ANALYZING THE IMPACT OF POLITICAL RHETORIC IN TRADITIONAL AND SOCIAL MEDIA

Rohit Aggarwal - rohitagg
Deepakanuraag Sathyanarayanan - deepakan
Adhithyakrishna Kovai Srinivasan - akovaisr
Department of Computer Science
University at Buffalo
Buffalo, NY 14214

APPLICATION URL : http://tweetengine.herokuapp.com/inforetrivers

OVERVIEW:

CROSS LINGUAL INFORMATION RETRIEVAL SYSTEM

The data was crawled from Twitter using the twitter search API. The crawled data was then processed using python script to extract the required information such as text, hashtags, user names, language etc. The geographical information such as locality, state and country were also added to the processed tweet JSON files using Google Maps Geo-encoding API. The processed tweet collection is then indexed in Solr. We developed a UI which we named as **Tweet Engine** such that Solr acts as the backend of it. We also integrated the analysis of search results to make it more interesting for the user. We also added the advanced search feature to allow the user to filter the tweet results based on the language and country. A query parser was also developed as part of this project to ensure that our IR system caters to the information need of the user.

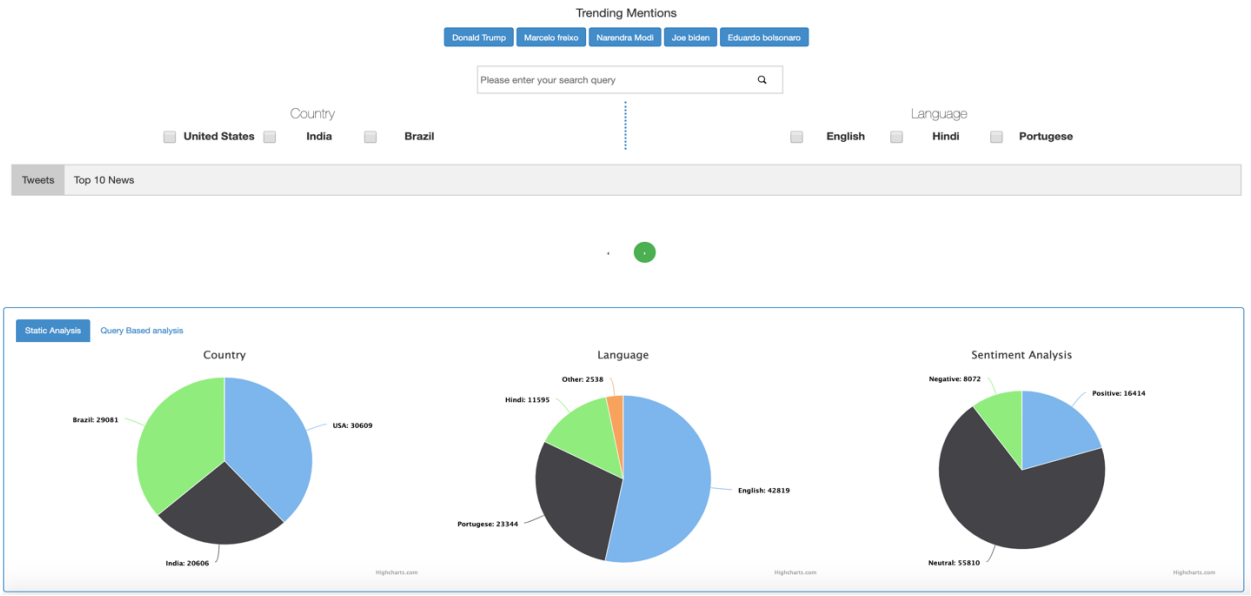


Fig – Home Page of IR System

FEATURES:

- - We crawled tweets in three different languages- English, Portuguese, Hindi.
- - UI displays 10 results in each page.
- - We implemented analysis of returned tweets corresponding to languages, locations, sentiment analysis, tweet timeline.
- - The results can be filtered based on the language and country according to user requirements.
- - User can query in a language and expect to obtain results in other languages.

IMPLEMENTATION DETAILS:

Preprocessing Data:

We implemented the preprocessing of data similar to the implementation we used for project 1 by using a python script. We gathered information for the fields such as user information fields- user-name, screen-name, profile-image, location and necessary fields such as date, text, emoticons, hashtags etc. The data is processed in order to index the data effectively corresponding to the required fields. The preprocessing of data also included the addition of location fields – country based on the geographical co-ordinate information present in the raw tweet files.

