From Playboy Cricketer to Islamist Politician?
An Analysis of Imran Khan's Tweets Pre and Post Aug 2018
Author: Maryam Khalid Shah

Background

Former captain of Pakistan's cricket team, Oxford graduate, ex-husband of a British billionaire tycoon's daughter and a friend of Princess Diana's, Imran Khan was known in Pakistan and around the world as a cricket hero, playboy and a sex symbol in the 90s and the early 2000s. Khan created his own political party in 1996, and in 2018, Khan, a self-proclaimed liberal (add footnote) was elected as Pakistan's Prime Minister.

As Prime Minister, he stressed on Islam being the fulcrum of the Pakistani state and society[1]. Some people however, think that he is a political opportunist using religion (Islam) to appeal to Pakistan's majority (approximately 96%) Muslim population.

Theory of Interest

One aspect of exploring whether Imran Khan is using religion strategically i.e. to appeal to the people, is to analyze his "religious tweets" (i.e. tweets with any religious elements) over time. It is important to note here that by religious tweets I do not mean tweets that only talk about religion. I also mean tweets that use religious phrases and elements e.g. a tweet talking about the Kashmir issue containing religious words. The reason why I want to consider the latter as well is because even the use of a few religious terms in otherwise non-religious tweets still signals to readers that religion is important to Imran Khan.

1. ***Has the number of Imran Khan's tweets that contain religious (specifically Islamic) elements increased over time?***
2. ***Have Imran Khan's (religious) tweets become <u>more</u> religious over time?***

The first question is not differentiating between tweets that are more versus less religious. Instead, it is looking at **all** tweets that contain Islamic elements, and determining whether the amount of these Islamic tweets have increased over time?

The second question is looking at whether the religious elements within his religious tweets have become more pronounced. **Specifically, has the number of Islamic terms within his religious tweets increased over time?** Religious tweets here would be considered to be all tweets that contain religious elements, not only tweets in which religion is the *dominant* element/topic.

Even though Khan used religious statements prior to getting elected, I was interested to find out whether his tweets became more religious after he became Prime Minister as compared to his

[1] https://quillette.com/2021/10/23/from-playboy-sports-star-to-islamist-politician-the-strange-turn-of-imran-khan/

pre-office tweets. Therefore, in addition to looking at his tweets over time generally, I compared his tweets before he assumed office (August 2018) and his tweets as Prime Minister.

Assumptions

This report is for anyone interested in Pakistani politics, including supporters and non-supporters of Imran Khan. An assumption going into the project was that some of Imran Khan's tweets would contain 'Islamic terms', and a major simplification or subjective decision that had to be made was regarding selecting terms that make tweets 'Islamic' – this is discussed in more detail throughout the report.

Data Collection

I scraped Imran Khan's tweets from December 2014 to May 2022 - this included an equal number of months (45) pre and post Aug 2018 when Imran Khan became Prime Minister. I used the Python package snscrape[2] to scrape the tweets from Imran Khan's official Twitter account[3] and store them in a JSON file.

While most of Imran Khan's tweets are in English, lately he has been posting the same message in an English and an Urdu tweet separately. This made data cleaning easier as I could remove Urdu tweets completely without having to translate them. The package langdetect[4] was used to detect the language of the 5259 total tweets scraped. Keeping only English tweets resulted in a final dataset of 3911 tweets.

Data Cleaning

I started by removing numbers from the tweets. Next, using a regular expression, I removed URLs (Uniform Resource Locators), as they are references to a location on the web but do not provide any additional information. Even though only English tweets were selected, some still included Urdu hashtags that needed to be removed. I first identified which tweets had Urdu hashtags, then removed the Urdu hashtags from those tweets. Tweets often contain mentions of other Twitter users as well, that does not add too much information. Therefore, I decided to remove all mentions, and converted the text to lowercase.

Even though only English tweets were kept and Urdu hashtags were removed, some tweets still contained Urdu words. This is because even though Urdu is written in an Arabic script, a lot of Urdu speakers use Roman letters to write Urdu words. For instance, while the tweet "Thank you for coming to the jalsa" is in English, it contains an Urdu word, "jalsa" (جلسہ) written in Roman letters. After manually going through a lot of tweets, a couple of Roman Urdu words were identified, as can be seen in the (shortened) table below. Each of the identified words was translated to English.

