Arab Data Bodies: Arab Futurism Meets Data Feminism

By Mustapha Bouchaqour
NYC College of Technology
New York

Mentor: Professor Laila Shereen Sakr
Department of Film and Media Studies. University of California Santa Barbara

Mentor: Professor Wole Oyekoya
Associate Professor at Hunter College
New York

ABSTRACT

Arab data bodies project is a mixed reality cinematic world of historical data and public spaces. It is basically a VR documentary of the Egyptian Uprising of 2011 that enables participant to experience the historical moment virtuality on the ground. The datasets used in this project come from R-Shief server. The R-Shief is an incredible archive that has been collecting millions of social media posts in 30 languages since 2008 including data that is related to social movements that occurred in Arab countries since 2011.

Since the project is centered on data, it is safe to say that data analysis seems to be the first steps toward the project's goal. The use of tools such as Python, Gephi, Unity3D unlock many features and insights that data may consist of. In these papers I'll walk you through the steps taken in order to achieve the project's goal which is creating a demo that will show how the Tweet text analysis will be distributed over several avatars.

KEYWORDS

Mixed Reality, datasets, data analysis, Gephi, Unity3D, Avatars

1 INTRODUCTION

In 2011, a strong wave of Arab revolution started. The story started in Tunisia when Mohamed Bouazizi who set himself on fire on 17 December 2010 in Ben Arous after being harassed by municipal officials catalyzed the Jasmine Revolution in Tunisia. This revolution later inspired a wider pro-democracy protest movement in the Middle East and North Africa known as the Arab Spring. That was a little background about the story as a big picture.

However, based upon this story, a huge dataset was collected and safeguarded in R-Shief server. The data consists of 87,707,630 records. Each records represent a tweet text that is in 58 languages recorded between March 2011 and June 2013. The virtual story we aim to address at these papers is to graph the sentiments analysis usng Tweet data and analysis the output as an avatar's reaction given at specifc date.

To do so, we have to get familiar with the datasets as a first step. Python was the tool I used to extract, analyze, and visualize the data. I used Gephi to create a network interaction using Tweet's nodes and edges. The sentiments analysis was the input for creating an ML-Agents that will be implemented in Unity3D. For the next sections, I will emphasis the road map I took to analyze data and finalize the demo or the project's story

2 TEAM-WORK

My research mentors are Professor Laila and Professor Wole. We are at least 10 active members in the team including PhD students and other professors from university of California and Egypt. My team ultimate goal is to create a game that will play as future trend for the data gathered. In other words, the team's story idea is to have a future picture that shift 2011 to 2111. Thus, seeing the data in a century from 2011,

It is an Arab futuristic world where the history of the 21st century is one where data and artificial intelligence have created "data bodies (DB)". In a hundred years from now, human, and non-human subjectivities are created from data to data births individuality.

The mechanism of this story sets in the future that locates the 2011 Arab Uprisings as the birth of the digital activism we witnessed grow globally throughout the twenty-first century—from Tunisia to Cairo to Occupy Wall Street, from 5M and 12M in Spain to the Umbrella Revolution in Hong Kong, and more… The player enters a public mass gathering brimming with the energy of social change and solidarity. The player has from sunrise to sunrise to interact with "data bodies."

My team is in stage of designing a blueprint for the game. However, I was so limited with the time given during the research period. Thus, I got to act the way that will allow me to have enough time to understand the ultimate 's goal, at the same time to accelerate the data processing, and applying some virtual reality. For the next sections, I will explain in more details what was the process I took through building my individual project.

3 R-SHIEF: INTRO TO DATASET

As I access R-Shief where data was collected, I found out that dataset was structured and centered around Tweet table. The R-Shief Twitter data is held in a MYSQL database. The data is split between a few different tables including but not limited to:

- tweet
- hashtag
- user
- language
- source
- url

3.1 Data Description

The tweet table glues everything together. It has the following columns:

- twitter_id
- geo
- source
- from_user_id
- to_user_id
- lang_code
- created_at

The user table has the following:

- user_id

- username
- profile_image_url

The 'hashtag' table has the following:

- hashtag_name
- Definition
- Related_Country
- Started_Collecting
- Stopped_Collecting
- hashtag_id

The 'url' table has the following:

username
profile_image_url

The 'hashtag' table has the following:

- hashtag_name
- Definition
- Related_Country
- Started_Collecting
- Stopped_Collecting
- hashtag_id

