Multilingual Search Engine

Sri Sai Bhargav Koritala
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
srisaibh@buffalo.edu

Suhash Bollu
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
srisaibh@buffalo.edu

Bhakti Jadhav
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
bhaktiha@buffalo.edu

Sandeep Kumar Yadlapalli
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
syadlapa@buffalo.edu

Abstract

The objective of this project is to analyze impact of political rhetoric of influential persons of interest from 3 countries namely India, Brazil and USA by monitoring social media. For analyzing impact, we have examined the response of around 100,000 tweets of influential personalities. We have built a complete end-to-end IR solution involving content ingestion, search, sentimental and topic analysis. We have also done analytics on collected tweets and added visualizations such as tag clouds for topic, hashtags, distribution of tweets over geography, country and languages.

1 Introduction

Information on the web is growing at an exponential rate in various languages and forms. Since the evolution of internet, it has been used for communicating information over the geography. Thought English remains the most widely used mode of communication there are also considerable number of documents on web in other languages also. We have collected and analyzed tweets in English, Hindi and Portuguese.

We have crawled profiles of 15 Persons of Interest from 3 countries from Twitter and indexed the tweets and replies in 3 languages using Solr.

1.1 Tweet Crawling and Indexing

Below are some key features and functionalities implemented by us in this project :

1. **Search Window** : The dashboard provides a search window wherein user can type search query and get relevant results from Apache Solr.

2. **Twitter User Information** : Corresponding to tweets retrieved from Solr, Person of Interest details like name and verification are provided.

3. **Tweet Information** : The tweet link is provided which upon clicking redirects the user to the tweet page.

4. **Time Series** : The tweets include timestamps (using Solr date fields), Banana provides different levels of granularity allowing for visualization of historical trends.

5. **Heat Map** : Heat Map shows the density. We have selected row field as Country and column field as Topic which gives us distribution of topics over country. For example: Heat Map can be used to depict political scenarios and issues prevailing in the country.

6. **Date based Filtering** : This filter allows user to retrieve tweets for a date range.

7. **Country based Filtering** : The user can retrieve tweets based on the country by applying this filter.

8. **Language based Filtering** : Using this filter user can retrieve tweets pertaining to a language.

9. **Topic based Filtering** : This filter allows user to retrieve tweets for a topic.

10. **Sentiment based Filtering** : The tweet words are analyzed for finding out the sentiments related to it. This filter helps us to retrieve tweets based on Positive, Negative or Neutral sentiments.

11. **Country Distribution Graph** : This graph shows the distribution of retrieved tweets over the geographic location.

12. **Language Distribution Graph** : This graph is plotted for the distribution of retrieved tweets over multiple languages.

13. **Hashtag based Tag Cloud** : For good visualization and UI experience we have also added a tag cloud containing hashtags according to retrieved tweets.

14. **Pagination** : We have also implemented pagination to display 10 tweets per page. A user needs to navigate to next page to see more tweets on the same query.