Google Maps Geocoding API:

We used the Geocoder python package to extract geographical information from the location attribute in the raw JSON file. The Google Maps geoencoding API provides a direct method to access these services using HTTP requests which is included in the python script used for preprocessing of raw tweet files. The state and country of the user are encoded into the processed JSON file. This information of users can be used to analyze the correct cities of the users who tweet the most about the search query.

Google Translate API:

We used the Google Translator API which is a cloud-based machine translation service. We used the API to detect the language of the given query and translate it to other languages such that the returned results are relevant to the user

Tweet Engine custom query parser:

We integrated this concept of implementation of detection and translation of queries by building our own query parser which is extended from the extended dismax query parser which is known as edismax query parser, an improved version of dismax query parser. The standard query parser, which is the default query parser, is intolerant of syntax errors as it expects the query to be well-formed. On the other hand, DisMax query parser is known to be a more forgiving parser as it is useful for directly passing in a user- supplied query string. The edismax query parser is an improved version of the dismax query parser with

additional features. We boosted the tweets which are in the same language as the language of the query so that this language is favored over others. In addition to this, if the query contains hashtag, the tweets with hashtags are boosted.

Faceted Search:

Faceting is the arrangement of search results into categories based on indexed terms. This technique of using faceted fields can be exploited in order to perform the analysis of the returned search results in terms of language count, location of the user. Various charts have been created for the visual representation of the analyzed data.

The concept of faceted searching is also used to narrow down the search results by setting parameters such as language, country etc. in order to get much more relevant results for the user.
The top trending mentions corresponding to the query can also be determined by the concept of faceted search.

Tweet Engine UI:

We developed a UI for our application by using HTML, Node js and JavaScript. We used JavaScript Highcharts api for making the pie chart and time line graph for the visualization of analytical data displayed on the webpage. The results page displays 10 results at a time in various pages similar to the design observed in Google Search Engine. The advanced search option allows the language based and country-based filtering of the tweets. In individual tweets, various information has been displayed such as user name, country, sentiment of the tweet and text of the tweet using the data from processed tweet files.

News API:

We used the News API which is a cloud-based machine translation service. We used the API to detect the given query and relevant news articles to the tweets made by different personalities published in news websites such as New York Times and CNN are fetched and displayed.

Sentiment Analysis:

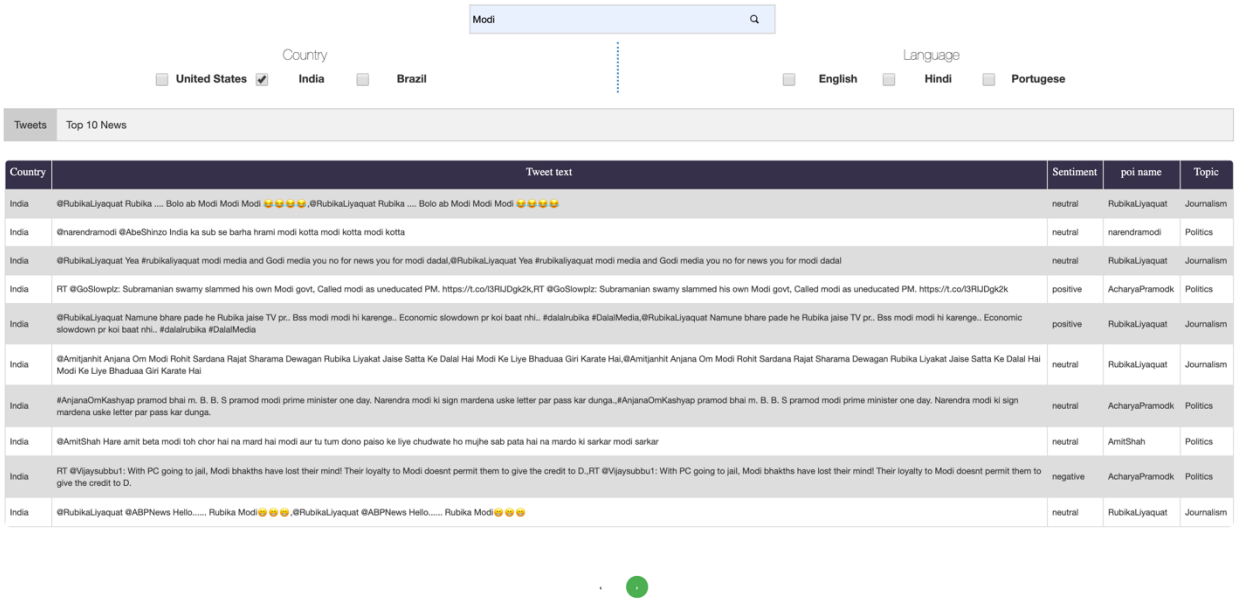
We implemented the simplest form of Sentiment analysis on the tweet results set using a word list called AFINN-111.The file AFINN-111 is the latest version which consists of 2477 words and phrases. The approach we used compares the terms in tweet text with the list of words in AFINN-111 file which contains pairs of word and precomputed sentiment score associated with the word (ranging from -5 to +5). The sentiment of the tweet is determined by adding up the sentiment components of the individual words in the tweet. For the results returned for the user-given query, the number of positive, neutral and negative tweets present in the search results is displayed.

