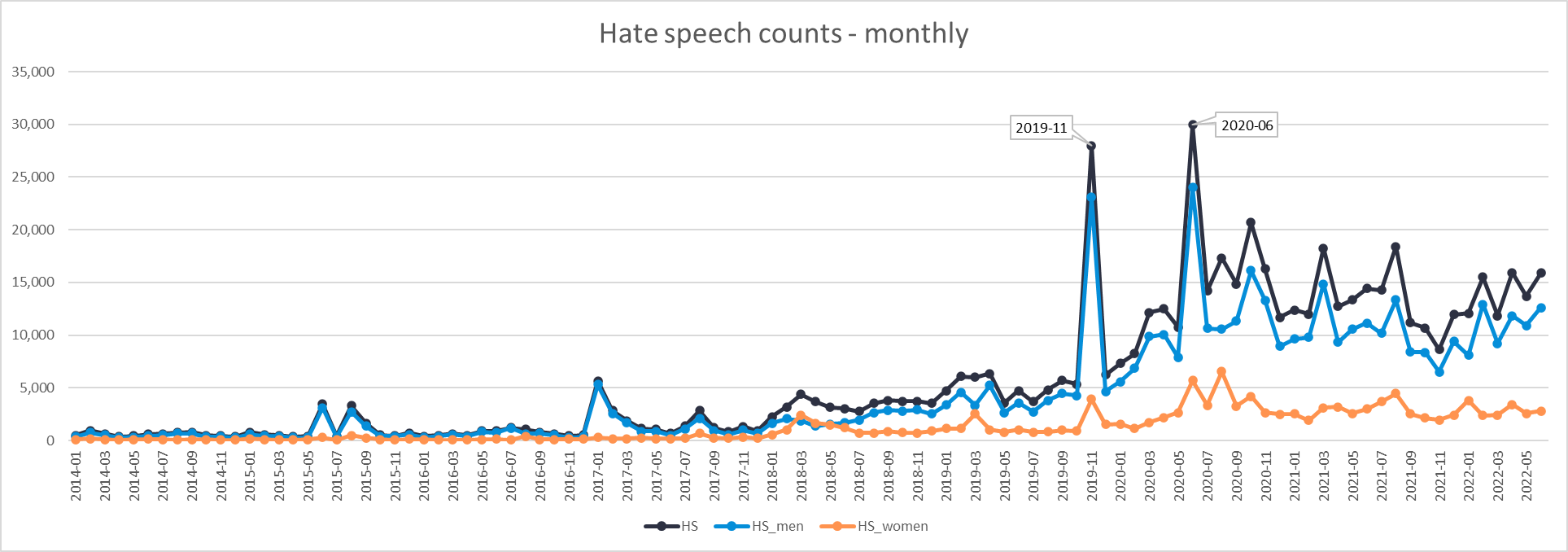


Figure 10. Monthly hate speech tweets count

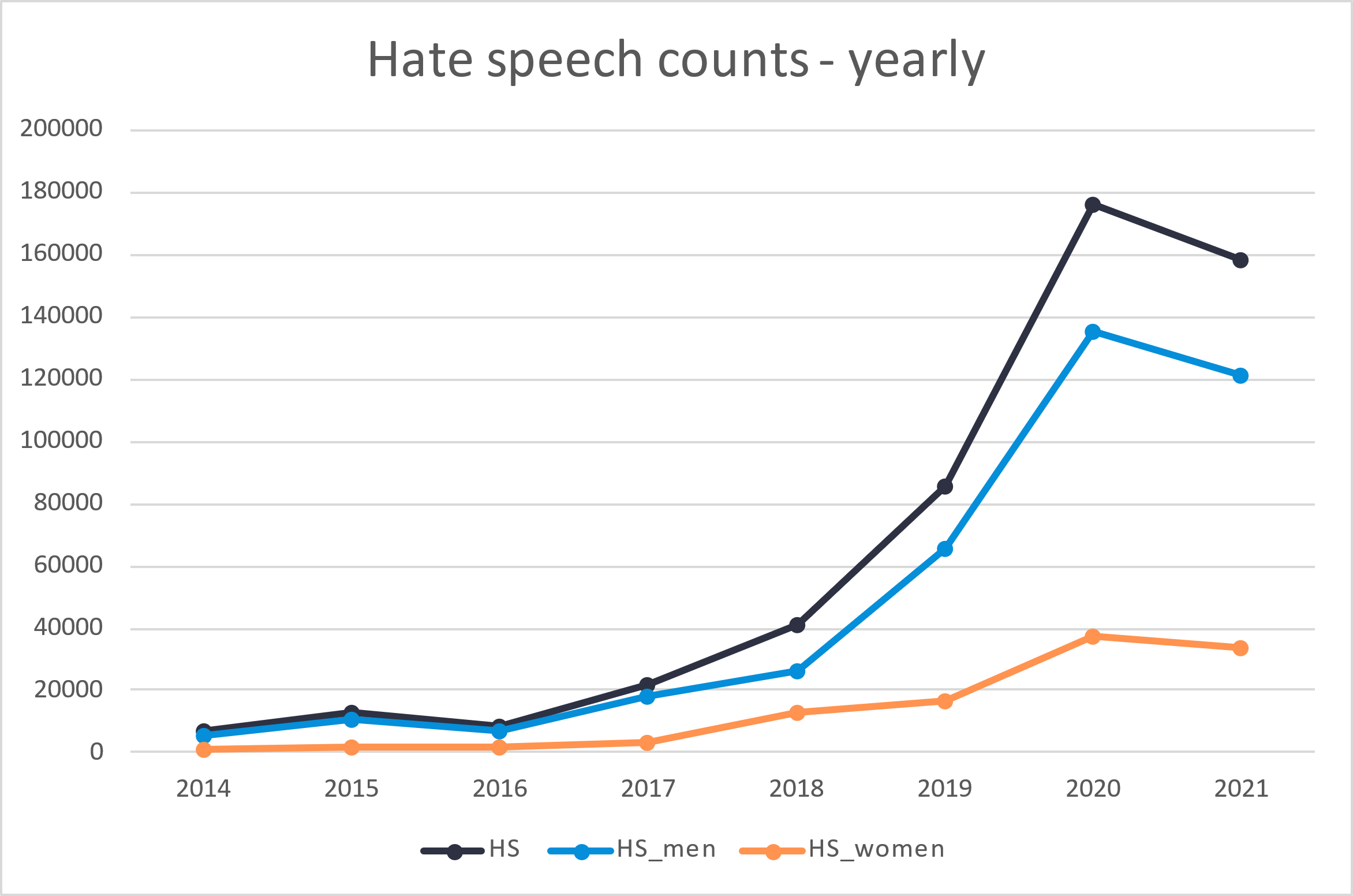


Figure 11. Yearly hate speech tweets count.

### Frustration type frequencies

For the tweets, frustration frequency was captured as number of tweets with negative tone in the specific timeframe. Figures 13 and 14 show the frequency of negative tweets by every category. The Tweets about migrants were predominant during all years, followed by government and geopolitics. huge increases

Unusual increases in Tweets about the Colombian government happened in January 2019 and June 2021. The same during March 2020 and October 2020 for migrants. Comparing them with the hate speech frequencies they follow the same pattern, showing the

