# Data Collection:

We used the **Scrapy** library in Python to create a spider which would crawl through the **Dawn Newspaper** website’s archives and access the links to the articles. The archive pages were used to extract the categories, excerpts, and dates of the articles, whereas the actual article’s webpage was used to extract it’s headings and text. This required 2 spiders:

* news.py : for scraping the article’s webpage (Data Collection/scrapy project/news\_crawler/news\_crawler/spiders/news.py)
* cats\_spider.py

(Data Collection/cats\_spider.py)

**cats\_spider.py** was run locally, whereas news.py was uploaded online to Scrapy’s online Scrapinghub platform, which run the spiders on their own servers.

A time delay of 3 seconds was given between Get requests, which prevented the spider from being blocked, which could have resulted in the loss of news articles.

**merge.py** was then used to merge results from both spiders by based on the link of the articles. The links were sometimes redirected, and thus for those links in the results from **cats\_spider.py** that had no match, the links were redirected to their using Python’s request library, and then matched.

Finally, the file: **ALL\_complete.csv** was created which included all articles on the Dawn Newspaper’s website from 1st July, 2017 to 31st May, 2018. This was the dataset used to run and explore LDA and it’s possible uses for online Pakistani newspaper analysis.

Topic Modelling:

\*The following codes can be found in Yearly LDA/code

The following codes are run in the given order:

**create\_docs.py**: sorts out the data from ALL\_complete.csv according to the date, as well as remove any irrelevant categories, given the kind of analysis you wish to do, hence creating data.csv which well then be used by the following code.

**LDA\_1.py**: uses the nltk and genism libraries to process the data by removing stopwords and punctuation, as well as stemming the words.

**LDA\_2.py**: runs the actual model on the corpus with varying number of topics (all even numbers between 2 and 40)

**lda\_choose.py**: finds the coherence values of the each LDA model and thus allows us to choose the better models to use (the higher the coherence value, the better)

**doc\_to\_topic\_dist.py**: creates document to topic distribution, which has the perecentage representations of all topics for each document

**lda\_extract\_topic\_rep\_csvs.py**: creates list of document headings along with the representation of a topic for that topic, for all topics, with a separate csv file for each topic

# **Results**

### (LDA was run multiple times on the dataset, but only the more prominent results are shown)

#### **Prominent Results No. 1 (PR-1)**

* The dataset includes all categories and the entire range of available articles.
* The number of topics was set to 20.
* Top words from each topic the more interesting, relevant and interpretable topics:
* Headings of the articles in descending order of the topic representation were viewed, and a consistent manual threshold was set for all topic models.

*Note that while some topics remain particularly consistent, those with spikes do not necessarily have a very concrete reason behind it. This could be due to chance but also, when looking at a difference of 10-15 articles, there isn’t necessarily a particular reason behind that.*

##### Topic 0 (Cricket articles)

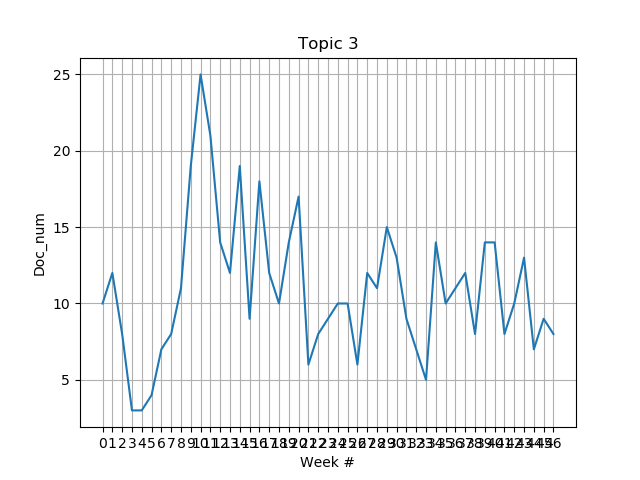
*Keywords:*

test, England, match, cricket, first, run, play, ball, say, team

##### Topic 3 (Rohingya Crisis related articles)

##### *Keywords:*

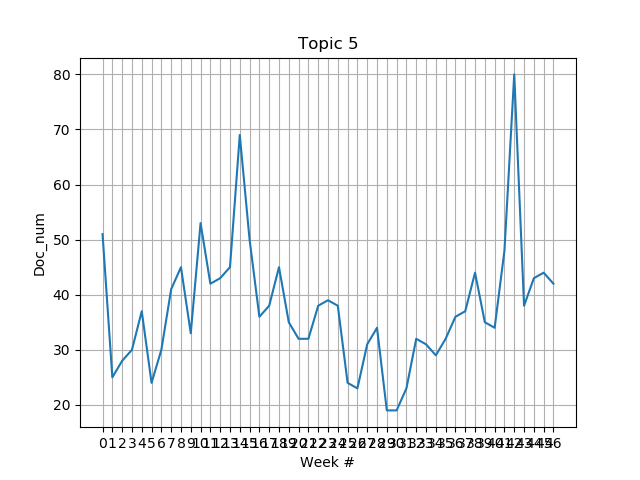
say, myanmar, rohingya, Bangladesh, international, britain, year, country, government, refugee



##### Topic 5 (US foreign relations)

##### *Keywords:*

say, trump, iran, president, nuclear, russia, united\_state, china, state, russian

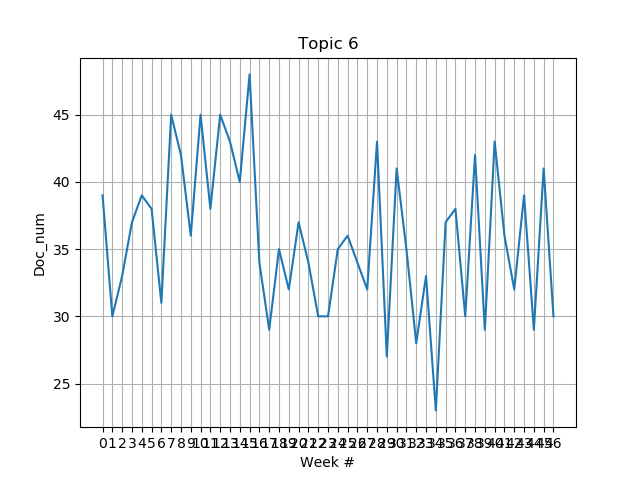


* *Spike # 1: Trump announces sanctions on North Korea, Florida shooting, US-Russia tussle over Ghouta*

##### Topic 6 (Panama Papers Case related articles)

##### *Keywords:*

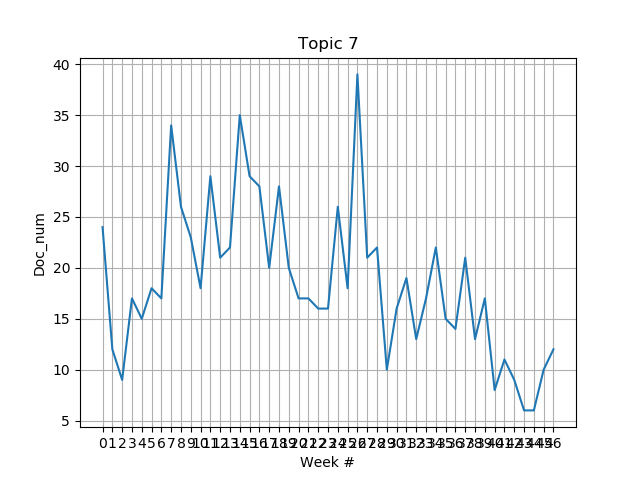
court, justice, case, say, nab, chief, judge, order, also, former



