MAPPING HAPPINESS AND FRUSTRATION INDEX ON TOP OF THE WORLD MAP THROUGH TWEETS

Karthikeyan Murugesan*, Raja V Palanisamy^A & Chandra P Khatri^A

* *School of Biology, Georgia Institute of Technology*
^A *School of Computational Science & Engineering, Georgia Institute of Technology*

INTRODUCTION

Microblogging sites are a continuous source of diverse user generated data; spanning from politics, weather, the entertainment industry, user sentiments, upto random unclassifiable babble. Twitter is one such site that has grown paranormal in this field with providing a rich repertoire of information to study community behaviour and user sentiments . Once tapped , this data could be analyzed semantically to delimit emotional polarities from each other. Furthermore, a succinct visual representation of the data in real time would highly appeal to the user to follow event detection patterns in the form of sentiments.

Hence in the current study, we have made a framework through which we obtain live feeds from twitter in real-time. Then through this framework we mine the data obtained and perform natural language processing, text classification and sentiment analysis on the data obtained. After cleaning and obtaining the data (sentiments and their locations), through this framework we eventually map 'happiness' and 'frustration' on top of the World Map.

LITERATURE REVIEW

People as a part of large community are acting as sensors and generating lots of meaningful information. Twitter data has been used for predicting stocks[1]. Recently, researchers have developed the efficient techniques to infer the sensible information from social-media during any crisis such as earthquakes and mass emergency[2,3,4]. It has been shown that people switch to social-media, particularly Twitter and Facebook during the incidents.

Shulz et al.[5] have analyzed the real-time incidents on small scale using Microblogs. Riberio et al., [6] have recently implemented the system to observe the events and conditions using Twitter in Portuguese.

Bontcheva and Rout[7] have surveyed all the existing technologies and research done so far in regards to analyzing the social media streams, particularly Twitter. Eventually processing and visualizing the data is also very important. Marcus et al.[8] have depicted the methods and tools which can be used to visualize the data obtained through tweets.

Although in past two years, few researchers have started using Twitter as a source to obtain the data for information retrieval, but many of the above mentioned studies are not for real-time data. Most of the existing applications and tools are for static (historic data).

Furthermore, their approach is not dynamic because most of them confine only to some local region and therefore includes the parameters which fits in best for those regions.
Sentiment analysis is a hard and tricky concept that is widely being pursued by researchers across the world. Sentiment viz[9] is an application on similar lines which estimates and visualizes sentiments from twitter feeds , but it lacks a concrete natural language processing filter for slangs and sarcasm. SentiGraph, CyberEmotions project[10], Sysomos, Radian6[11] are other tangentially relevant commercial social media monitoring applications which focus mainly on web analytics of product brands and not on sentiments per se. This field has not progressed much in non-english languages either.

We are trying to create our own database of words, which would be very specific to our objective. Then we would be performing NLP and customizing the computations for better results. Recently,researchers have tried making sense of social-media streams with the help of NLP and Semantics[12][13] Studies specifically dealing with Twitter streams are of our interest[14][15]

PROPOSED METHOD

The technologies which we are using for the current study are:

- *Python:* For obtaining, mining and processing data throughout the project.
- *Tweepy Python Library:* Used for obtaining structured data from Twitter, based on some parameters (Location, keywords).
- *SQLite:* For storing key-words and their sentiments scores. Act as dictionary, where words are key and sentiment scores are values.
- *Natural Language Toolkit (NLTK):* Used for tokenization and entity extraction from the data obtained.
- *CartoDB:* Open source tool for visualization of geospatial data on the local server. Various dependencies of cartodb is explained in the ‘Completed Tasks’ section.