AFINN-111 file was available only in English language. This file was parsed and translated into the other languages in order to create sentiment based word-lists for other languages. By doing this, we can now derive the sentiment analysis for tweet results of all languages.

Topic Analysis:

We implemented the topic modelling using **latent Dirichlet allocation** (**LDA**). **Latent Dirichlet allocation** (**LDA**) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

SCREENSHOTS OF INFORMATION RETRIEVAL SYSTEM:



Country	Tweet text	Sentiment	poi name	Topic
India	@RubikaLiyaquat Rubika .... Bolo ab Modi Modi Modi 😂😂😂😂,@RubikaLiyaquat Rubika .... Bolo ab Modi Modi Modi 😂😂😂😂	neutral	RubikaLiyaquat	Journalism
India	@narendramodi @AbeShinzo India ka sub se barha hrami modi kotta modi kotta modi kotta	neutral	narendramodi	Politics
India	@RubikaLiyaquat Yea #rubikaliyaquat modi media and Godi media you no for news you for modi dadal,@RubikaLiyaquat Yea #rubikaliyaquat modi media and Godi media you no for news you for modi dadal	neutral	RubikaLiyaquat	Journalism
India	RT @GoSlowplz: Subramanian swamy slammed his own Modi govt, Called modi as uneducated PM. https://t.co/I3RlJDgk2k,RT @GoSlowplz: Subramanian swamy slammed his own Modi govt, Called modi as uneducated PM. https://t.co/I3RlJDgk2k	positive	AcharyaPramodk	Politics
India	@RubikaLiyaquat Namune bhare pade he Rubika jaise TV pr.. Bss modi modi hi karenge.. Economic slowdown pr koi baat nhi.. #dalalrubika #DalalMedia,@RubikaLiyaquat Namune bhare pade he Rubika jaise TV pr.. Bss modi modi hi karenge.. Economic slowdown pr koi baat nhi.. #dalalrubika #DalalMedia	positive	RubikaLiyaquat	Journalism
India	@Amitjanhit Anjana Om Modi Rohit Sardana Rajat Sharama Dewagan Rubika Liyakat Jaise Satta Ke Dalal Hai Modi Ke Liye Bhaduaa Giri Karate Hai,@Amitjanhit Anjana Om Modi Rohit Sardana Rajat Sharama Dewagan Rubika Liyakat Jaise Satta Ke Dalal Hai Modi Ke Liye Bhaduaa Giri Karate Hai	neutral	RubikaLiyaquat	Journalism
India	#AnjanaOmKashyap pramod bhai m. B. B. S pramod modi prime minister one day. Narendra modi ki sign mardena uske letter par pass kar dunga.,#AnjanaOmKashyap pramod bhai m. B. B. S pramod modi prime minister one day. Narendra modi ki sign mardena uske letter par pass kar dunga.	neutral	AcharyaPramodk	Politics
India	@AmitShah Hare amit beta modi toh chor hai na mard hai modi aur tu tum dono paiso ke liye chudwate ho mujhe sab pata hai na mardo ki sarkar modi sarkar	neutral	AmitShah	Politics
India	RT @Vijaysubbu1: With PC going to jail, Modi bhakths have lost their mind! Their loyalty to Modi doesnt permit them to give the credit to D.,RT @Vijaysubbu1: With PC going to jail, Modi bhakths have lost their mind! Their loyalty to Modi doesnt permit them to give the credit to D.	negative	AcharyaPramodk	Politics
India	@RubikaLiyaquat @ABPNews Hello..... Rubika Modi😊😊😊,@RubikaLiyaquat @ABPNews Hello..... Rubika Modi😊😊😊	neutral	RubikaLiyaquat	Journalism