The 'url' table has the following:

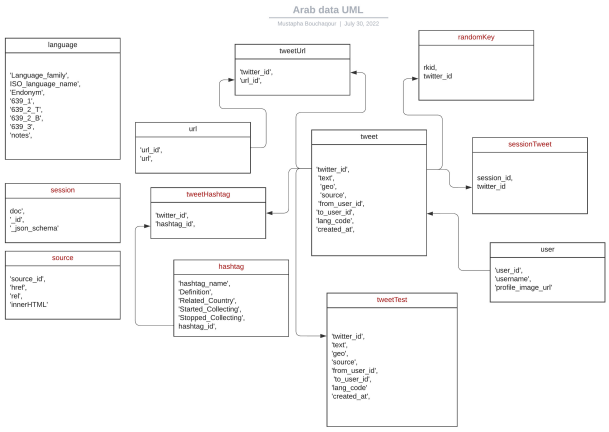
- url_id
- url

You can look at a tweet's user's info by INNER JOIN'ing the tweet table with the user table on the from_user_id column of the tweet table.Because tweets and hashtags, and also tweets and URLS, have a manytomany relationship, they are associated by INNER JOIN'ing on these assocation tables:

- tweetHashtag
- tweetUrl

4 DATA ANALYSIS

Figure 1: Data UML



As it was mentioned earlier, this project is a data-driven project, we kind of want to design a roadmap that will help us understand data in depth. After extracting the data using Python-SQL codes, Figure 1 shows data UML.

Next, we'll apply several analysis over data.

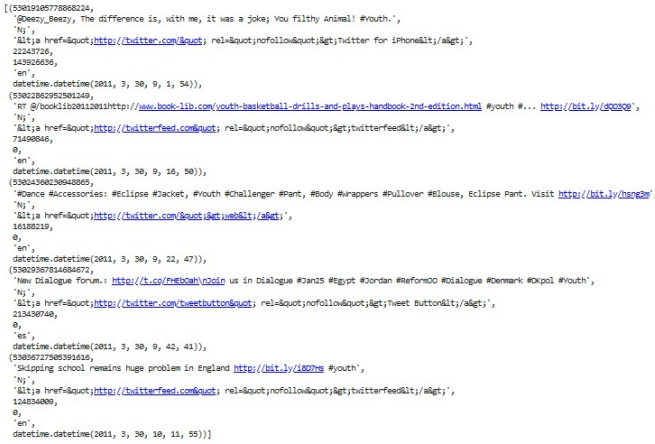
4.1 Time-Series Analysis Concept

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. Our objective is to show how variable variables change over time. In other words, time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Time series analysis typically requires a large number of data points to ensure consistency and reliability.Time series data can be used for forecasting—predicting future data based on historical data.

Given the huge datasets we are dealing with, we wanted to apply time series over the tweet-table to see the change over time.

Figure 2: Showing the first 5 rows in the Tweet table



Reducing the size of data requested from the tweet table by applying filters over data using time searching-word.

Timeframe: August 2011 Keeping only the tweet that contains the word Syria.

In any future analysis applied in the tweet table, we will keep the same filter conditions.

4.2 Graphing Tweet Data Over Time

The two graphs below show the distribution number of tweets in a day within the month August 2011. The function used to graph the time series takes a date (The date you want to analyze) and/or the language you want to filter over.

Figures 1 and 2 show the distribution number of tweets on a given date.

Figure 1 shows that on August 17, 2011, the number of tweets reached the minimum (less than 5 tweets) at 1pm, while the frequency reached the maximum (more than 70 tweets) at 7pm.

Figure 2 shows the frequency of tweets given the same date as August 2011, however, we picked only the English language. At