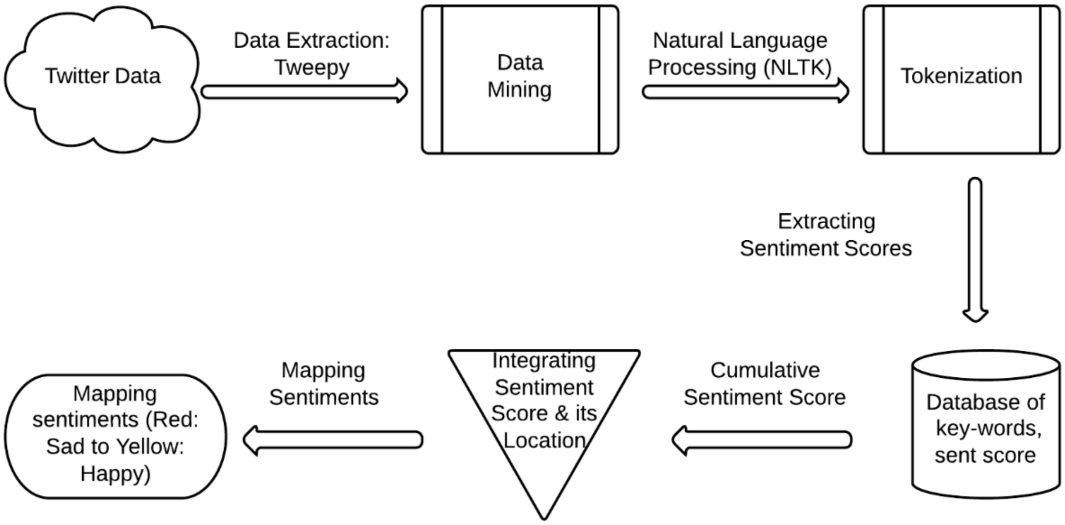

Twitter Data
Data Extraction: Tweepy
Data Mining
Natural Language Processing (NLTK)
Tokenization
Extracting Sentiment Scores
Database of key-words, sent score
Cumulative Sentiment Score
Integrating Sentiment Score & its Location
Mapping Sentiments
Mapping sentiments (Red: Sad to Yellow: Happy)

Figure 1: Methodology

DATA EXTRACTION AND MINING

We are using Tweepy Python library to extract the data from Twitter in structured form. This results in latency issues and limits on client requests[16]. Not all tweets are precisely geo annotated [17][18][19] and the geo annotation of the such tweets depends on self reported location fields.

Twitter feeds are usually very short sentences (140 characters). Hence, selected keyphrases can themselves better explain the content of the tweet. Keywords can be considered as a means of dimensionality reduction, thus making the dataset smaller. For twitter data, literature[20] discusses keyword extraction approaches based on co-occurrence of words (forming a graph of terms with edges derived from the distance between occurrences of a pair of terms and assigning weights to vertices). It has found that this method of keyword extraction was found to perform favourably on Twitter data compared to methods which relied on text models[21].

Tweepy uses Twitter Rest and Twitter Search APIs for obtaining the data. To access the data, a user needs to generate access and token keys. These keys are then integrated with Tweepy Python library. Tweepy returns the data in JSON format. A lot of related information (user_name, user_id, time, location, user_profile details) are also retrieved along with Tweets. For current study, only geo-location and Tweets are retrieved through the data.

We have designed the framework in such a way that we process the data as soon as we receive it, hence we don’t need to save it. Every time a Tweet is obtained, we tokenize it and chop the words. Stop words (e.g. “the”, ”a”, ”is”, ”that” etc.) are removed. Remaining words are then searched in already existing database of key-words, and if the word exist then its sentiment score is obtained. This is done for all the words in obtained Tweet. Eventually, a cumulative sum of score (Sentiment) is appended with the geo-location.