Fig – Query Modi is made using advanced search feature to give tweets only from country India

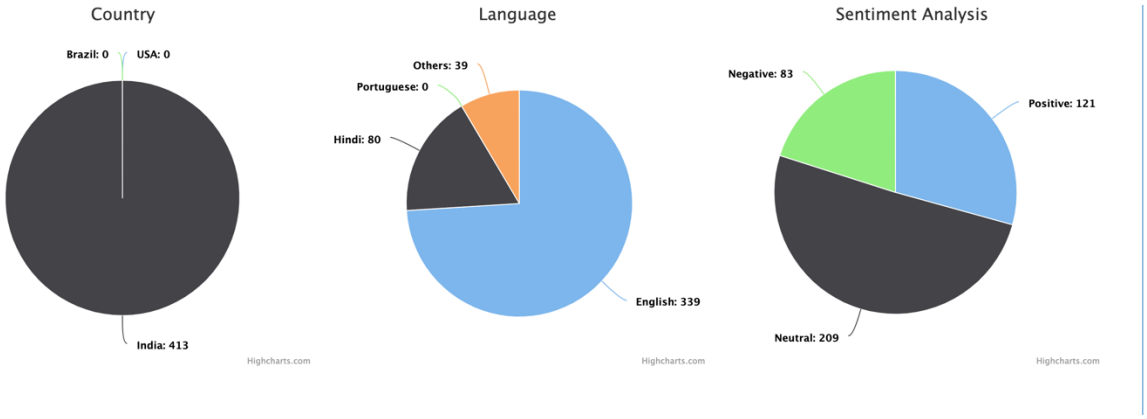


Fig – Pie Charts Displaying various analysis of the tweets for the Query Modi for country India

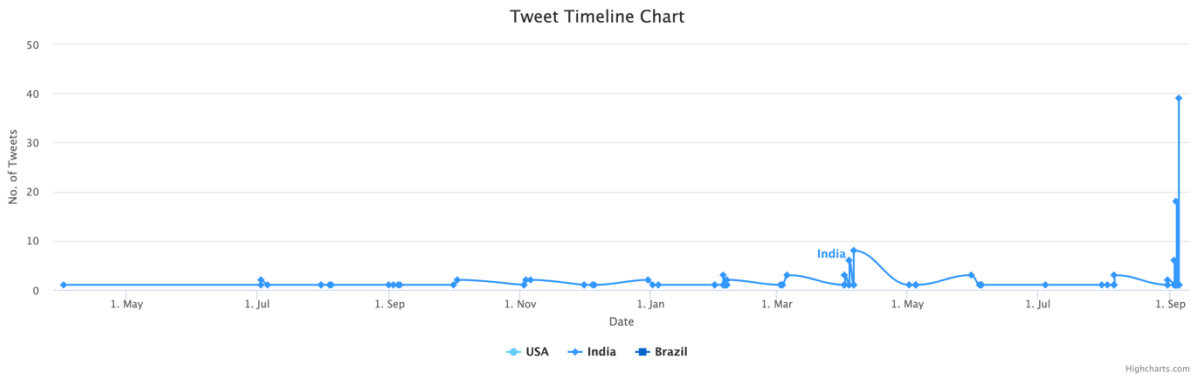


Fig – Timeline Graph Showing the date on which the tweets returned by the system were made