2 Implementation Details:

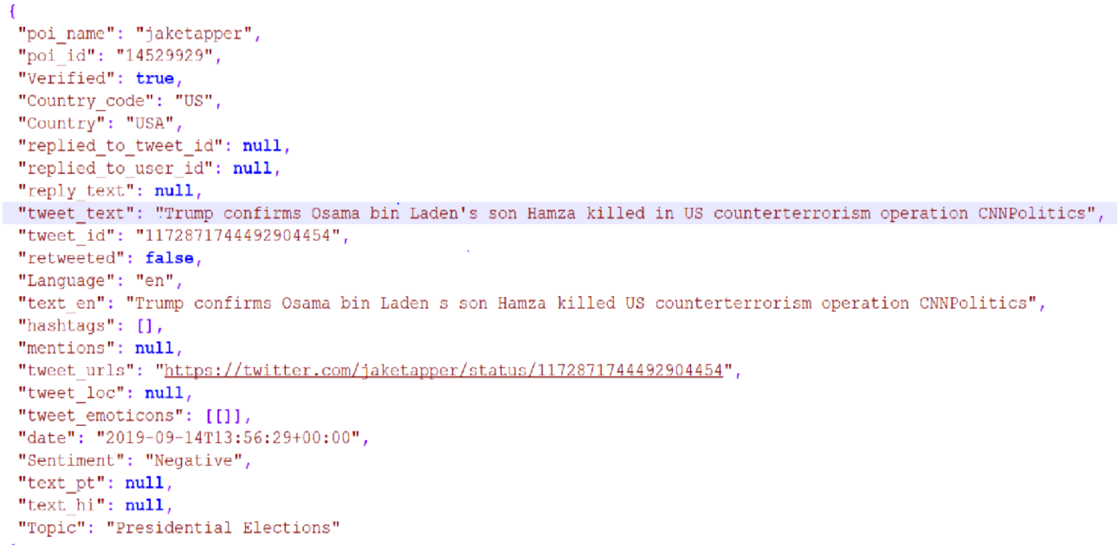
2.1 Tweet Crawling and Indexing :

We have used the tweets with collected for Project in the form of JSON files. There are 100,000 tweets for following Person of Interests :

1. realDonaldTrump
2. Alexandregarcia
3. BernieSanders
4. BolsonaroSP
5. CarlosBolsonaro
6. ewarren
7. jairbolsonaro
8. jaketapper
9. KamalaHarris
10. ManuelaDavila
11. narendramodi
12. PiyushGoyal
13. rajnathsingh
14. rashtrapatibhvn
15. yadavakhilesh

The example of tweet with field names is as given below :

Sentiment and Topic fields are added after processing the tweets. You can see that the **Sentiment** of the given tweet is **Negative** and **Topic** is **Presidential Elections**.



```
{
"poi_name": "jaketapper",
"poi_id": "14529929",
"Verified": true,
"Country_code": "US",
"Country": "USA",
"replied_to_tweet_id": null,
"replied_to_user_id": null,
"reply_text": null,
"tweet_text": "Trump confirms Osama bin Laden's son Hamza killed in US counterterrorism operation CNNPolitics",
"tweet_id": "1172871744492904454",
"retweeted": false,
"Language": "en",
"text_en": "Trump confirms Osama bin Laden s son Hamza killed US counterterrorism operation CNNPolitics",
"hashtags": [],
"mentions": null,
"tweet_urls": "https://twitter.com/jaketapper/status/1172871744492904454",
"tweet_loc": null,
"tweet_emoticons": [[]],
"date": "2019-09-14T13:56:29+00:00",
"Sentiment": "Negative",
"text_pt": null,
"text_hi": null,
"Topic": "Presidential Elections"
}
```

Figure1: Tweet JSON

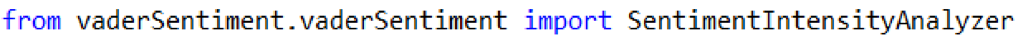
2.2 Tweet Filtering:

The tweets contain unwanted data such as URL, special characters which when present in tweet lowers the precision. So, we have filtered out the URLs and special characters present in the tweet while content ingestion from Twitter.

2.3 Sentimental Analysis:

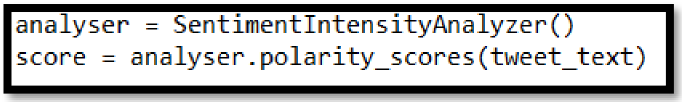
Sentiment Analysis is a sub-field of Natural Language Processing (NLP). Sentiment analysis is contextual mining of text which identifies and extracts subjective information in tweets and helping us understand the social sentiment of the people on particular tweet. The aim of this is to gauge the attitude, sentiments, evaluations, attitudes and emotions of a POI and other people based on the computational treatment of subjectivity in a tweet and replies.