DATA VISUALIZATION

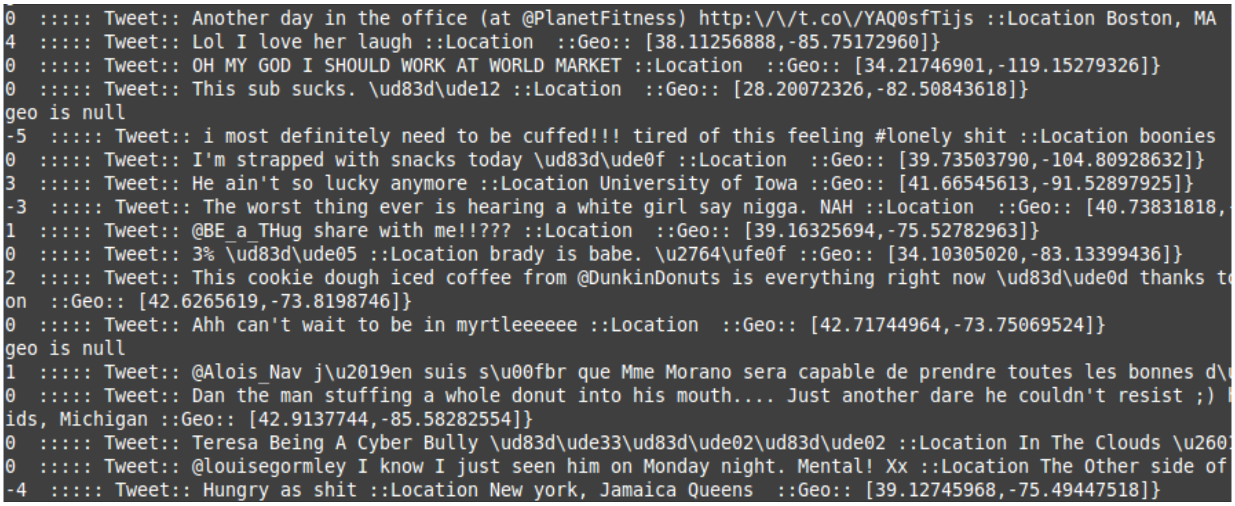
We use an open-source tool called CartoDB[22] which is specifically used for mapping geo-spatial data. It has the following components,

- A User Interface
- A database for storing geospatial data
- SQL API for SQL queries over HTTP
- A Map tiler

CartoDB is written in Ruby and has plenty of dependencies, which are described in the next section. CartoDB can be run on local server with user-interface to upload, edit and visualize the geo-data.

COMPLETED TASKS:

1) Data Extraction and Data Mining: (Completed)



```
0  ::::: Tweet:: Another day in the office (at @PlanetFitness) http:\/\/t.co\/YAQ0sfTijs ::Location Boston, MA
4  ::::: Tweet:: Lol I love her laugh ::Location  ::Geo:: [38.11256888,-85.75172960]}
0  ::::: Tweet:: OH MY GOD I SHOULD WORK AT WORLD MARKET ::Location  ::Geo:: [34.21746901,-119.15279326]}
0  ::::: Tweet:: This sub sucks. \ud83d\ude12 ::Location  ::Geo:: [28.20072326,-82.50843618]}
geo is null
-5  ::::: Tweet:: i most definitely need to be cuffed!!! tired of this feeling #lonely shit ::Location boonies
0  ::::: Tweet:: I'm strapped with snacks today \ud83d\ude0f ::Location  ::Geo:: [39.73503790,-104.80928632]}
3  ::::: Tweet:: He ain't so lucky anymore ::Location University of Iowa ::Geo:: [41.66545613,-91.52897925]}
-3  ::::: Tweet:: The worst thing ever is hearing a white girl say nigga. NAH ::Location  ::Geo:: [40.73831818,
1  ::::: Tweet:: @BE_a_THug share with me!!??? ::Location  ::Geo:: [39.16325694,-75.52782963]}
0  ::::: Tweet:: 3% \ud83d\ude05 ::Location brady is babe. \u2764\ufe0f ::Geo:: [34.10305020,-83.13399436]}
2  ::::: Tweet:: This cookie dough iced coffee from @DunkinDonuts is everything right now \ud83d\ude0d thanks t
on  ::Geo:: [42.6265619,-73.8198746]}
0  ::::: Tweet:: Ahh can't wait to be in myrtleeeeee ::Location  ::Geo:: [42.71744964,-73.75069524]}
geo is null
1  ::::: Tweet:: @Alois_Nav j\u2019en suis s\u00fbr que Mme Morano sera capable de prendre toutes les bonnes d\
0  ::::: Tweet:: Dan the man stuffing a whole donut into his mouth.... Just another dare he couldn't resist ;)
ids, Michigan ::Geo:: [42.9137744,-85.58282554]}
0  ::::: Tweet:: Teresa Being A Cyber Bully \ud83d\ude33\ud83d\ude02\ud83d\ude02 ::Location In The Clouds \u260
0  ::::: Tweet:: @louisegormley I know I just seen him on Monday night. Mental! Xx ::Location The Other side of
-4  ::::: Tweet:: Hungry as shit ::Location New york, Jamaica Queens  ::Geo:: [39.12745968,-75.49447518]}
```

Figure 2: Mined Data (Sentiment Score, Tweet, Geo Location)

Tweets are obtained based on location and their Sentiment is computed. Figure 2 represents the mined data.