[2] https://github.com/JustAnotherArchivist/snscrape
[3] https://twitter.com/ImranKhanPTI?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor
[4] https://pypi.org/project/langdetect/

Urdu Word	English Translation
jalsa	protest
haqeeqi	real
kutchery	court
junoon	enthusiasm

The tweets also included a lot of abbreviations which were expanded, as can be seen in the table below.

Abbrevation	Actual Word
mtg	meeting
ppl	people
abt	about
int	international
pak	pakistan
govt	government

Religious/Islamic Terms

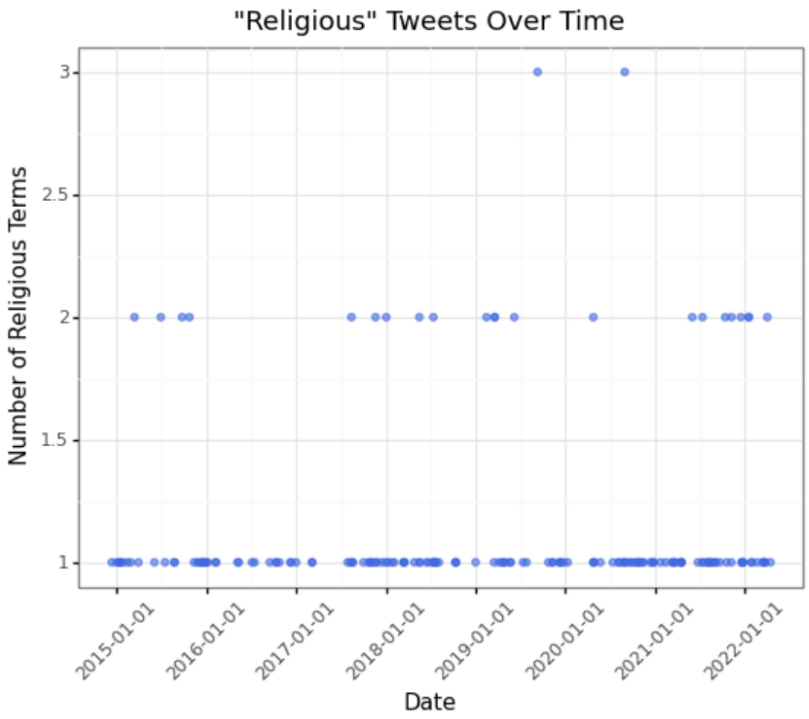
More than a 1000 of Imran Khan's tweets were manually scanned in order to identify Islamic terms. These included 'martyrs', 'karbala', 'imam hussain', 'prophet', 'baatil', 'inshaallah', 'riyasat-i-madina' and more. Since I wanted certain religious names and phrases to be captured in the data exploration, I rewrote 'hazrat ali' as 'hazrat-ali' and 'imam hussain' as 'imam-hussain'.

Data Exploration

All tweets combined contained 621,423 words. The most common word in Imran Khan's tweets was Pakistan, followed by people, government, pti (the name of his political party), and Nawaz Sharif (his major political opponent).

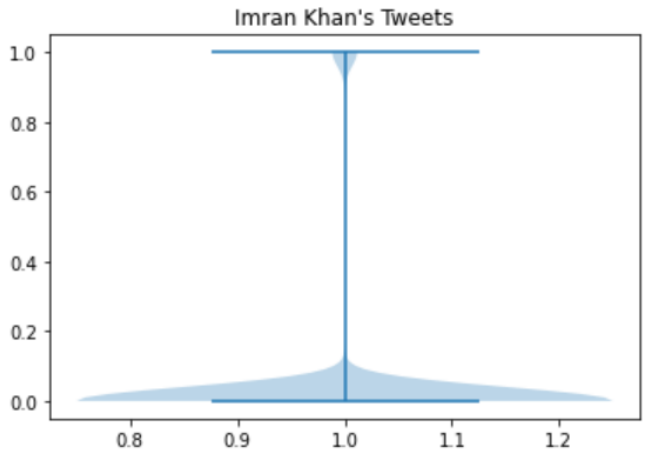
I created a few new columns to aid data exploration. These included a column storing the total number of words in each tweet, a binary variable to identify whether a tweet was religious (i.e. contained one or more of the religious terms identified), a column for the number of religious terms in each tweet, and lastly a column to store the percentage of a tweet that contained religious terms (i.e. number of religious terms in a tweet / total words in a tweet).