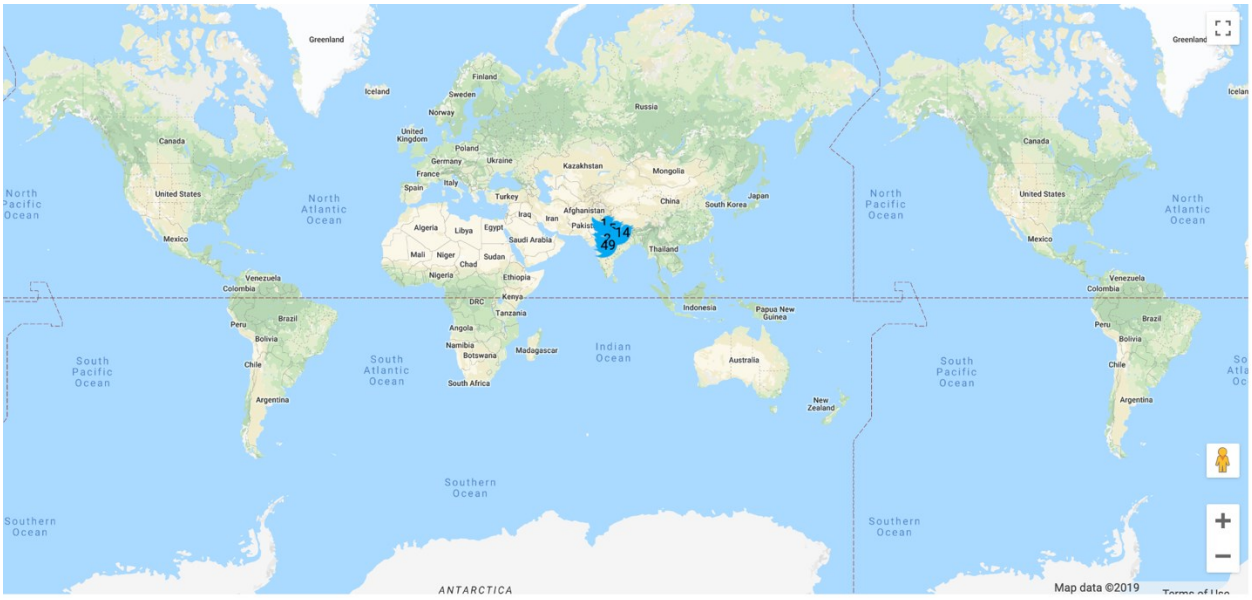
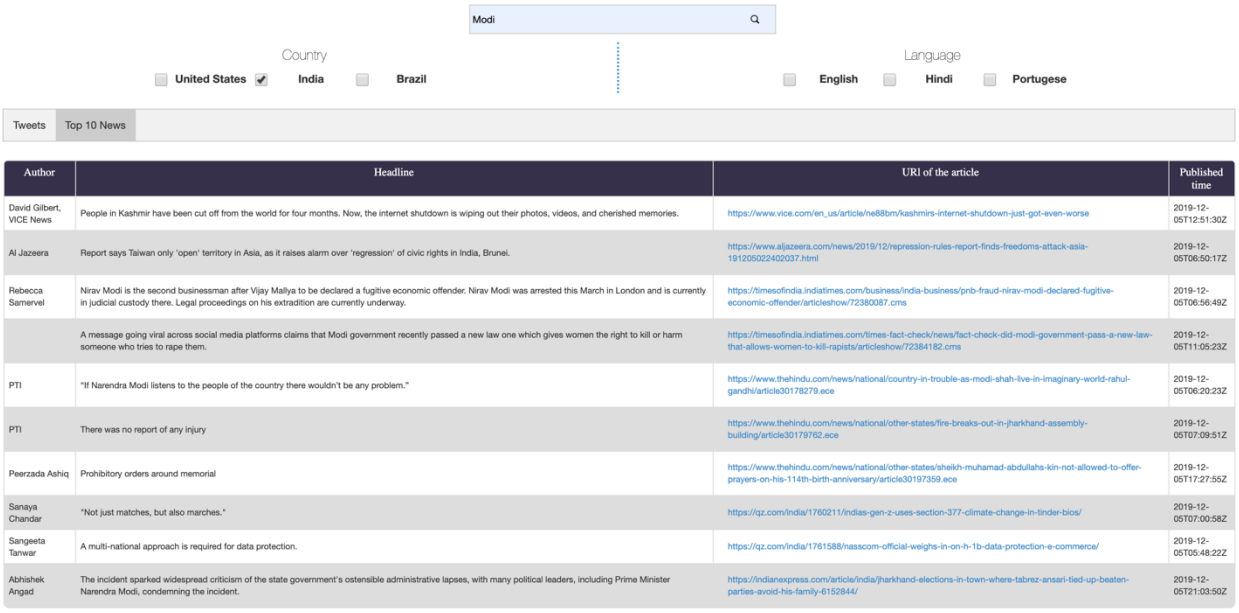