##### Topic 7 (Pakistan Foreign Relations related articles)

##### *Keywords:*

pakistan, say, country, political, party, would, also, government, state, military

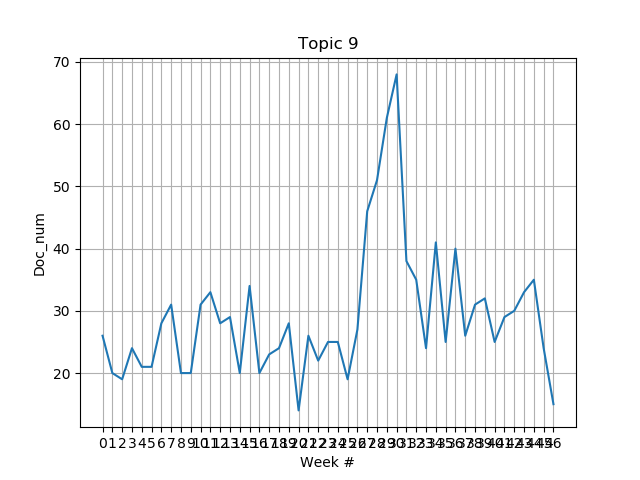


* *Spike # 1: US suspends aid, puts Pakistan on special-watchlist and Trump tweets about Pakistan*
* *Spike # 2: Pakistan placed on FATF grey-list, steps up foreign relations efforts*
* *Spike # 3: Concerns over concerns over construction of kashianga dam in india, articles and public statements on cpec due to upcoming cpec summit 2018*

##### Topic 9 (killings, kidnappings, assaults)

##### *Keywords:*

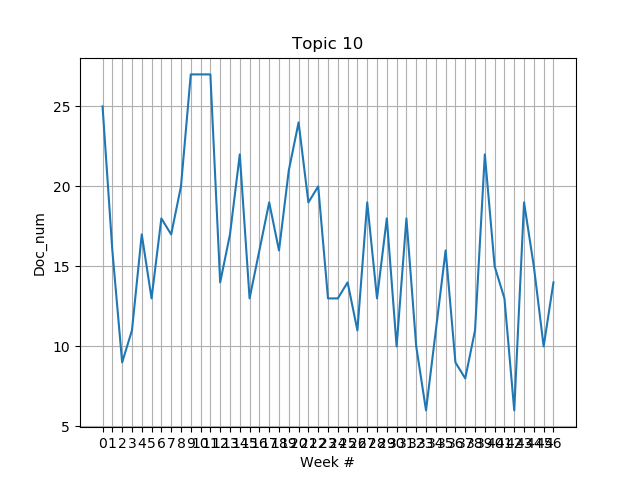
police, say, case, report, arrest, court, suspect, also, officer, investigation



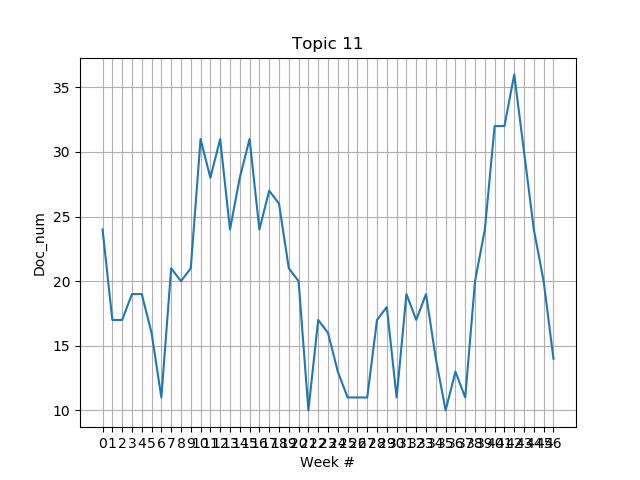
* *Spike # 1: Here, it can be seen that there is a particular rise around January. This could imply that it was due to the Zainab murder case, but this required further research.*

##### Topic 10 (Parliament related news)

##### *Keywords:*

say, party, election, government, minister, would, assembly, chief, pml, pti

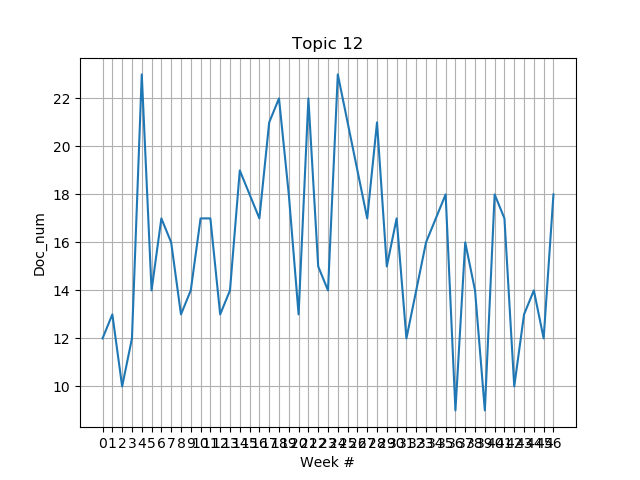
##### Topic 11 (Kashmir-related news)

*Keywords:* india, indian, kashmir, state, journalist, say, medium, people, protest, delhi

##### Topic 12 (Syria, Afghanistan, Iran, Saudi Arabia)

##### *Keywords:*

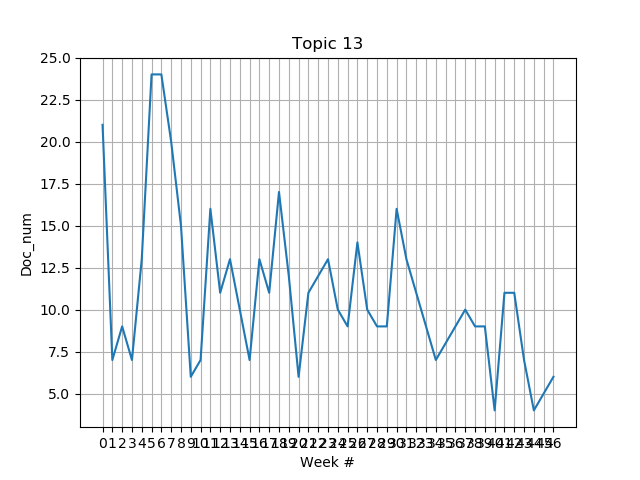
Say, muslim, kill, year, government, people, attack, force, group, state



##### Topic 13 (NAB, corruption charges)

##### *Keywords:*

say, official, land, project, also, airport, railway, security, investigation, authority