I started by looking at the number of religious terms in religious tweets over time.

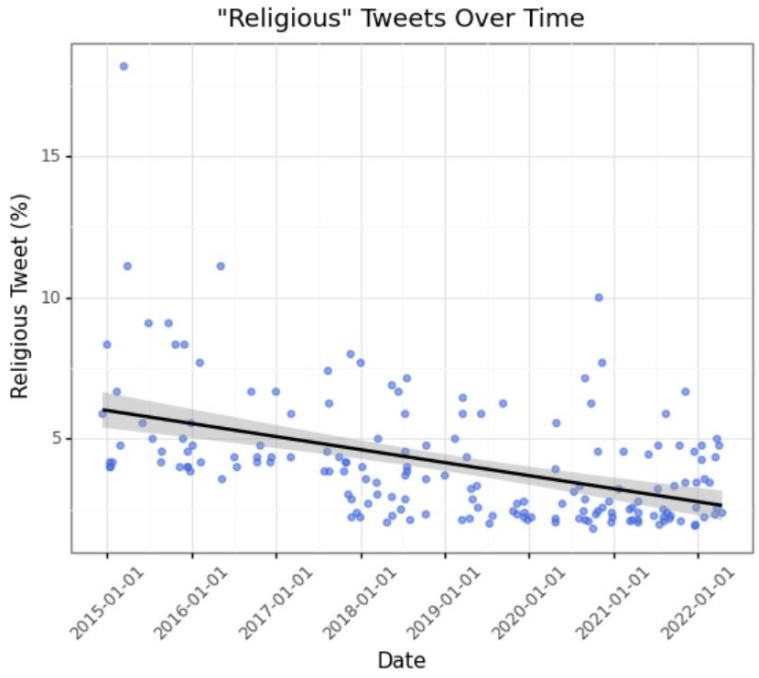


As can be seen in the visualization above, tweets both pre and post August 2018 included one religious term on average. Tweets with more than one religious term however increased post 2018. Overall, the number of religious tweets before August 2018 was less than in the post August 2018 period (this was confirmed by looking at the exact number of religious tweets pre and post August 2018).

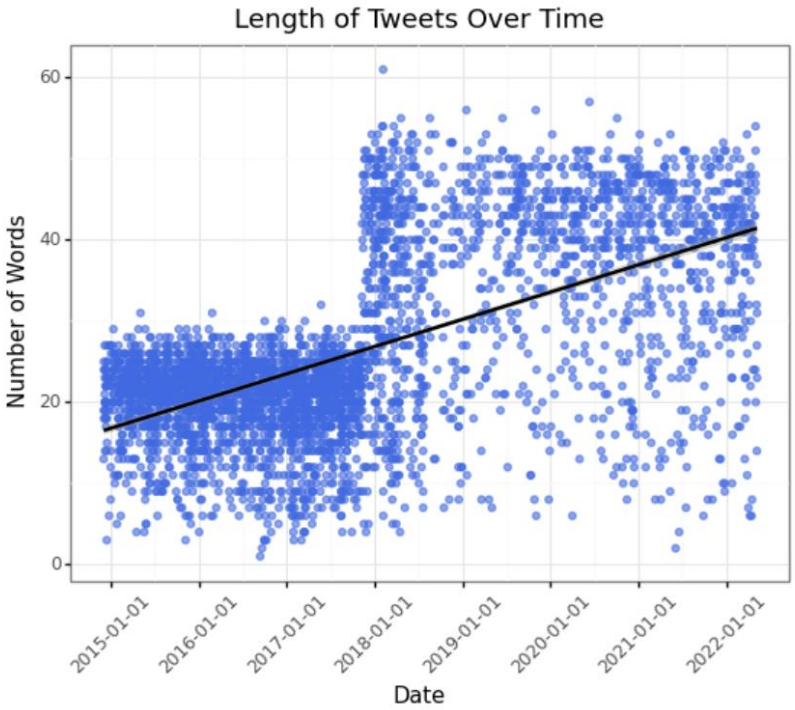
Most of Imran Khan's tweets were not religious however, as can be seen in the violin chart below (the majority of the tweets are where the variable 'rel' is equal to zero (denoting not religious).



Next, I explored how the religious proportion of a tweet changed over time.