Vader (**Valence Aware Dictionary and Sentiment Reasoner**) is a lexicon and rule-based sentiment analysis tool that is ***specifically attuned to sentiments expressed in social media***. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as positive, negative or neutral.



```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

We used **polarity_scores()** method to obtain the polarity indices for the given sentence.



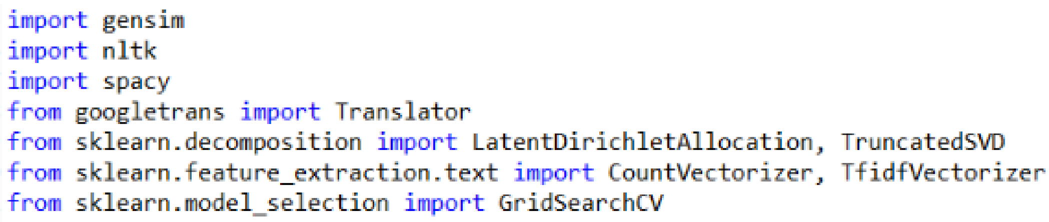
```
analyser = SentimentIntensityAnalyzer()
score = analyser.polarity_scores(tweet_text)
```

2.4 Topic Analysis:

Knowing the evolution or the segmentation of an account's followers can give actionable insights into real time concerns of users. Carrying topic analysis of followers of politicians, journalists and social activists can produce a complementary view of opinion polls.

The goal of topic analysis is to find out what is being are tweeted about through Topic Modeling of timelines.

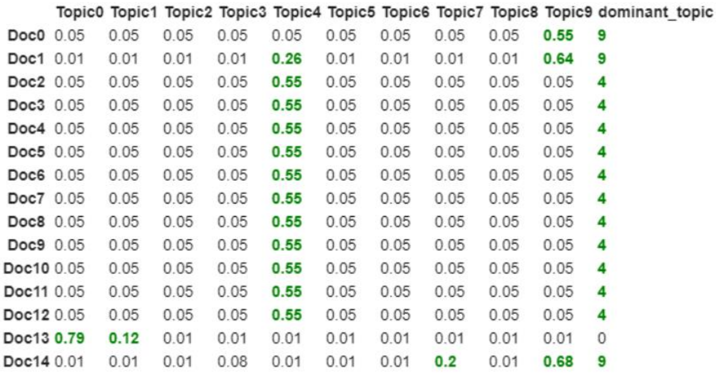
For topic analysis, we explored LDA topic modelling method which is unsupervised model. We used the following packages for implementation :