#### **Prominent Results No. 2 (PR-2)**

#### LDA was also run on separate months

**Note that here, many of the major events in the separate months were revealed, though if we are to properly automate this process, we need to find a way to differentiate between a relevant topic vs an improper classification of certain articles which do not necessarily have an underlying event, let alone a theme.**

#### October 2017:

* The dataset includes all categories and all articles in the month October 2017.
* The number of topics was set to 18.
* Topic 3: NAB, ECP and PTI related news
* Topic 4: Elections Act 2017, Khatm-e-Nabwooat issue and events relevant to Nawaz Sharif’s disqualification
* Topic 5: Middle-eastern crisis, including Gaza, Libya and Hammas (increasing tenstions between Hammas and Israel over the Gaza strip as well as increasing tensions between Libya and Saudi Arabia)
* Topic 6: News related to Tillerson’s foreign trips to Central Asia, CPEC and US-Pak-India relations
* Topic 8: Cricket-related news
* Topic 9: Economy related news
* Topic 12: Mass-death related incidents .e.g. wildfires and terrorism
* Topic 15: Qandeel murder case as well, including Mufti Qavi’s trials, as well as other reports on sexual harassment or child abuse
* Topic 17: Catalan and Kurd independence

#### November 2017:

* The dataset includes all categories and all articles in the month October 2017.
* The number of topics was set to 20.
* Topic 3: Includes case against Dar, but no consistency
* Topic 2: cricket-related news
* Topic 5: Faizabad sit-in
* Topic 6: Middle eastern crisis, as well as some news on Iran and Afghanistan
* Topic 8: Killings and murders, including honor killings and child abuse
* Topic 9: Investment projects
* Topic 11: Pak-Afghan-US relations (increased pressure on Pakistan to deal with terrorist factions, as well as tensions between Afghanistan and Pakistan over border patrol)

**Comments:**

##### *Note that for October 2017, LDA was able to classify articles on the following major events of the month:*

* Elections Act 2017 and events relevant to Nawaz Sharif’s disqualification
* Increased tensions between Hammas and Israel over the Gaza Strip
* Tillerson’s visit to Central Asia
* Mufti Qavi’s trial for his alleged involvement in Qandeel Baloch’s murder
* Catalan independence movement

##### *And for November 2017:*

* The Faizabad sit-in
* Increased pressure from US on Pakistan to deal with terrorist factions

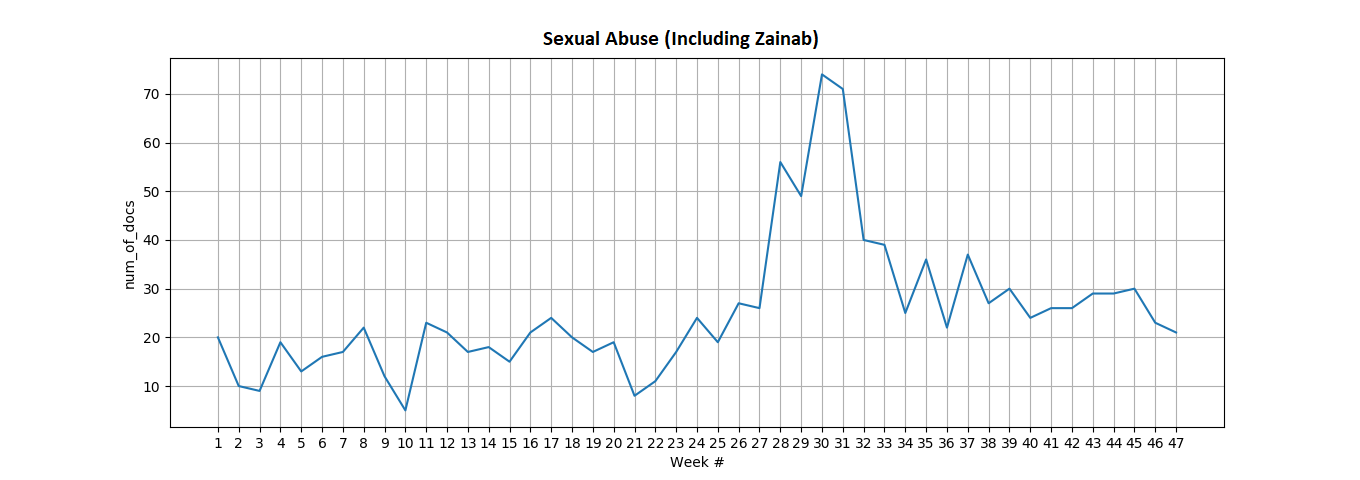
#### **Prominent Results No. 2 (PR-2)**

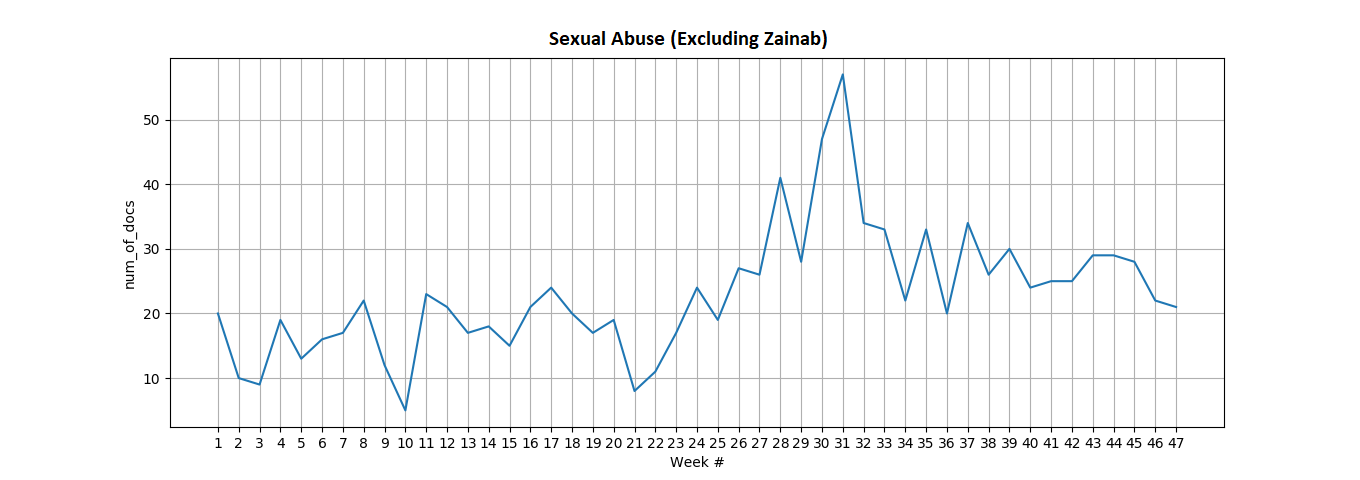
#### Guided LDA

\*The entire code for this can be found in Guided LDA/guidedlda.py

**As previously mentioned, Topic 9 provided some interesting results which required further investigation. The guidedlda library was used to allow for running LDA while also specifying particular keywords for certain topics.**