The plot above shows that over time, the proportion of religious terms in religious tweets decreased. Since we saw in one of the previous plots that the number of religious terms actually increased post 2018, a possible reason for this downward trend is lengthier tweets over time (leading to a low proportion of religious terms). This was confirmed by plotting the length of tweets over time (as can be seen below).



The length of tweets generally increased close to 2018. The most likely reason for this is Twitter expanding its character count from 140 to 280 in November 2017.

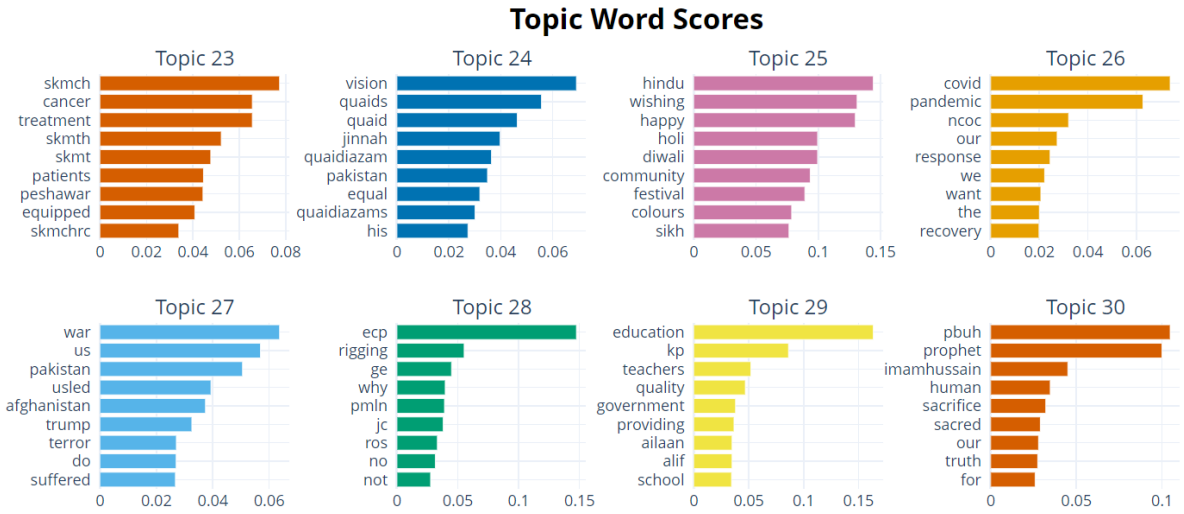
Looking at other variables over time, it was found that the number of replies, retweets and likes increased over time, both for Imran Khan's tweets in general, and for his 'religious' tweets specifically.

Guided Topic Modeling

I used Guided BERTopic and created a seed topic list containing religious/Islamic terms that were passed through the model.

Guided Topic Modeling or Seeded Topic Modeling is a collection of techniques that guides the topic modeling approach by setting a number of seed topics that the model will converge to. By defining the topics BERTopic is more likely to model the defined seeded topics. However, BERTopic is merely nudged towards creating those topics. In practice, if the seeded topics do not exist or might be divided into smaller topics, then they will not be modeled. Thus, seed topics need to be accurate in order to accurately converge towards them[5].

59 topics were created, with the highest frequency of a topic being 374. The topic of interest was 30th in the list, with a frequency of 22. Topic words scores were analyzed in order to interpret each topic (a few of these can be seen in the image below):