Figure 3: Tweet over time for the date: 08/17/2011

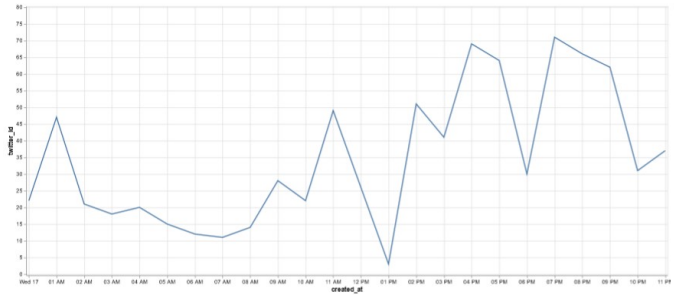
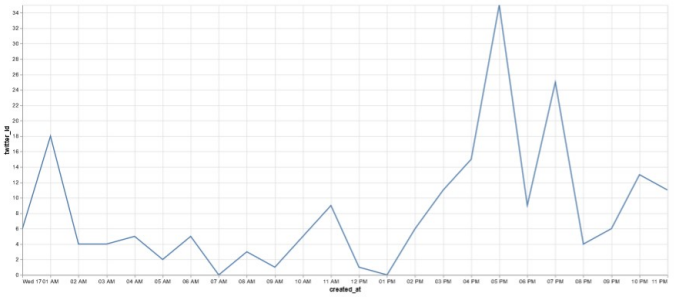


Figure 4: Tweet over time for the date: 08/17/2011 Language: English



exactly 7AM and 1PM there were almost 0 tweets, however, at 5pm the maximum occurred and was around 34 tweets at that specific hour. We may conclude that People tend to tweet either early in the morning or in evening time.

4.3 Time Series Analysis: Heatmap

With data we have, we aim to check and compare the number of tweets per day within August 2011. For that reason, we will use a calendar heatmap.

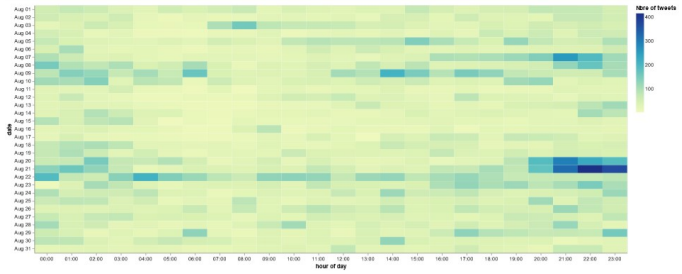
A calendar heatmap uses colored cells, typically in a single base color hue and extended using its shades, tones, and tints like shades of blue from light to dark. It shows a relative number of events for each day in a calendar view. In our case we are arranging data over days while we check the frequency of tweets for each hour a day. That enables us to quickly recognize daily and weekly patterns.

For the sake of simplicity, we created a function that takes a specific date and language and prints out a heatmap graph that starts from that given date. Figure 5 shows the heatmap of tweet data given no language filter.

Reading the heatmap we may conclude that on August 20 and 21 the number of tweets started increasing around 8pm and reached the maximum at 10pm - 416 tweets.

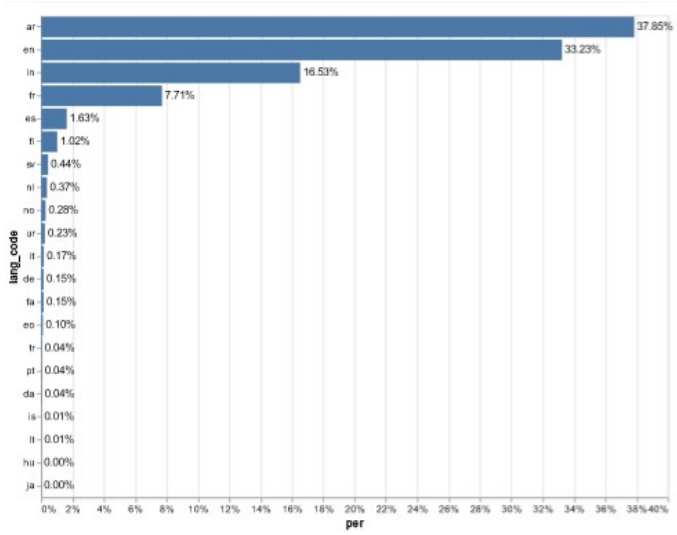
People are more likely to tweet more around August 20 and 21. According to Wikipedia in August 20 “The death toll from the previous day rose to 34, and the Syrian army renewed a siege on Homs with army tanks, firing at the local population to keep them from rallying”, and in August 21 “In a media interview, Assad claimed he wanted to pursue reforms and pursue "terrorists". Assad warned against foreign intervention. Two people died in Hama when Shabbiha randomly opened fire on civilians in the street. The Syrian opposition gathered in Syria for talks on creating a rival government”

Figure 5: Heatmap during August 2011



4.4 Tweet Languages

Figure 6: Language Used for Tweet Data



We have used Languages as a filter for most time series analysis functions, however, it is time to check the languages that are used by people for the same data we have.