2) Computations & Back-end work: (Completed)

2.1 Building database:

- Words (2500 positive and negative words)
- Assigning weights to each word based on its sentiment

2.2 Computation:

- Natural Language Processing: Tokenizing each word, key words extraction from the Tweets obtained
- Sentiment Analysis (based on mined text data, obtaining weight of each word from database, cumulative sum of sentiments for each tweet)
- For each tweet of interest, obtaining overall sentiment and location as final output

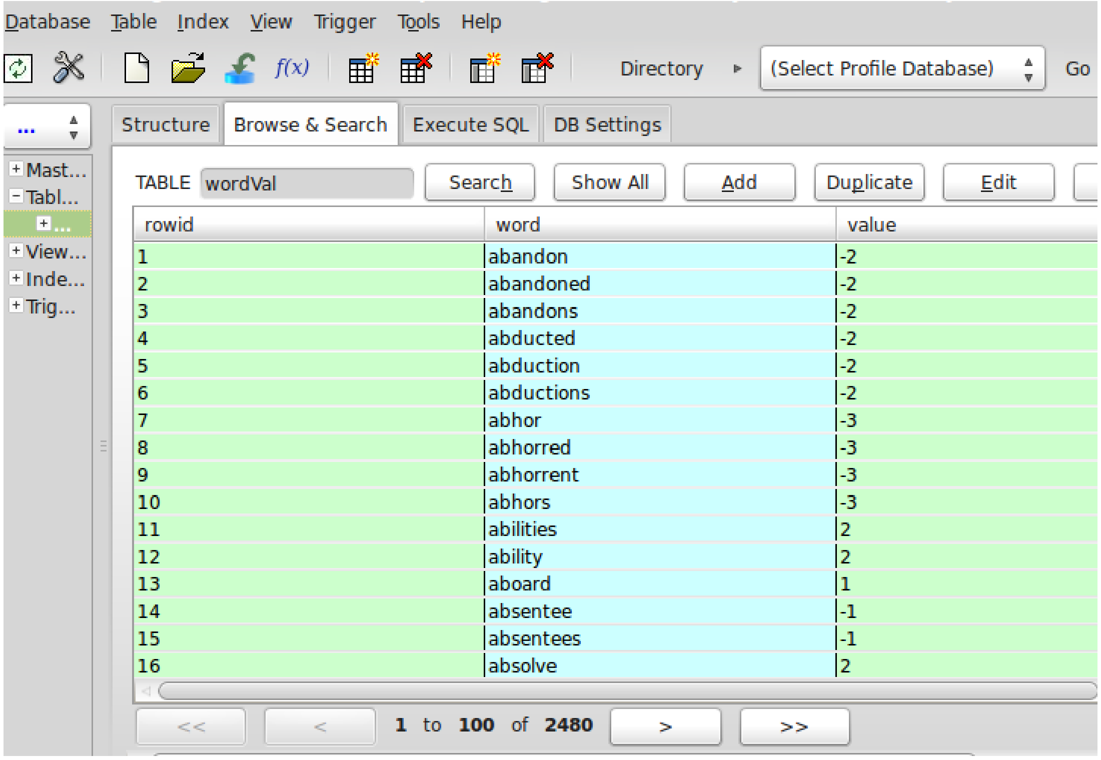


TABLE wordVal

rowid	word	value
1	abandon	-2
2	abandoned	-2
3	abandons	-2
4	abducted	-2
5	abduction	-2
6	abductions	-2
7	abhor	-3
8	abhorred	-3
9	abhorrent	-3
10	abhors	-3
11	abilities	2
12	ability	2
13	aboard	1
14	absentee	-1
15	absentees	-1
16	absolve	2

1 to 100 of 2480

Figure 3: Database of words and their sentiment scores

Use Case:
Tweet: "Happy by Pharrel Williams! Too good."
Sentiment = Cumulative Sent score = 3 (Happy) + 0 (by: stop word) + 0 (Pharrel/Williams: Proper Noun) + 0 (Too: stop word) + 2 (Good)
Senti Score = 5