* The dataset includes only the categories Pakistan, Newspaper and Blogs on all available dates.
* The number of topics used was 20.
* Keywords were only specified for Topic 0, and they were as follows:
* ['rape', 'raped', 'rapes', 'gangrape', 'gangraped', 'harassment', 'harasser', 'harasses', 'harassing', 'harassed', 'harass', 'harassers', 'harrassed', 'harrassment', 'sexual']
* The textmining library was used to create the appropriate document-term matrix for the guidedlda library, which did not have a very efficient stemming code, hence the redundancy in the keywords. Though, it should be noted that the vocabulary was searched for all words containing the series of letters: ‘rape’, ‘harras’, ‘haras’ and ‘sexual’. Then words to be found as relevant keywords were added to the list. Hence, essentially there were only 4 keywords: ‘rape’, ‘gangrape’, ‘harrass’, and ‘sexual’.
* Guided LDA was run twice, once with all articles, and the other without any articles containing the word ‘Zainab’ in the heading, so as to eliminate any possible bias for a certain spike in articles due to the Zainab case. In both cases though, the spike remained, showing that there was a clear rise in sexual abuse cases during that time.



**Comments:**

This was expected. The Zainab case was a huge story and really affected the overall views and reactions of the Pakistanis toward sexual harassment.

One thing though, was quite interesting. While doing a brief overview of the headings of these articles around the time of the spike, we found that famous murder cases like the Mashal Khan, Shahzeb Khan, Perween Rahman and NaqeebUllah Mehsud murder cases were also being mentioned along side within the same topic.

So we mapped out the mentions of these cases within articles, and the results are as follows:

|  |  |
| --- | --- |
| Month # | Month name |
| 1 | July 2017 |
| 2 | August 2017 |
| 3 | September 2017 |
| 4 | October 2017 |
| 5 | November 2017 |
| 6 | December 2017 |
| 7 | January 2018 |
| 8 | February 2018 |
| 9 | March 2018 |
| 10 | April 2018 |
| 11 | May 2018 |

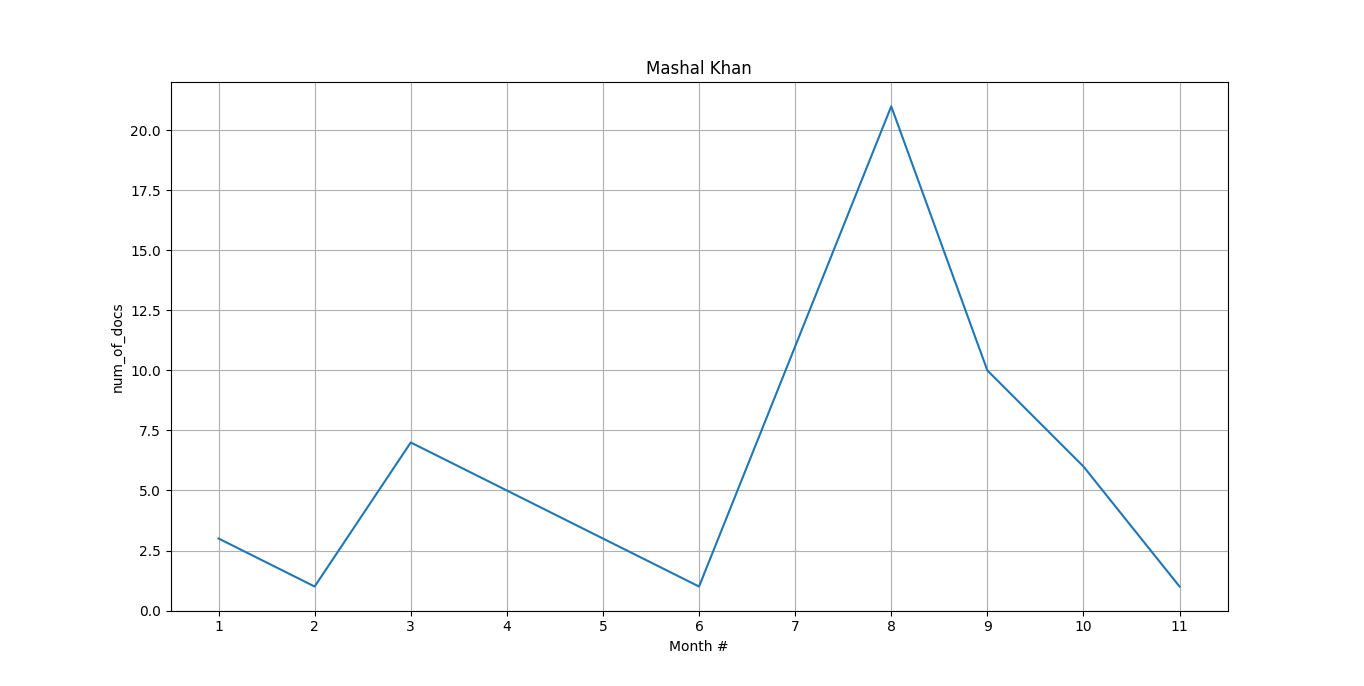


Figure 1: Mashal Khan Murder case articles (Keywords: ‘Mashal Khan’ or ‘Mashal’ and ‘blasphemy’)