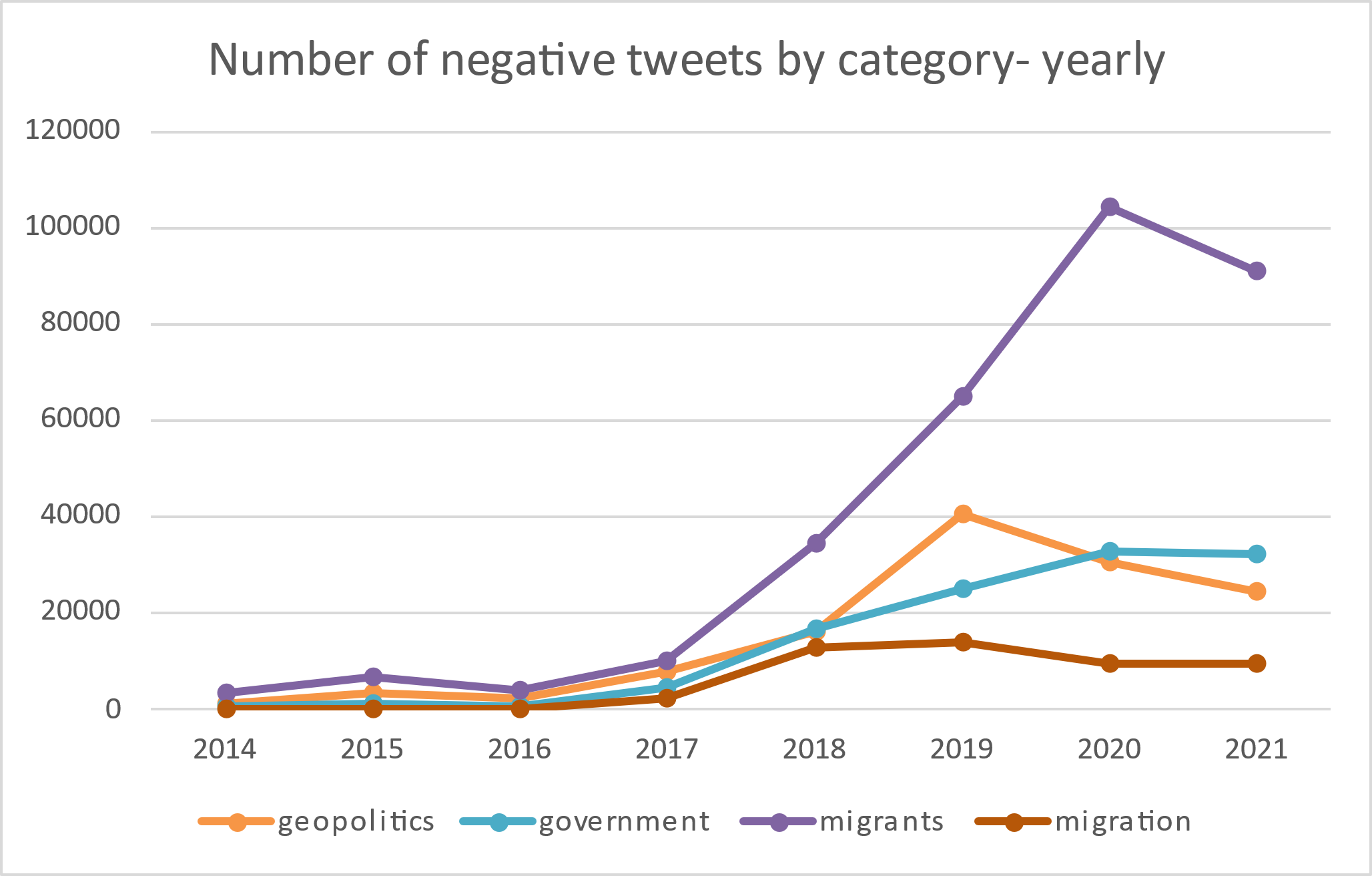


Figure 12. Yearly frequencies by frustration type

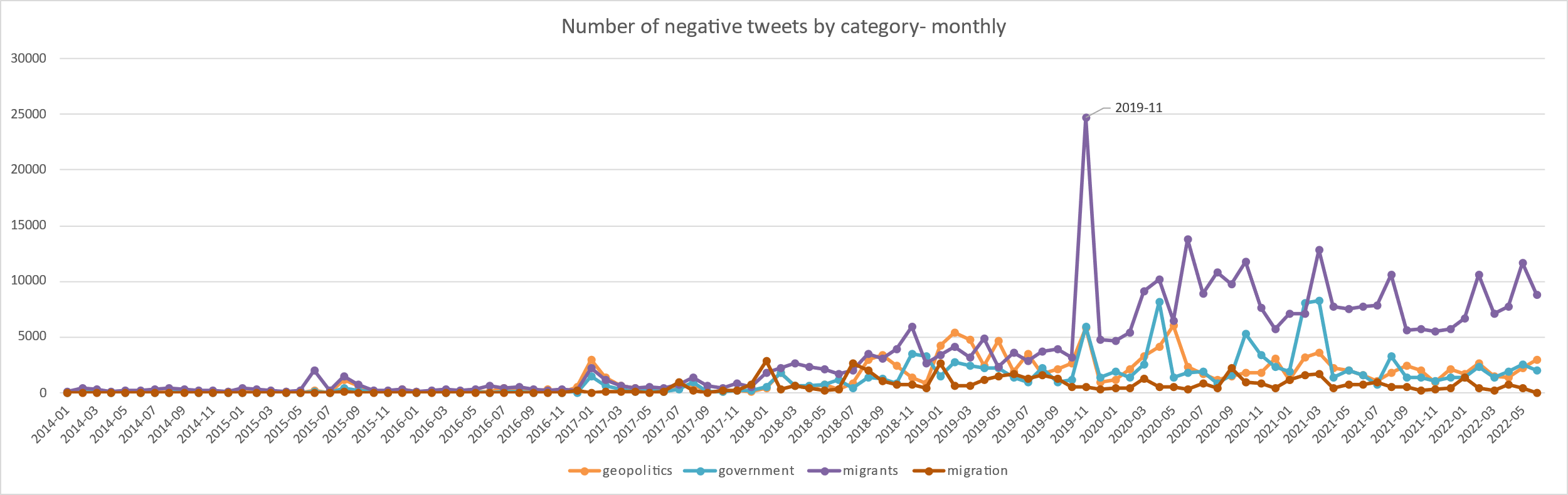


Figure 13. Monthly frequencies by frustration type

1. [Los Insultos y groserías más populares de Colombia | Groserías de Colombia (colombiamagica.co)](https://www.colombiamagica.co/entretenimiento/insultos-y-groserias-populares-en-colombia) [↑](#footnote-ref-1)