```
import gensim
import nltk
import spacy
from googletrans import Translator
from sklearn.decomposition import LatentDirichletAllocation, TruncatedSVD
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import GridSearchCV
```

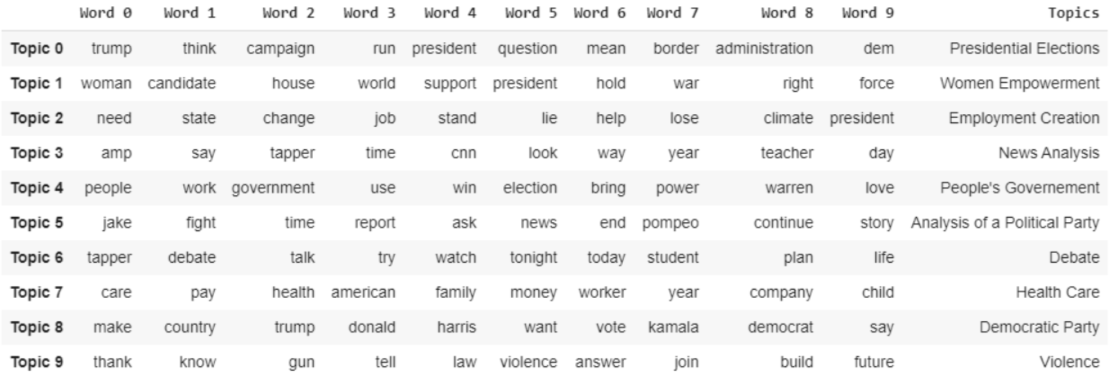
For tweets, the order of transformations is:
sent_to_words() –> Stemming() –> vectorizer.transform() –> best_lda_model.transform().

1. tokenize each sentence into a **list of words**, removing punctuations and unnecessary characters altogether.
2. ***Spacy*** package we used here for stemming.
3. To classify a document as belonging to a particular topic, a logical approach is to see which topic has the highest contribution to that document and assign it.
4. We have greened out all major topics in a document and assigned the most dominant topic in its own column.



	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	dominant_topic
Doc0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.55	9
Doc1	0.01	0.01	0.01	0.01	0.26	0.01	0.01	0.01	0.01	0.64	9
Doc2	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc3	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc4	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc5	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc6	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc7	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc8	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc9	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc10	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc11	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc12	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05	0.05	0.05	4
Doc13	0.79	0.12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0
Doc14	0.01	0.01	0.01	0.08	0.01	0.01	0.01	0.2	0.01	0.68	9

Figure2: Document-Topic Matrix



	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Topics
Topic 0	trump	think	campaign	run	president	question	mean	border	administration	dem	Presidential Elections
Topic 1	woman	candidate	house	world	support	president	hold	war	right	force	Women Empowerment
Topic 2	need	state	change	job	stand	lie	help	lose	climate	president	Employment Creation
Topic 3	amp	say	tapper	time	cnn	look	way	year	teacher	day	News Analysis
Topic 4	people	work	government	use	win	election	bring	power	warren	love	People's Governement
Topic 5	jake	fight	time	report	ask	news	end	pompeo	continue	story	Analysis of a Political Party
Topic 6	tapper	debate	talk	try	watch	tonight	today	student	plan	life	Debate
Topic 7	care	pay	health	american	family	money	worker	year	company	child	Health Care
Topic 8	make	country	trump	donald	harris	want	vote	kamala	democrat	say	Democratic Party