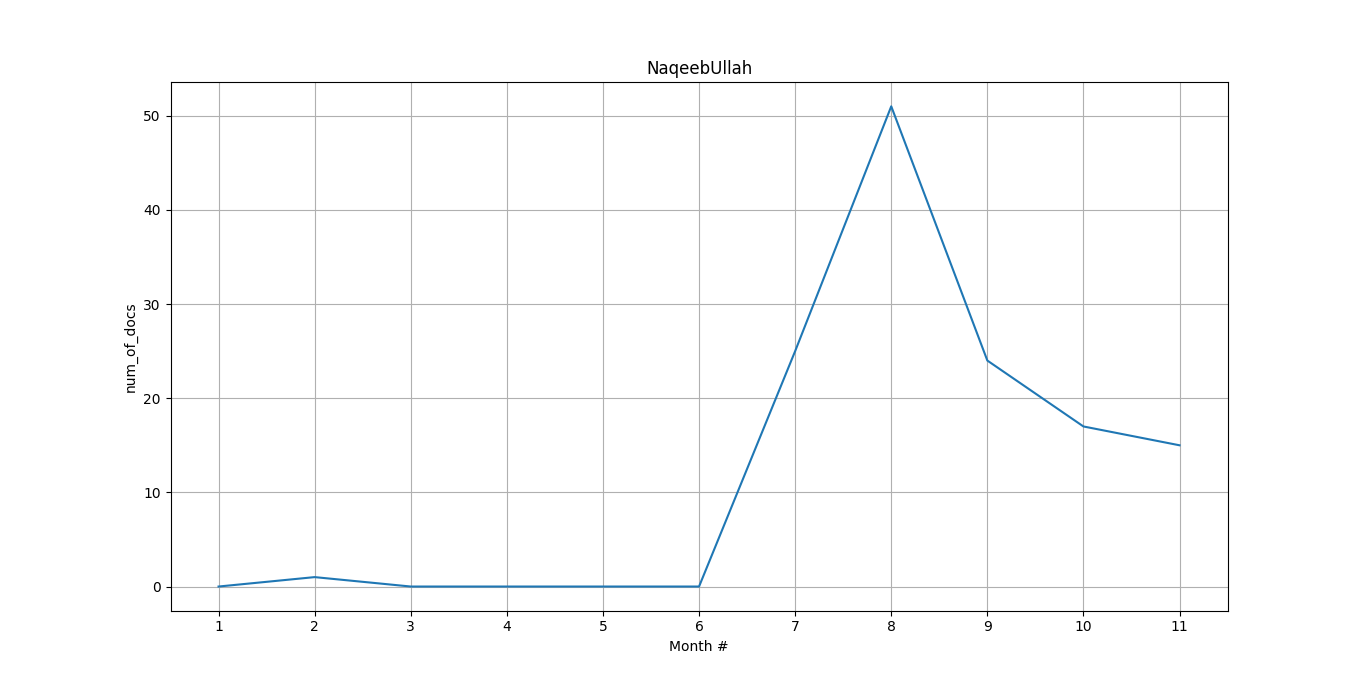


Figure 2: NaqeebUllah Masood murder case (Keywords: 'NaqeebUllah' or 'Naqeebullah' or 'Naqeeb Ullah')

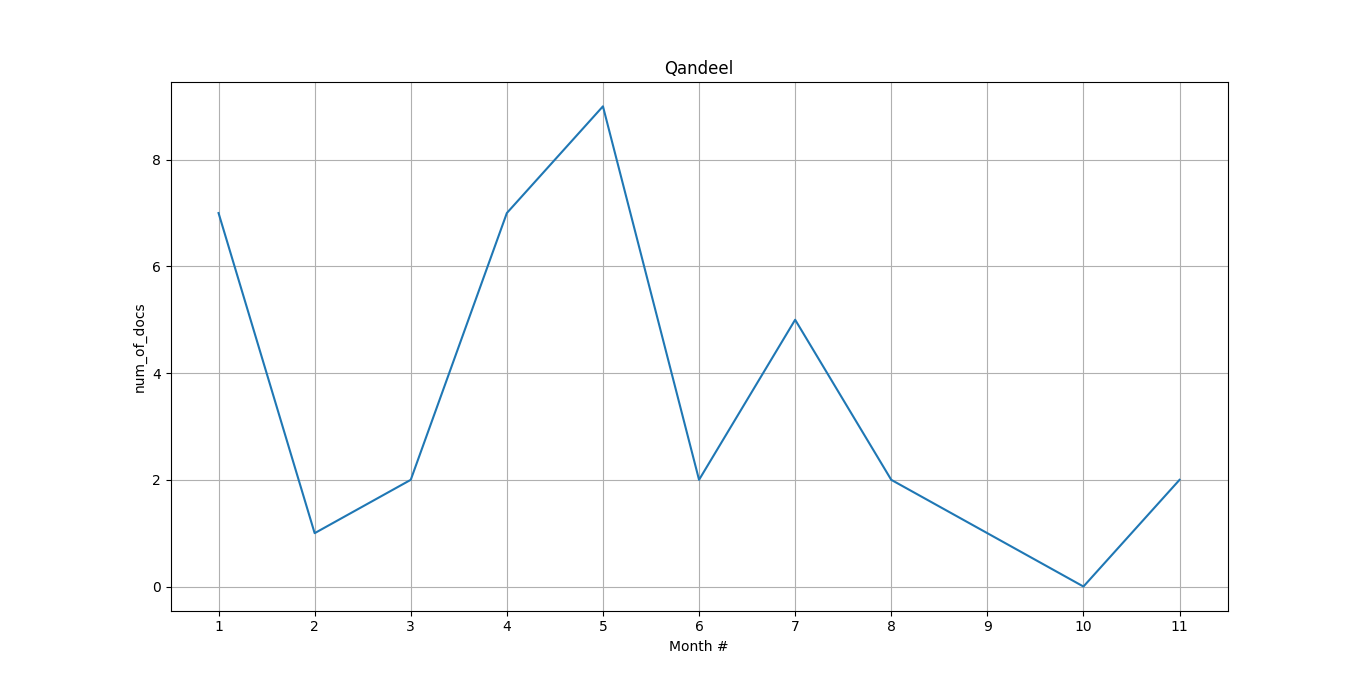
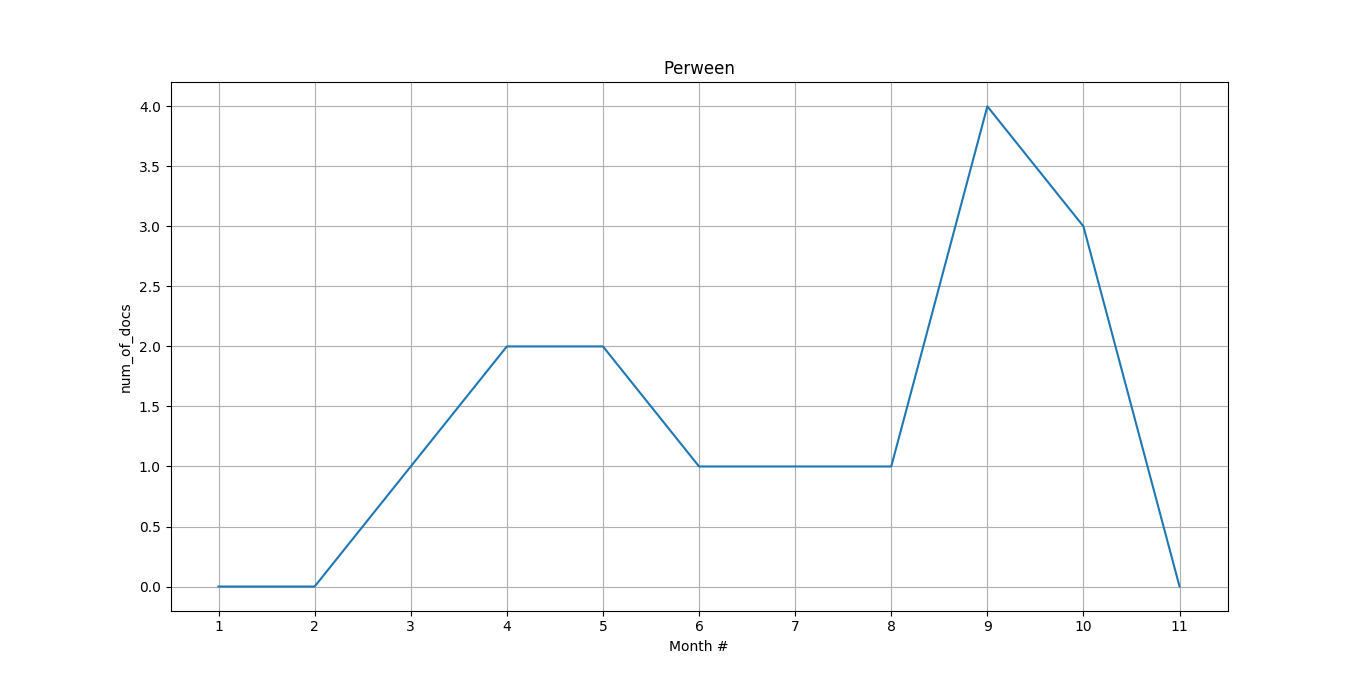