3) Visualization: (In Progress)

3.1 Building Framework for development of application **(Completed)**

The CartoDB application interface required the installation of a variety of software packages,libraries and dependencies. The application is available open source with the source files on github[23]

The following dependencies had to be installed for CartoDB to run locally -

Table 1: Dependencies for CartoDB

CartoDB-SQL-API	This component powers the SQL queries over HTTP.
GDAL 1.10	Raster support
GEOS 3.3.4	Geometry function support
Mapnik 2.1.1	API for creating and styling map tiles
NodeJS 0.8	Tiler API and SQL API
PostGIS 2.0.x	Geospatial extension for PostgreSQL
Postgres 9.1	Relational database
Redis 2.2	Required for Windshaft & SQL API
Ruby 1.9	Language that CartoDB uses
Varnish 2.1	Web application accelerator
Windshaft	Powers CartoDB Maps API

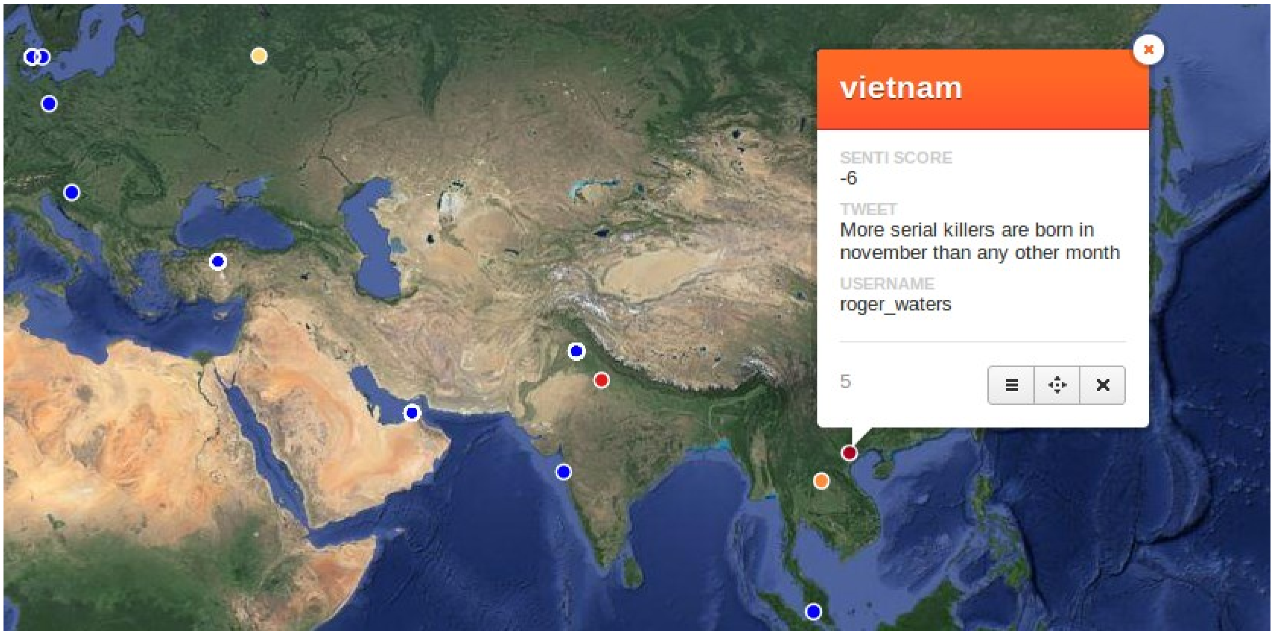
The user interface is up and running and the user can now create ,upload ,edit and visualize the geospatial data

Currently working on running test geo annotated data through the CartoDB API

3.2 Generating the basic visualization **(Completed)**

We have been experimenting with the static version of CartoDB , georeferencing and representing the tweets on the world map.