Arabic and English seem to dominate the languages used in tweet data.This result was not surprising, however, people also use French to communicate and some other languages including but not limited to Turkish, and some local dialects.

5 ANIMATE DYNAMIC GRAPHS WITH GEPHI

5.1 Preparing The Data

Gephi has several options for loading network data from a database or as graph file types such as .graphml or .gexf. For dynamic graphs, however, the simplest option is to load data into Gephi from correctly labeled and formatted spreadsheets. In network graph terminology, “nodes” represent individual Twitter users and “edges”

	twitter_id	text	geo	source	from_user_id	to_user_id	lang_code	created_at
0	97819131096735744	RT @USAforFreeLibya: Ramadan Mubarak to all Mu...	a:2:{s:11:"coordinates";a:2:{i:0;d:32.78107800...	<a href="http://twitter.com/#!/downloa...	363184099	0	in	2011-08-01 00:01:12
1	97819154215747584	RT @DrFiraa_89: شكرًا على تهنئتك...	N;	<a href=" http://blackberry.com/twitter...	363184099	0	in	2011-08-01 00:01:18
2	97819412761022464	#syria#اهالي درعا تحرر الأوطان #Yemen #libya ...	N;	<a href="http://twitter.com/">...	256549942	0	en	2011-08-01 00:02:20
3	97819445648568322	RT @essamz: اذا سقال لك شاء الله يشدة الى (ال...	N;	<a href=" http://www.echofon.com/"...	206243312	0	ar	2011-08-01 00:02:27
4	97819539819081728	RT @defusertt: the first sahur: praying for th...	N;	<a href=" http://www.tweetdeck.com&quot...	225417448	0	en	2011-08-01 00:02:50

Figure 7: Tweet Table

represent the retweet connections between users. I start with nodes and edges csv files, created using Python code, however, it seems to be another easy way using NetworkX in Python.

Let's take a look at tweet data when we used dataframe to read it Figure 8.

The main three columns we are going to use are from_use_id, to_use_id,created_at, and Lang_code (See description above with tweet data).

Gephi has several options for loading network data from a database or as graph file types such as .graphml or .gexf. For dynamic graphs, however, the simplest option is to load data into Gephi from correctly labeled and formatted spreadsheets. In network graph terminology, "nodes" represent individual Twitter users and "edges" represent the retweet connections between users. I start with nodes and edges csv files, created using Python code, however, it seems to be another easy way using NetworkX in Python.

Gephi requires a nodes spreadsheet with the first column specifically named "Id" containing the Twitter user ids, the second should be "Label" and contain the Twitter user screen names, in my case I kept the use_id as a label. You may also add any other columns but this step is optional.

Next, for the edges spreadsheet, similarly to the nodes, Gephi expects specifically labeled and ordered columns during import. The first two required columns are "Source" and "Target", representing the Twitter user pair engaged in retweeting. The third column should be "Type", which for this Twitter data I have is "language" since we are dealing with retweets. The fourth column should be "Label", which in this case is a simple index, but I won't add anything to this column, so I will have a fourth column in total. The fourth column then is the most important, it should be named "Timeset" and contain the creation time of the retweet — specifically in iso format. The "Timeset" column is the time variable and will be used to animate the network graph in Gephi. It is the value that comes from the column "created_at". Finally, the edges dataframe can be saved as a csv for import into Gephi.

Using the steps described above we were able to graph dynamic data for the period August 2011. The colors you see show the languages people used to tweet with.

Here is the link to a YouTube Video that shows the tweet interactions: https://www.youtube.com/watch?v=qUbvjoPJoCM

We can also graph some static visualization using Gephi Figure 8.

6 SENTIMENTS ANALYSIS

Sentiment Analysis (also known as opinion mining or emotion AI) is a sub-field of NLP that tries to identify and extract opinions within a given text across blogs, reviews, social media, forums, news etc. Sentiment Analysis can help craft all this exponentially

Figure 8: Tweet Nodes and Edges

growing unstructured text into structured data using NLP and open source tools. For example the tweet data we have is a treasure trove of sentiment and users are making their reactions and opinions for every topic under the sun. We will focus on classifying the tweet-text into positive or negative sentiment.