Figure 3: Perween Rahman murder case (Keywords: 'Perween Rahman' or 'Perween Rehman')

Figure 4: Qandeel Baloch murder case (Keywords: ‘Qandeel Baloch’)

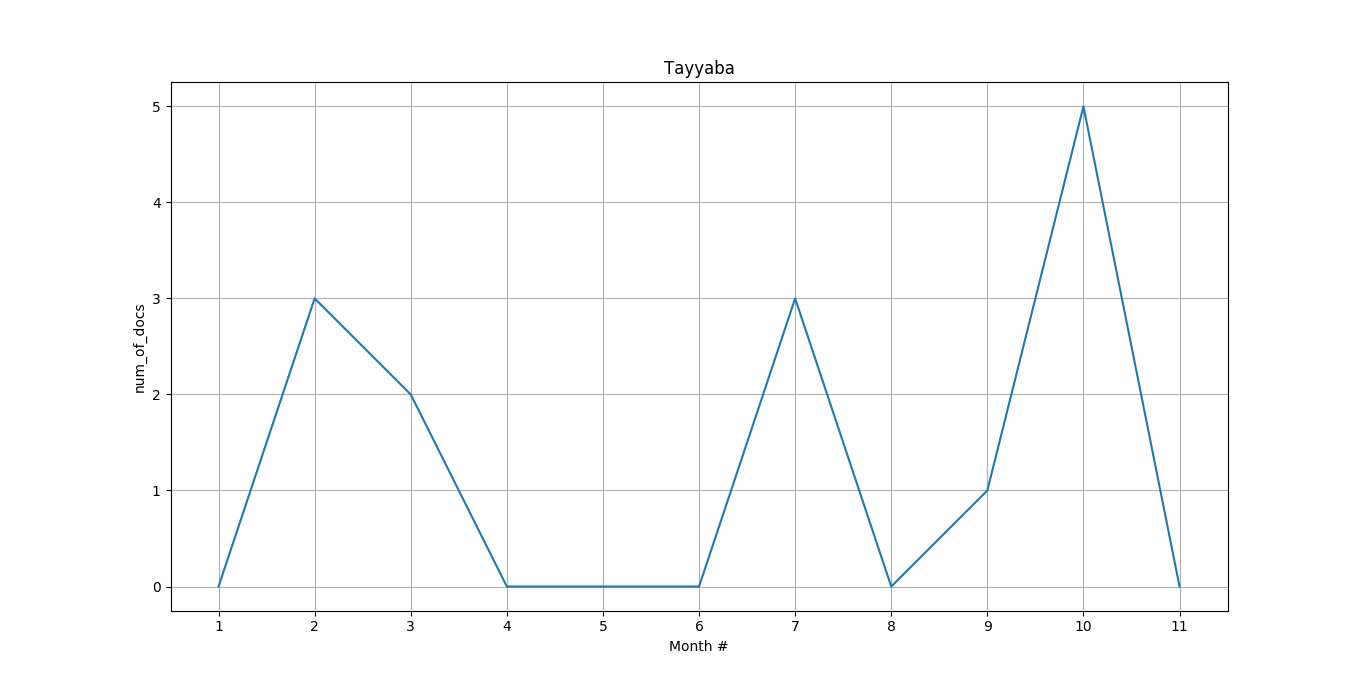


Figure 5: Tayyaba torture case (Keywords: 'Tayyaba')

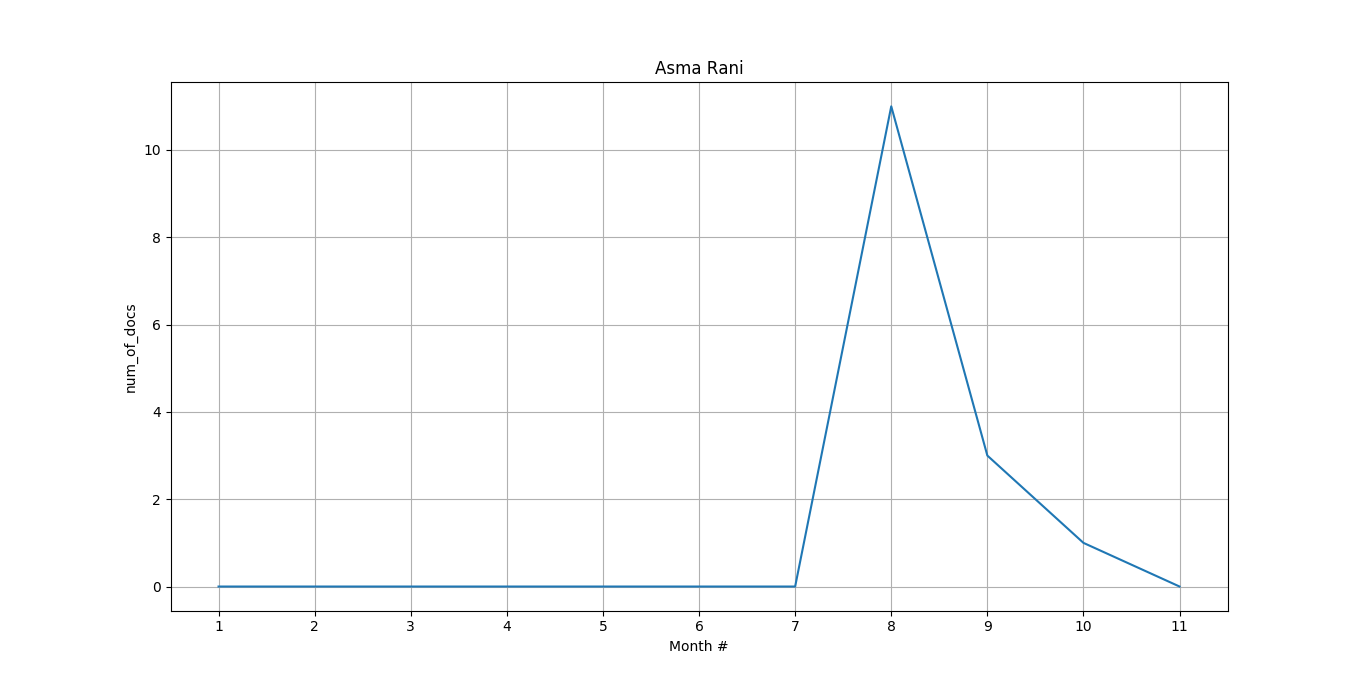


Figure 6: Asma Rani murder case (Keywords: 'Asma Rani')