The following happy, sad and neutral tweets are shown in Figure 4.



vietnam

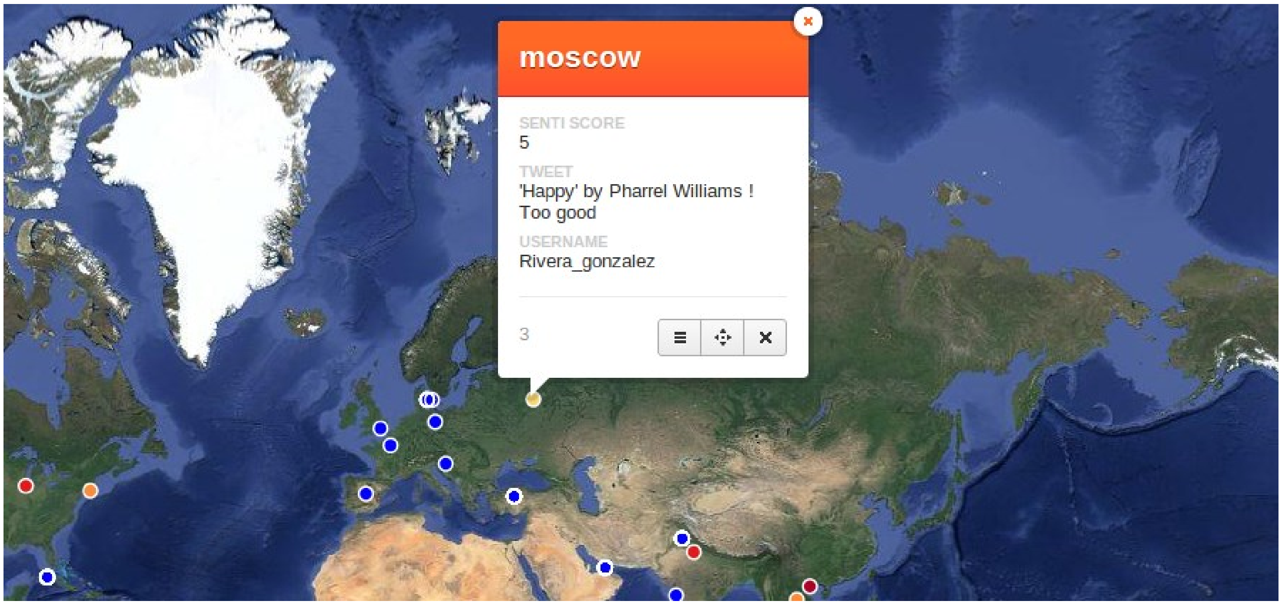
SENTI SCORE
-6

TWEET
More serial killers are born in november than any other month

USERNAME
roger_waters

5

4(a): A sad tweet



moscow

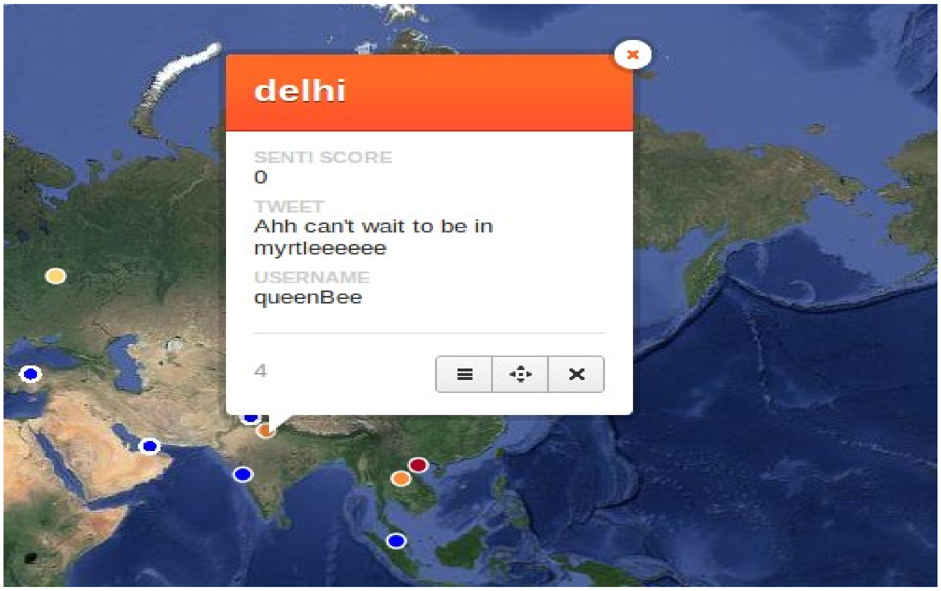
SENTI SCORE
5

TWEET
'Happy' by Pharrel Williams ! Too good

USERNAME
Rivera_gonzalez

3

(b): A happy tweet



delhi

SENTI SCORE
0

TWEET
Ahh can't wait to be in myrtleeeeee

USERNAME
queenBee

4(c): A Neutral tweet

Figure 4: CartoDB visualization

USPs

- ★ Extensive database of keywords for sentiments
- ★ Sentiment scores for emoticons ! :)
- ★ Location based ‘sentiment mapping’
- ★ Dynamic visualization