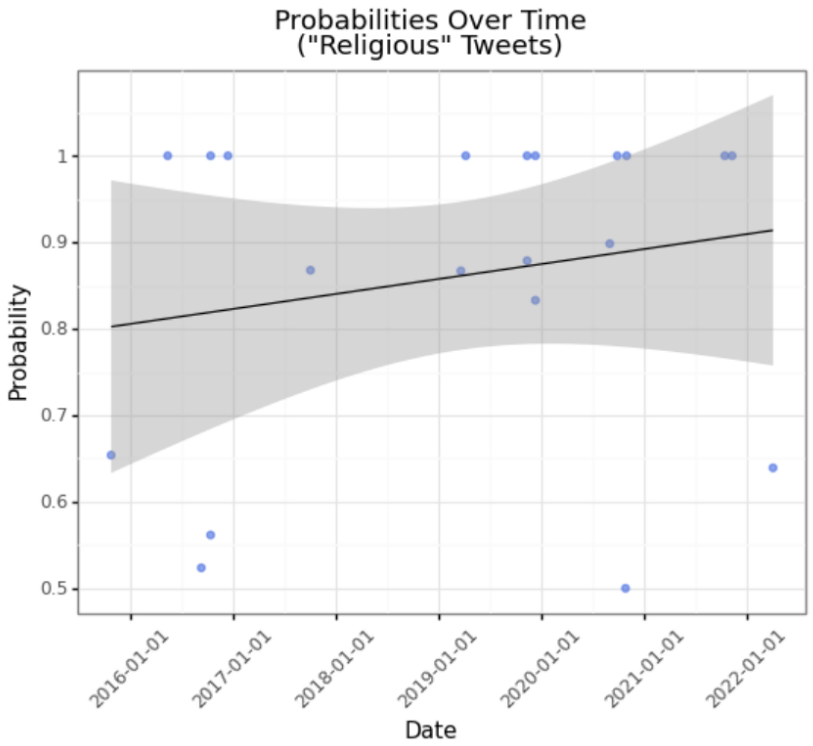


Topic 0 was about Imran Khan's main political opponent, Nawaz Sharif, Topic 1 was about India (Pakistan's main 'enemy') and the Kashmir issue (Kashmir is a piece of land that Pakistanis and Indians have been fighting over since 1947), and Topic 2 was about Imran Khan's own political party, PTI.

[5] https://maartengr.github.io/BERTopic/getting_started/guided/guided.html

A heatmap showed the similarity between topics (based on the cosine similarity matrix between topic embeddings). It showed that the topic capturing religious/Islamic terms is the most similar to the one about Islamophobia (similarity score of 0.76).

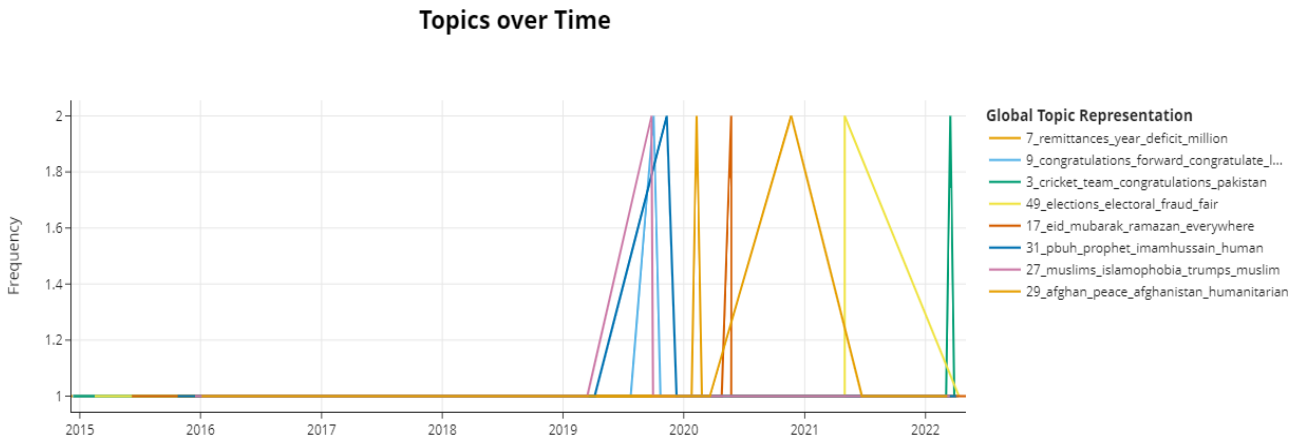
I then looked at only the tweets where the dominant topic was the one of interest, and plotted the probabilities (of a tweet having this topic) over time (chart below).



Even though the probability of a tweet having the 'religious' topic increased over time, we don't have enough data (n less than 30) to fully support this trend.