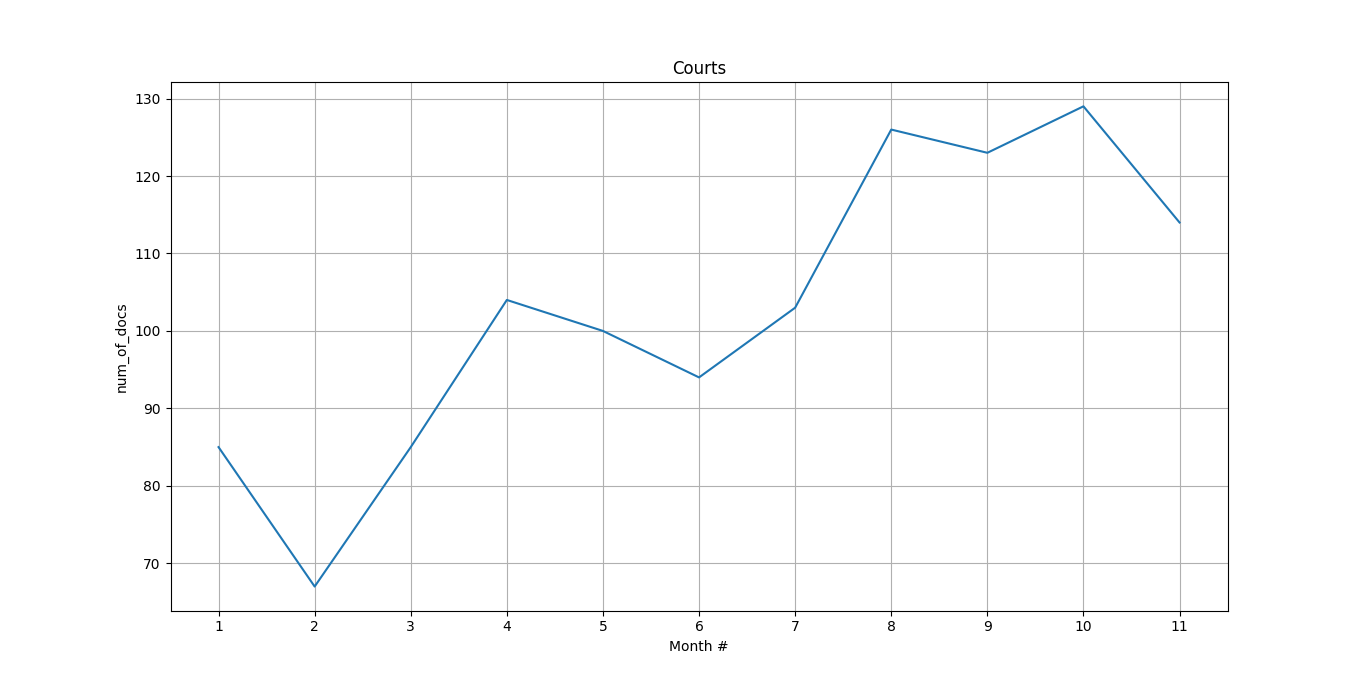
**Comments:**

There are a few things to note here. Firstly, the a number of cases, particularly the Mashal Khan case, Perween Rahman case, and also the Tayyaba and Qandeel Baloch cases began to gain significantly extra attention around the time of the Zainab case despite the fact that they had been going on for quite some time before as well. As a hypothesis, it can be said that there could be a certain bias towards these cases, either from the media or the justice system.

For example, it is quite clear that corruption cases against politicians which were being delayed for years, suddenly started to be looked into after the Panama Papers case. The data’s proof of this can be seen in the following two graphs:

Figure 7: Mentions of the courts including articles related to the Zainab case (Keywords: ‘SHC’ or ‘LHC’ or ‘SC’ or ‘court’ or ‘Court’ in text or heading)

Figure 7: Mentions of the courts including articles related to the Zainab case (Keywords: ‘SHC’ or ‘LHC’ or ‘SC’ or ‘court’ or ‘Court’ in text or heading)



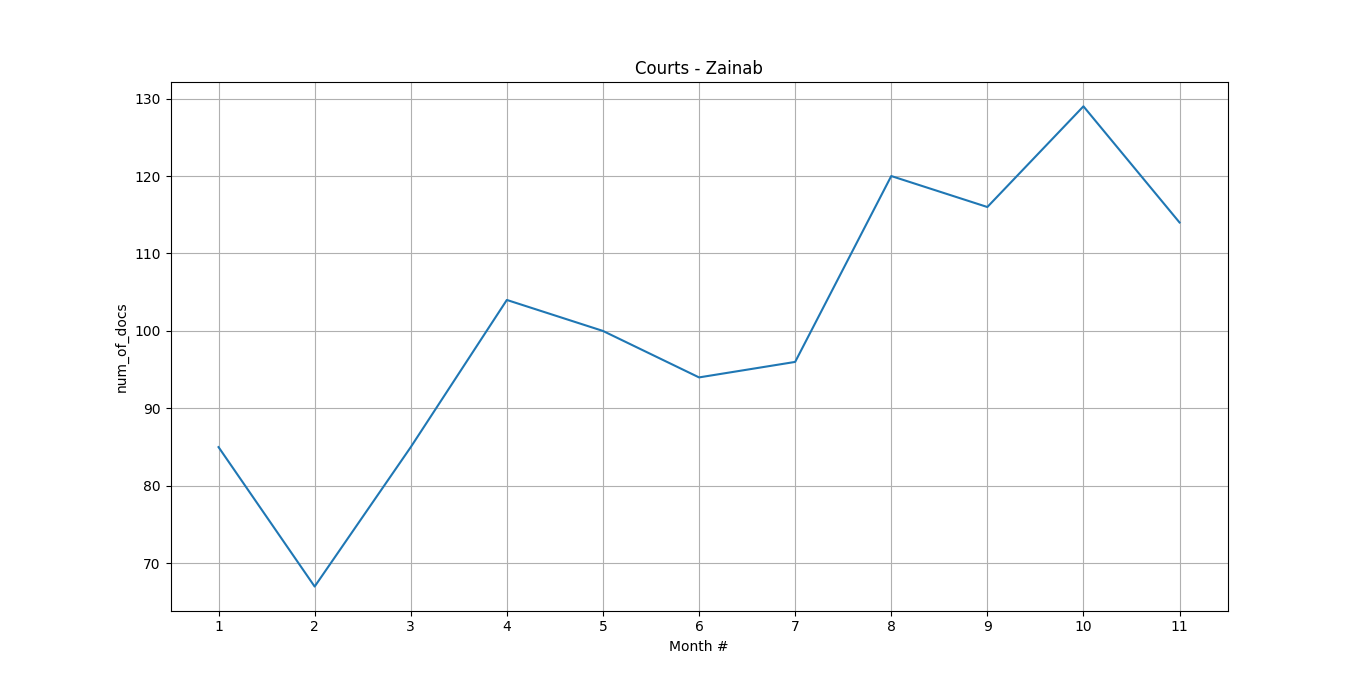


Figure 8: Mentions of the courts including articles related to the Zainab case (Keywords: ‘SHC’ or ‘LHC’ or ‘SC’ or ‘court’ or ‘Court’ in text or heading and NO Zainab in heading)