Topic 9	thank	know	gun	tell	law	violence	answer	join	build	future	Violence

Figure3: Words corresponding to Topics

3 Architecture

We have used Banana dashboard for achieving complete IR solution. The Banana is forked from Kibana and works with data stored in Apache Solr to create a rich and flexible UI where various visualizations are stored on dashboard.

The dashboard contains control for search query inputs and quantitative displays over the results of the query and runs as a client-side application for visualizations like charts, maps, time series etc.

Banana Dashboard

1. Downloaded the open source banana framework from Lucidworks.
2. Integrated it with Solr already installed on AWS server.
3. Usually we run Apache Solr on server mode. But as Banana architecture requires collection which is set of cores and not a single core. So we had to run Solr on cloud mode for getting it hosted.

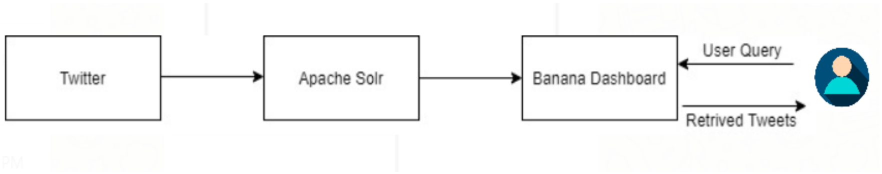


Figure3: Model Architecture

4 Graphical User Interface

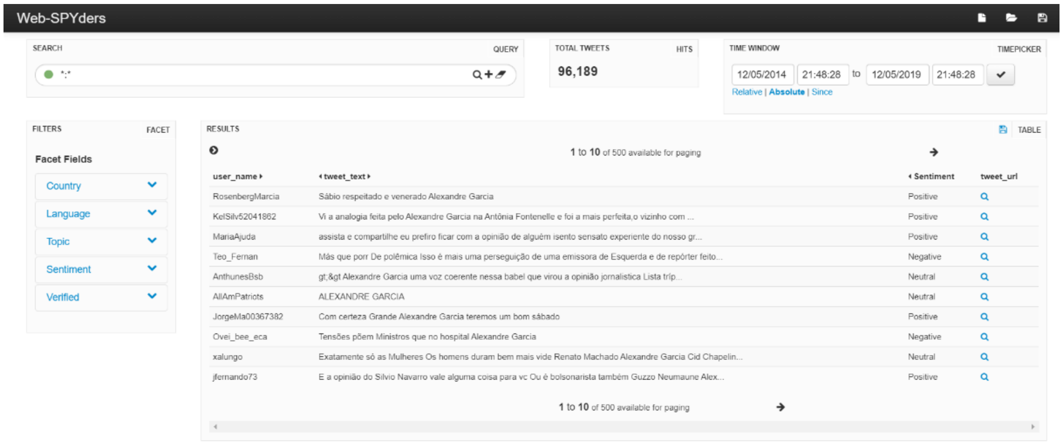
Given below are some snapshots of our web-application named Web-SPYders. The landing page consists of a Search window wherein the user can type the query and hit search icon to get the results.

Landing Page:

Search Bar allows the user to lookup a query that the user needs information about. This also displays the total number of tweets in corpus. Date is the default filter which is applied on the data. The fields that we are interested in include user_name, tweet, the corresponding sentiment and the link to the tweet as a hyperlink.

Search Filters

1. Country – India, Brazil, USA
2. Language – Hindi, Portuguese, English
3. Topic – Topics that were inferred from the topic analysis by using genism
4. Sentiment – Neutral, Positive, Negative
5. Verified – if the user is verified then true and false if not.



Web-SPYders

SEARCH QUERY

:

TOTAL TWEETS HITS

96,189

TIME WINDOW TIMEPICKER

12/05/2014 21:48:28 to 12/05/2019 21:48:28

Relative | Absolute | Since

FILTERS FACET

Facet Fields

Country

Language

Topic

Sentiment

Verified

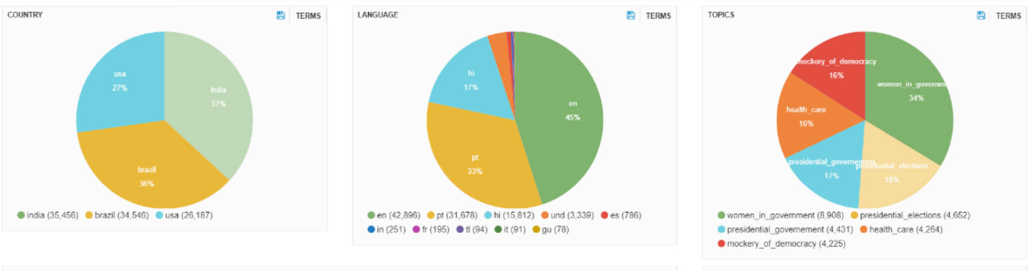
RESULTS TABLE

1 to 10 of 500 available for paging

user_name	tweet_text	Sentiment	tweet_url
RosenbergMarcia	Sábio respeitado e venerado Alexandre Garcia	Positive	🔍
KelSilv52041862	Vi a analogia feita pelo Alexandre Garcia na Antônia Fontenelle e foi a mais perfeita,o vizinho com ...	Positive	🔍
MariaAjuda	assista e compartilhe eu prefiro ficar com a opinião de alguém isento sensato experiente do nosso gr...	Positive	🔍
Teo_Fernan	Más que porr De polêmica Isso é mais uma perseguição de uma emissora de Esquerda e de repórter feito...	Negative	🔍