Fig – Map Showing the locations of the user from where the tweet is made



Author	Headline	URI of the article	Published time
David Gilbert, VICE News	People in Kashmir have been cut off from the world for four months. Now, the internet shutdown is wiping out their photos, videos, and cherished memories.	https://www.vice.com/en_us/article/ne88bm/kashmirs-internet-shutdown-just-got-even-worse	2019-12-05T12:51:30Z
Al Jazeera	Report says Taiwan only 'open' territory in Asia, as it raises alarm over 'regression' of civic rights in India, Brunei.	https://www.aljazeera.com/news/2019/12/repression-rules-report-finds-freedoms-attack-asia-191205022402037.html	2019-12-05T06:50:17Z
Rebecca Samervel	Nirav Modi is the second businessman after Vijay Mallya to be declared a fugitive economic offender. Nirav Modi was arrested this March in London and is currently in judicial custody there. Legal proceedings on his extradition are currently underway.	https://timesofindia.indiatimes.com/business/india-business/pnb-fraud-nirav-modi-declared-fugitive-economic-offender/articleshow/72380087.cms	2019-12-05T06:56:49Z
	A message going viral across social media platforms claims that Modi government recently passed a new law one which gives women the right to kill or harm someone who tries to rape them.	https://timesofindia.indiatimes.com/times-fact-check/news/fact-check-did-modi-government-pass-a-new-law-that-allows-women-to-kill-rapists/articleshow/72384182.cms	2019-12-05T11:05:23Z
PTI	"If Narendra Modi listens to the people of the country there wouldn't be any problem."	https://www.thehindu.com/news/national/country-in-trouble-as-modi-shah-live-in-imaginary-world-rahul-gandhi/article30178279.ece	2019-12-05T06:20:23Z
PTI	There was no report of any injury	https://www.thehindu.com/news/national/other-states/fire-breaks-out-in-jharkhand-assembly-building/article30179762.ece	2019-12-05T07:09:51Z
Peerzada Ashiq	Prohibitory orders around memorial	https://www.thehindu.com/news/national/other-states/sheikh-muhamad-abdullahs-kin-not-allowed-to-offer-prayers-on-his-114th-birth-anniversary/article30197359.ece	2019-12-05T17:27:55Z
Sanaya Chandar	"Not just matches, but also marches."	https://qz.com/india/1760211/indias-gen-z-uses-section-377-climate-change-in-tinder-bios/	2019-12-05T07:00:58Z
Sangeeta Tanwar	A multi-national approach is required for data protection.	https://qz.com/india/1761588/nasscom-official-weighs-in-on-h-1b-data-protection-e-commerce/	2019-12-05T05:48:22Z
Abhishek Angad	The incident sparked widespread criticism of the state government's ostensible administrative lapses, with many political leaders, including Prime Minister Narendra Modi, condemning the incident.	https://indianexpress.com/article/india/jharkhand-elections-in-town-where-tabrez-ansari-tied-up-beaten-parties-avoid-his-family-6152844/	2019-12-05T21:03:50Z

Fig – Relevant News Articles published in various News Websites

VIDEO DEMONSTRATION:
We demonstrated the functionality of our IR system in a video available at:
http://tweetengine.herokuapp.com/static/img/output.mp4

PROJECT CONTRIBUTIONS:

TEAM MEMBER	CONTRIBUTIONS
Deepakanuraag Sathyanarayanan	Pie Charts, TimeLine Graphs, Map Analysis and Language Translation,Video Explanation
Adithyakrishna kovai srinivasan	Complete UI Design and Implementation, news articles, search integration.
Rohit Aggarwal	Preprocessing of Tweets, Query Parser, Sentiment Analysis and Report.

REFERENCES:

1) Geocoding API - https://developers.google.com/maps/documentation/geocoding/start
2) Google Translate API - https://cloud.google.com/translate/docs/
3) Sentiment Analysis - https://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
4) High Charts - https://api.highcharts.com/highcharts/
5) News Articles - https://newsapi.org/
6) Topic Analysis - https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24