**This clearly shows a rise in work by the courts. But then we look at the same mentions for police as well:**

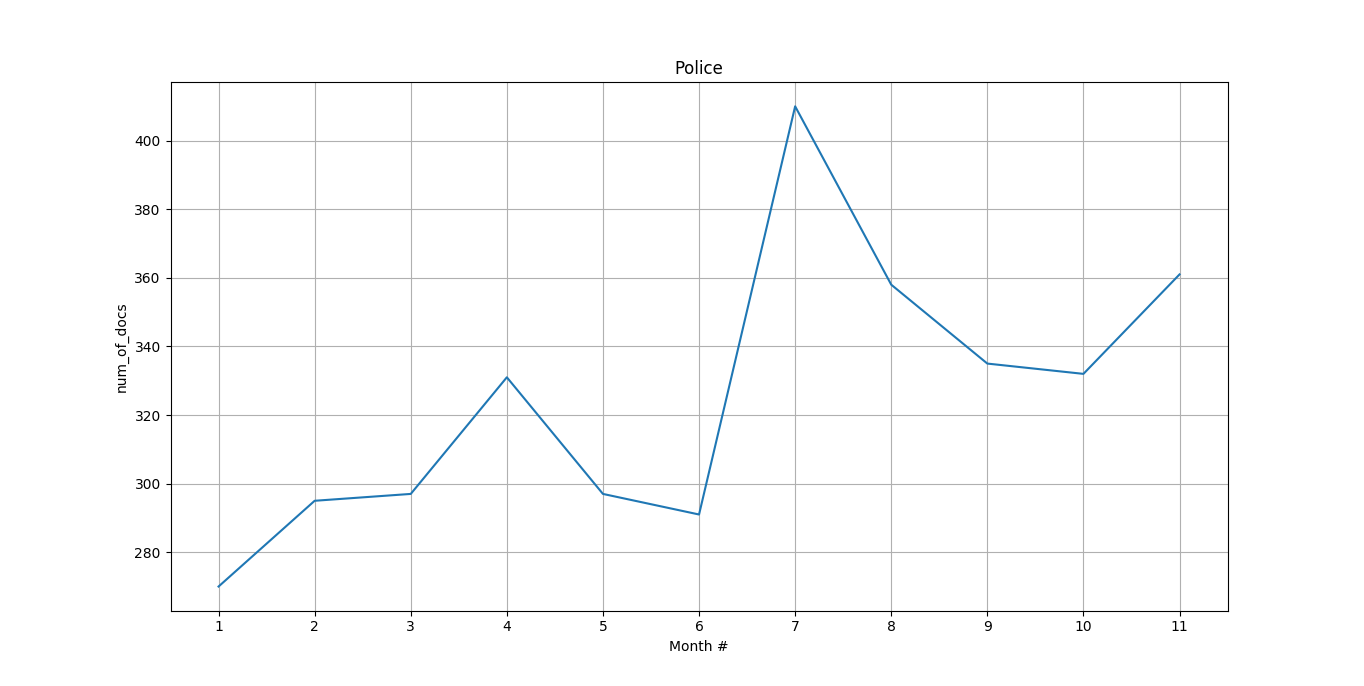


Figure 9: Mentions of the police including articles related to the Zainab case (Keywords: 'Police’ or ‘police’ in text or heading)

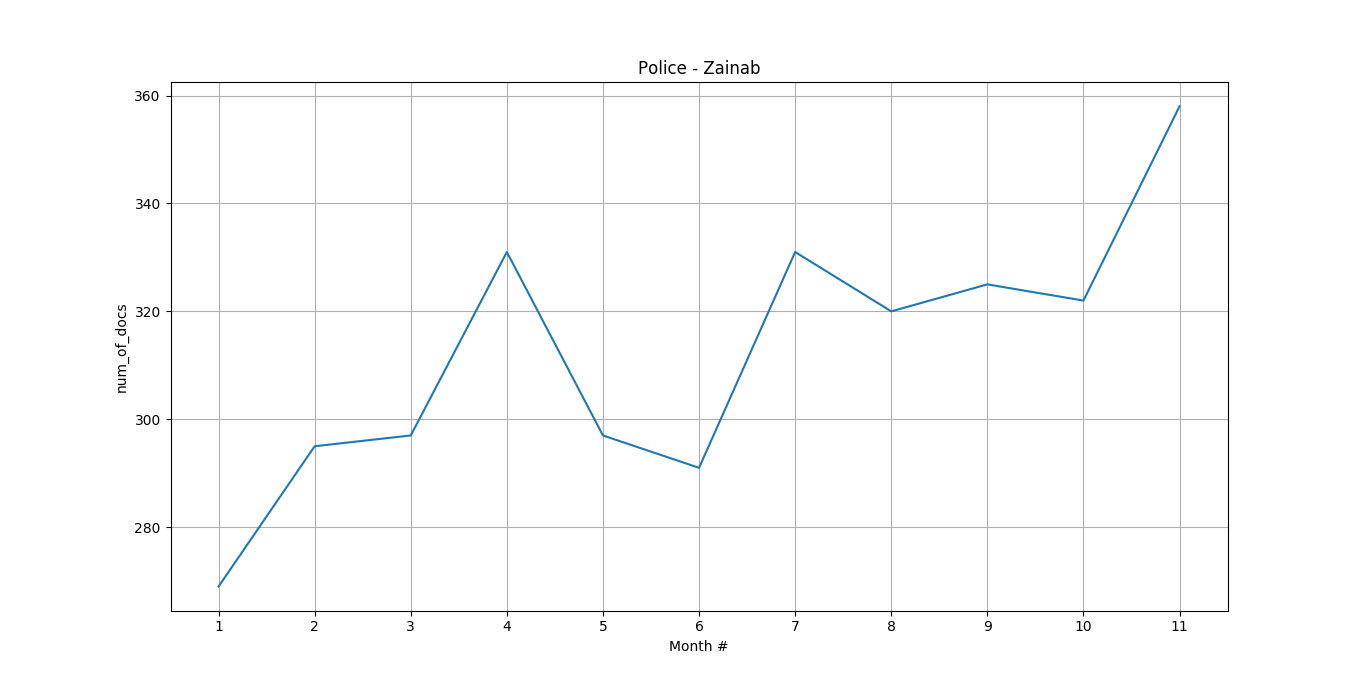


Figure 10: Mentions of the police including articles related to the Zainab case (Keywords: 'Police’ or ‘police’ in text or heading and NO 'Zainab' in heading)

**Comments:**

From the data, we can hypothesize that the police are somewhat influenced by media attention. This in turn, could lead to the courts being more active, because it is only when people are more vocal about reporting crimes to the police, followed by investigations by the police that are hyped up by the media, that cases can be properly looked into by the court.

Also, note how round about the same time as the Zainab case, two other major cases, including the Naqeeb Ullah case and Asma Rani cases regained momentum as well. The cases possibly gained more attention than their actual time of occurrence, due to the reasons mentioned above, and can also help explain the sudden hype among the Pakistani population