AnthunesBsb	gt;&gt Alexandre Garcia uma voz coerente nessa babel que virou a opinião jornalistica Lista tríp...	Neutral	🔍
AllAmPatriots	ALEXANDRE GARCIA	Neutral	🔍
JorgeMa00367382	Com certeza Grande Alexandre Garcia teremos um bom sábado	Positive	🔍
Ovei_bee_eca	Tensões põem Ministros que no hospital Alexandre Garcia	Negative	🔍
xalungo	Exatamente só as Mulheres Os homens duram bem mais vide Renato Machado Alexandre Garcia Cid Chapelin...	Neutral	🔍
jfernando73	E a opinião do Silvio Navarro vale alguma coisa para vc Ou é bolsonarista também Guzzo Neumaune Alex...	Positive	🔍

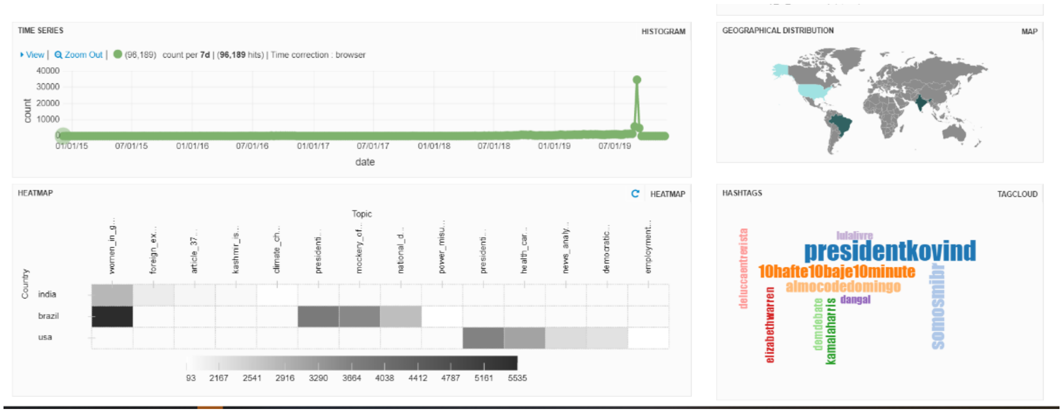
1 to 10 of 500 available for paging

Distribution of tweets across the Country, Language and Topics are represented by a pie chart, that gives the percentage of the corresponding filter in the data.



Time Series shows the distribution of tweets across the year, month and day and their corresponding counts.

Geographical distribution gives a fair view of the tweets in map.



5 Conclusion

The basic idea behind creating the project is to correlate and analyze the basics of an Information Retrieval modelling and to create in harmony an union of the search system essentials. The focus of the application is to attend the informational need of the user. We have been able to cater the keyword search and the analysis which has been the focus of this project. As to draw more visualizations from the data that is presented i.e. corpus data and the query results, we have managed to draw meaningful conclusions by applying concepts of machine learning like topic and sentimental analysis. Although we are limited by the data that we present in the corpus, we have used advanced pre-processing and post-processing techniques that are visually more conclusive than the data itself. To amalgamate all the resources and the knowledge that have been put to work for this application, we have a basic and informative Web Search System, that can be built to cater more of the navigational and transactional aspects of web search.

6. Team Contributions

Team Member	UBIT Name	UB PersonID	Tasks
Bhakti	bhaktiha	50321322	Front end &Backend, Visualization(banana) Solr
Bhargav	srisaibh	50316982	LDA Analysis, Sentiment Analysis
Suhash	suhashbo	50317925	Front end &Backend, Visualization(banana) Solr
Sandeep	syadlapa	50313888	Data Collection& Formatting, Translation

References

[1] https://doc.lucidworks.com/lucidworks-hdpsearch/2.5/Guide-Banana.html

[2] https://medium.com/@yanlinc/how-to-build-a-lda-topic-model-using-from-text-601cdcbfd3a6

[3] https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

[4] https://medium.com/@alexisperrier/topic-modeling-of-twitter-timelines-in-python-bb91fa90d98d