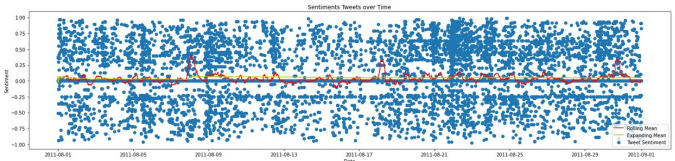


Figure 9: Sentiment Analysis: Tweet Agust 2011 Date

We can notice from Figure 9 some interesting things right off the bat here.

- There are a lot of tweets with a sentiment score of more than 0.
- We have A LOT of data.
- The mean seems somewhat stable across our data
- There seems to be areas of higher density where more tweets are occurring

Let's see if we can get a little bit clearer picture of our sentiment over time. Overall our data is noisy, there is just too much of it. Taking a sample of our data might make it easier to see the trends happening. We'll use the pandas sample() function to retain just a 10 percentage of tweets data .

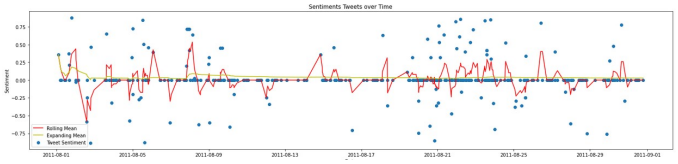


Figure 10: Sentiment Analysis: 10% of Tweet August 2011 Data

Figure 10 is much better, allowing us to actually see some dips and trends in sentiment over time.

- The first week of August graph shows more negative points
- The following weeks of August show more positive sentiment
- The middle of August we have a neutral sentiment on average

Generally, people were optimistic during the period of August 2011. People hope to see some changes coming or maybe to see some of their requests would be approved and taken into consideration by the Syrian regime. However, we all now know that is not the case. The Syrian people have suffered a lot since the revolution started and we hope that sacrifice will bring hope and peace.

Let us dive into some positive and negative sentiments.

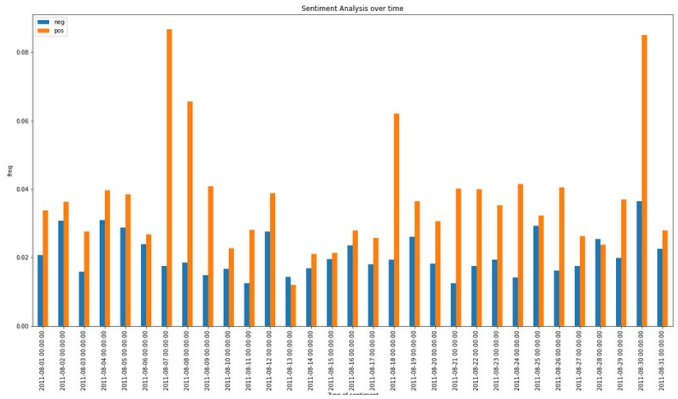


Figure 11: Positive and Negative Sentiment

7 WHAT'S THE STORY?

Team ultimate goal: This is a story set in the future that locates the 2011 Arab Uprisings as the birth of the digital activism we witnessed grow globally throughout the twenty-first century—from Tunis to Cairo to Occupy Wall Street, from 5M and 12M in Spain to the Umbrella Revolution in Hong Kong, and more…The player enters a public mass gathering brimming with the energy of social change and solidarity. The player has from sunrise to sunrise to interact with "data bodies."

My goal: The story I worked on is based on creating an animation using several avatars then use the output of sentiments analysis to check the avatar's reaction.

7.1 Static Expression With Solo Avatar

Figure 12: Projecting Sentiment Analysis output Over Character

Creating Text input in Unity3D to capture the date then going over the data given to script file to pick the percentage of negative and positive that were an output for sentiment analysis. The avatar then will change its face to react upon the data given to it.

- **A**ffective computing: Facial expressions can not only be analyzed, but also be used to generate animation, purely on data.
- **A**nimators: Capture facial expressions with a standard webcam and use it to animate any compatible avatar.

8 CONCLUSION

Arab Data Bodies (ADB) is a data-driven game whose objective is to familiarize, foster ideas and educate players about people power, protest movements, collaboration, empathy, community, and political action.

Player is able to see the avatar's emotion during the period of August 2011. There were a lot of challenges such as implementing different emotions randomly in several avatars. There is an approach to solve that challenge which is using ML-Agent to program solo avatars.

The solid final product should include a text where the player can input data, and several avatars that will react upon the sentiment analysis given that specific date. However, given the short time I have on this project I was unable to really handle that challenge.