PES UNIVERSITY		**Document Ref.:**	
		Version No.:	
		Date:	09/04/2019
		Copy No.:	
Project Name:	Detection of Events and Emerging Themes from Social Media Streams		
Project Code:	PW19COP01		
Status:	**Draft / Current /Superseded**		
Document Type:	**Controlled / Uncontrolled**		

Detection of Events and Emerging Themes from Social Media Streams

Detecting events of national importance, disease outbreaks, and other emergencies and displaying the same on a user's Twitter feed.

Prepared By:		**Reviewed By:**	
Name	**Date**	**Name**	**Date**
Vaishnavi Rao	09/04/2019		
Varsha R.	09/04/2019	**Approved By:**	
		Name	**Date**

Distribution List

Project Representative(s)	**PESU Representative(s)**
1. Vaishnavi Rao 2. Varsha R.	1. Prof. C. O. Prakash

TABLE OF CONTENTS

References

[1] Catching the Long-Tail: Extracting Local News Events from Twitter; Puneet Agarwal, Rajgopal Vaithiyanathan, Saurabh Sharma and Gautam Shroff

[2] Topic Extraction from News Archive Using TF*PDF Algorithm; Khoo Khyou Bun Mitsuru Ishizuka

[3] A Survey of Techniques for Event Detection in Twitter; Farzindar Atefeh and Wael Khreich

[4] Sentiment-Based Event Detection in Twitter; Georgios Paltoglou

[5] Emerging Event Detection in Social Networks with Location Sensitivity; Unankard Sayan, Xue Li, and Mohamed A Sharaf

[6] Predicting Crowd Behaviour with Big Public Data; N. Kallus

[7] Event Detection in Social Media Data; Wenwen Dou, Xiaoyu Wang, William Ribarsky, Michelle Zhou

[8] Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development Ekta, Paahuni Khandelwal, Priya Bundela, Richa Dewan

Change History

This section describes the details of changes that have resulted in the current Test Plan document.

#	Date	Document Version No.	Change Description	Reason for Change
1.				
2.				
3.				

1.0 Introduction

The micro-blogging platform, Twitter, has opened people up to an entirely new world of information-gathering and news-sharing, including but not limited to detailing everyday events and activities of the users, the "Tweeple". Twitter, then, becomes an exhaustive source of data for analysis as well as the first place people are increasingly turning to for daily alerts or updates, in the face of news organisations and other traditional forms of media people looked to earlier. A major challenge facing event detection using social media streams is to extract useful real-world events from the mundane and polluted text (abbreviated words, spelling and grammatical errors and text written in mixed languages). [3] Event detection requires the automatic answering of what, when, where, and by whom. After reporting on the most recent efforts in the area, it is clear that no method addressed all of these questions. [5] Investigating how to model the social streams together with other data sources, like news streams to better detect and represent events was another complication faced earlier. [6] A considerable effort is still required to achieve efficient, scalable and reliable systems for event detection, summarization, tracking and association. [7]

1.1 Scope

The project will depend on the efficient functioning of Twitter minus any crashes on any given day to be able to conduct real time analysis smoothly in the long run. A proper, working internet connection is required that keeps the end user connected with their surroundings in order to access Twitter easily. Additional limitations include false positives generated by the model and disinformation provided by people on Twitter. Twitter users are assumed to be sensors, who make observations, which are their actual tweets to detect a target event. The tweets are given a time values and a location as well. The detection of these events and other emerging topics is restricted to the Indian subcontinent.

2.0 Test Strategies

1. Unit testing:

Every single element of our project needs to be tested. For this, we first prepare a test dataset with the relevant tweets belonging to all categories 0, 1, 2, 3, 4. We generate a label for each of these tweets using all the algorithms available.

We then display a few tweets belonging to each category. We manually analyze the tweets and finalize on the algorithm to be chosen.

2. Integration testing:

This is the phase where we need to merge both the front-end with the Python back-end. We will then feed tweets whose labels we know and generate labels and compare with the actual labels. This will also help us in generating an accuracy for better understanding of how the algorithm chosen performs.

This stage hasn't been implemented yet because the front-end is yet to be completed.

3. System Testing

There really isn't a system testing stage as after integration, the entire system is completely built and hence becomes redundant.

The testing then happens in these few steps:

- Test on validation data (train_test_split)
- Test on test dataset
- Manually evaluate performance
- Test on live stream Twitter data
- Manually evaluate performance
- Perform integration testing to evaluate the performance of the front-end and correct errors if any
- Beta test data by displaying options to various users via the front-end
- Add any user requirements and send for production

3.0 Performance Criteria

1 terminal for the user to look at comparison of different algorithms and their

performances and 1 for display of tweets via the UI

- 1 user can use the UI. No login is required.
- The entire day's tweets need to be processed. Could be above 500MB to 4GB
- System only caters to tweets generated in the Indian subcontinent
- 1.6 million tweets will be used for training in the ratio of 80:20 for train test split.

After validation, live tweets from twitter.com will be fetched for users.

4.0 Test Environment

The test environment is basically right now jupyter notebook. The accuracy and the top results from running the algorithm in the test dataset has given us these outputs. These outputs have been manually analyzed to conclude the performance of our algorithm. The hardware environment is nothing but a PC with jupyter notebook installed. This will work with both Ubuntu and Windows systems.

5.0 Roles and Responsibilities

Data mining: Gathering data suitable for testing. This testing data should not be labelled and must have equal number of tweets belonging to all 5 categories.

Database management: Here, we need to organize the data properly and store them

Pre-processing: The test data also has to be pre-processed

Model Evaluation: The test data after being pre-processed and converted to the right vector form is passed into the model to obtain results

Testing: we use different models, observe results and conclude the performance and draw justifications for the behaviour of the results.

6.0 Test Schedule

1. Test on validation data (train_test_split)
2. Test on test dataset
3. Manually evaluate performance
4. Test on live stream Twitter data
5. Manually evaluate performance
6. Perform integration testing to evaluate the performance of the front-end and correct errors if any
7. Beta test data by displaying options to various users via the front-end
8. Add any user requirements and send for production

7.0 Test Tools Used

The test tools used are:

Python3

Jupyter Notebook

Pandas

Scikit-learn

Keras

Numpy

8.0 Acceptance Criteria

The test data is said to have passed the acceptance criteria if we can manually verify the validity of the data it is predicting

The test data was run on the best algorithm we found, the logistic regression algorithm with fine tuned parameters and the results were good for 0, 1, 2 and 4 and were weak for 3

9.0 Test Case List

This section shall clearly define the test cases that are planned for testing. The following information shall be mentioned: -

Test Case Number	Test Case	Required Output
0	None of the above	0
1	Natural Disasters	1
2	Disease Outbreaks	2
3	Terror Attacks	3
4	Extreme Weather	4

10.0 Test Data

Just happened a terrible car crash

Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all

Heard about #earthquake is different cities, stay safe everyone.

Forest fire near La Ronge Sask. Canada

All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected

13,000 people receive #wildfires evacuation orders in California

Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school

#RockyFire Update => California Hwy. 20 closed in both directions due to Lake County fire - #CAfire #wildfires

#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Colorado Springs areas

Typhoon Soudelor kills 28 in China and Taiwan

We're shaking...It's an earthquake

11.0 Traceability Matrix

CRS/HLD/LLD Reference Section No. and Name	Test Case Number