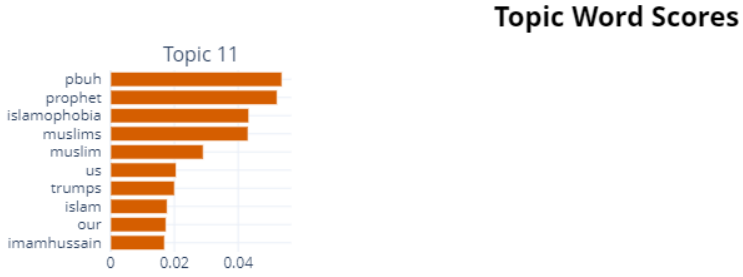
Dynamic Topic Modeling

Dynamic topic modeling showed that there was a spike in our topic of interest (Topic 31 in the image below) between mid-2019 till 2020. This coincided with the spike in topic 27 (about islamophobia). This made me curious to see that if the BERTopic model was not guided, would topics 31 and 27 essentially have converged into one topic? I explored this next.



Topic Modeling

As suspected, without guiding the topic model, it created one topic that captured all Islamic terms (as can be viewed in the image below). This does make sense, however the reason why I did not want to include the word 'muslims' and 'islam' in my seed topic list is because the word 'muslim' is a more generic word e.g. 'Eid Mubarak to all muslims' is equivalent to 'Merry Christmas to all Christians'. I therefore specified 'religious/Islamic' terms as those that mostly resonate with only the Muslims (in Pakistan) without having to specify the audience - this includes mentioning Imam Hussain or the Holy Prophet PBUH, or using Arabic Islamic words like baatil (falsehood).



Dynamic topic modeling showed the same peak that was seen earlier with the guided topic model i.e. tweets about Islam or tweets containing Islamic terminology increased in the first quarter of 2019 till just after mid-2019.

Text Classification

After the two research questions were answered through exploratory data analysis and topic modeling, I decided to explore an additional question: given the content of Imran Khan's tweets, can we predict whether his tweet would be "religious"/Islamic (based on our 'definition')? Please note that this is not related to the two original research questions, and was only explored as an additional element to the project.

I decided to use support vector machine classifiers (SVC and LinearSVC) to see which one would perform better. As seen earlier, the data was highly imbalanced (with only a few tweets categorized as 'religious'). Since SVCs are not effective at imbalanced classification, I undersampled the majority class. I was interested in trying another method and using SMOTE to oversample the minority class, but even though I have used SMOTE before, I could not figure out how to apply it to a text classification problem.

Randomly undersampling the majority class can result in losing information that may be valuable to a model. As expected, undersampling resulted in a very small dataset (n=342). The data was then split into a training and test set, and a pipeline was created including CountVectorizer(), TfidfTransformer() and the relevant classifier. The table below shows the results of the three different classifiers.

Classifier	Accuracy	Precision	F1 Score	Recall
SVM	0.85	0.87	0.86	0.85
Linear SVM	0.90	0.93	0.90	0.87

Linear SVM performed better than SVM on all performance metrics.

Conclusion

The first research question looked at whether the number of Imran Khan's tweets that contain Islamic elements have increased over time. In our exploratory data analysis, we saw that there were more tweets containing Islamic elements post August 2018 as compared to pre August 2018. Moreover, through dynamic topic modeling, we were able to see a spike in our topic of interest between mid-2019 till 2020.

The second research question looked at whether Imran Khan's Islamic/religious tweets have become *more* religious over time? Exploratory data analysis showed that the number of Islamic terms in tweets on average, was the same in both pre and post August 2018. However, more than one Islamic term in a tweet was more common post August 2018. Topic modeling also allowed us to look at tweets where the dominant topic was the one of interest, and to view the probabilities (of a tweet having this topic) over time. Even though the probability of a tweet having the 'religious' topic increased over time, it should be noted that the size of the dataset was quite small (less than 30).

Was Success Achieved?

Before starting this project, success was defined as follows:

- Creating a corpus of Imran Khan's tweets across 8 years with their corresponding dates
- Cleaning the corpus and getting it ready for analysis
- Using guided topic modeling correctly
- Being able to answer the 2 main questions I have using my topic modeling results.

Based on this criteria, success was achieved, although at times I had to lean on descriptive analysis of the text in order to answer the questions I had. Therefore, there is probably room for improvement in my topic modeling approaches and analysis.

What I Would Do Differently

If I had more time, I would like to explore GuidedLDA as well, and compare it with Guided BERTopic in terms of ease of implementation and analysis. While for this project I wanted to specifically look at Imran Khan's tweets since he became Prime Minister and his tweets before he took oath, I would also like to explore his tweets in general i.e. looking at all of his tweets, not just his tweets within the 8 years I identified.