FUTURE PLAN OF EVENTS

- ➔ We have currently set up a stationary version of the database, the goal now is to populate the PostGres database dynamically with the upcoming tweets.
- ➔ Study the capacity of dynamism possible in visualizing the tweets
- ➔ Understand the time delay between the tweet occurrence and its representation on the world map (latency issues)[8]. We hope to reduce the latency as much as possible
- ➔ Finally visualize the data through CartoDB

REFERENCES

1. Raquel Meyer Alexander, James K. Gentry: Business Horizons Volume 57, Issue 2, March–April 2014, Pages 161–167 **Using social media to report financial results**
2. L. Palen, K. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald (2010). **"A vision for technology- mediated support for public participation & assistance in mass emergencies & disasters."** In Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference.
3. Sudha Verma, Sarah Vieweg, William J. Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, Kenneth M. Anderson, (2010) **"Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency"**, AAAI Publications, Fifth International AAAI Conference on Weblogs and Social Media
4. Soudip Roy Chowdhury, Sihem Amer-Yahia, Carlos Castillo Muhammad Imran Muhammad Rizwan Asghar, **"Tweet4act: Using Incident-Specific Profiles for Classifying Crisis-Related Messages"** (2013), The 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)
5. Axel Schulz and Petar Ristoski, (2013) **"The Car that Hit The Burning House: Understanding Small Scale Incident Related Information in Microblogs"**, AAAI Technical Report WS-13-0 10.
6. Sílvio S. Ribeiro, Jr., Clodoveu A. Davis, Jr., Diogo Rennó R. Oliveira, Wagner Meira, Tatiana S. Gonçalves Jr. Gisele L. Pappa Jr. (2012) **"Traffic observatory: a system to detect and locate traffic events and conditions using Twitter"**
7. Kalina Bontcheva a; Dominic Rout (2012), **"Making Sense of Social Media Streams through Semantics: a Survey."** Semantic Web 0 (0) 1 1 IOS Press, Semantic Web journal, (2012)
8. Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, Robert C. Miller, (2013) **"Processing and Visualizing the Data in Tweets"**
9. www.csc.ncsu.edu/faculty/healey/tweet_viz/
10. www.cyberemotions.eu/
11. www.eurekalert.org/pub_releases/2013-09/uosc-ta090313.php
12. Bontcheva, K., & Rout, D. (2012). **Making sense of social media streams through semantics: a survey**. *Semantic Web*.
13. Ghiassi, M., Skinner, J., & Zimbra, D. (2013). **Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network**. *Expert Systems with Applications: An International Journal*, *40*(16), 6266-6282.
14. Hienert, D., Wegener, D., & Paulheim, H. (2012). **Automatic classification and relationship extraction for multi-lingual and multi-granular events from wikipedia**. *Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, *902*, 1-10.
15. Go, A., Bhayani, R., & Huang, L. (2009). **Twitter sentiment classification using distant**

supervision. *CS224N Project Report, Stanford*, 1-12.

16. Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2012). **Processing and visualizing the data in tweets**. *ACM SIGMOD Record*, *40*(4), 21-27.

17. www.social-media-monitoring-review.toptenreviews.com/

18. Mahmud, J., Nichols, J., & Drews, C. (2012, May). **Where Is This Tweet From? Inferring Home Locations of Twitter Users**. In *ICWSM*.

19. Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011, May). **Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles**. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 237-246). ACM.

20. Zanzotto, F. M., (2011, July). **Linguistic redundancy in twitter**. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 659-669). Association for Computational Linguistics.

21. Wu, W., (2010, June). **Automatic generation of personalized annotation tags for twitter users**.

22. www.cartodb.com

23. https://github.com/CartoDB/cartodb