Detecting Dengue Spreading in Sri Lanka based on News Articles

Group 17

E/13/183 – Nishara Kavindi

E/13/222 – Prabhashi Meddegoda

E/13/286 – Peshali Randika

Supervisor : Mr. D.S.Deegalla

Overview

- Background
- Related Works
- Problem Definition
- Design Justification
- System Design
- Methodology
- Implementation Justification
- Results and Evaluation
- Conclusion
- Future Directions
- Milestones
- Milestone Plan for the Second Phase

Background

Multiple approaches to monitor disease occurrences continuously.

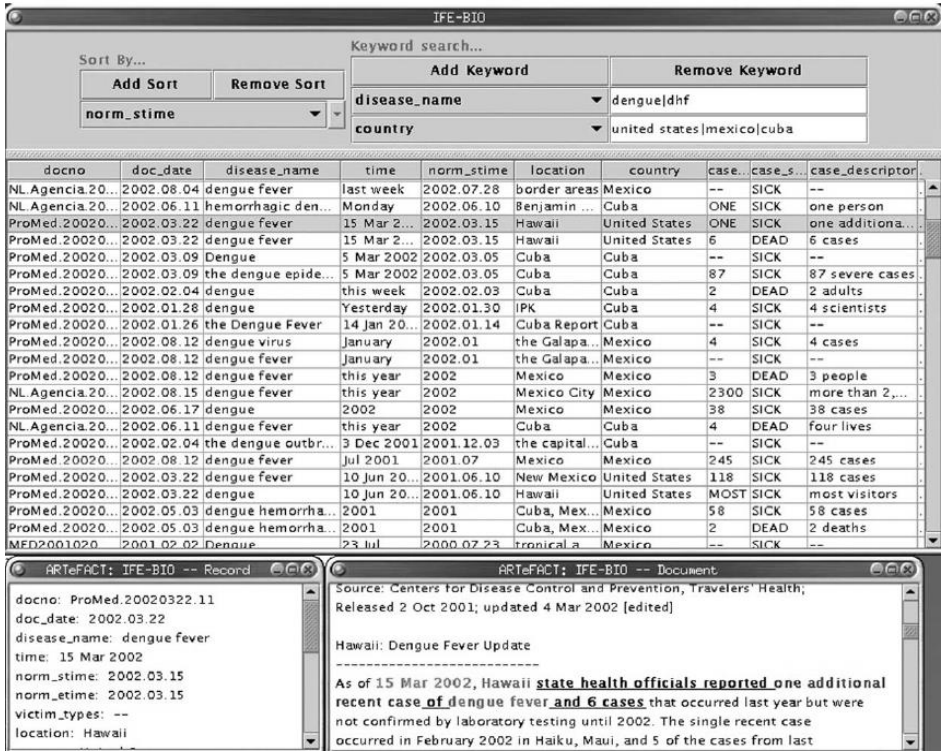
1. Indicator-based system
    - Traditional approach
    - Collect and analyze structured information reported by healthcare providers
2. Event-based system
    - Modern approach
    - Collect and analyze unstructured information reported by social media news and internet

The approach of monitoring diseases using data from online news articles is one of the modern approaches.

Related works

1. Information Extraction for enhanced access to disease outbreak reports

Authors : Ralph Grishman, Silja Huttunen, and Roman Yangarber
Computer Science Department, Courant Institute of Mathematical Sciences, New York University, New York, NY 10003-6806,USA



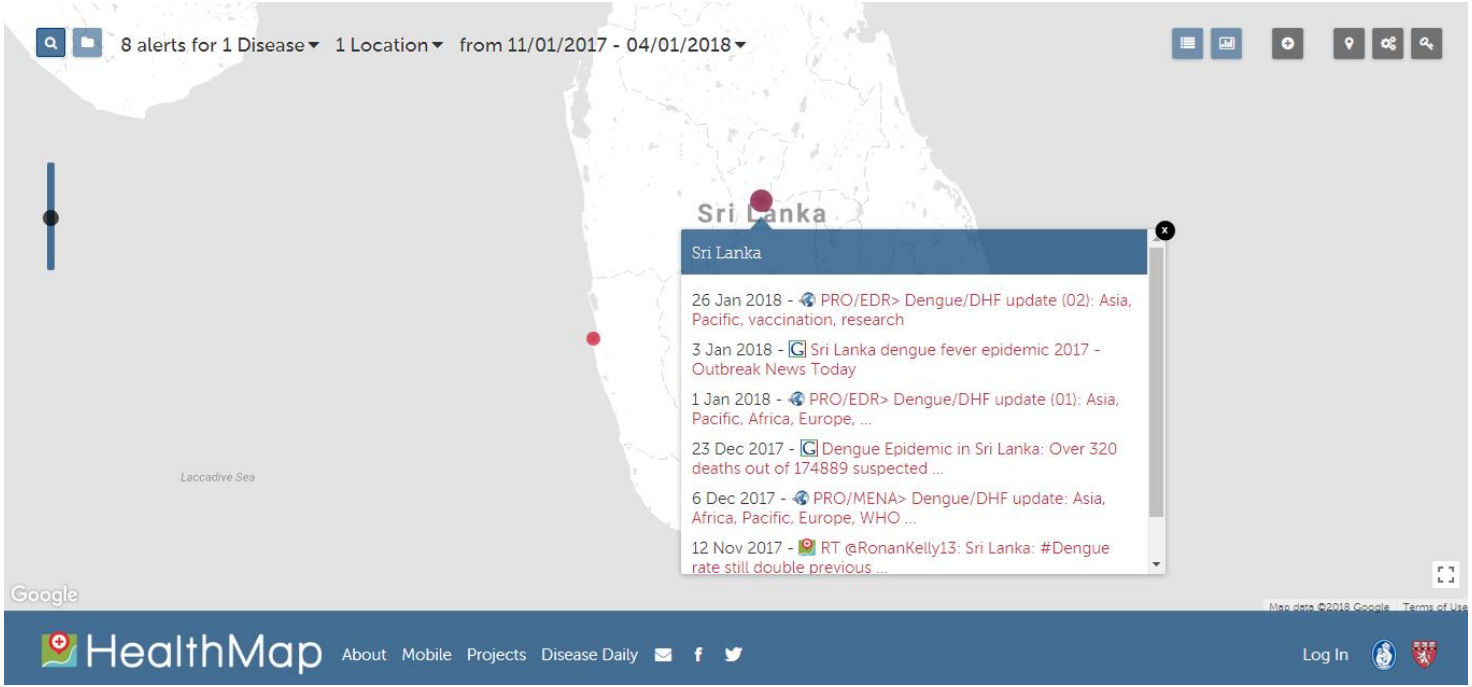
- This system gathers Web pages, extracts information about outbreaks, and presents the extracted information in a tabular form with links to the document.

2. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports

Authors : Clark C. Freifeld,Kenneth D. Mandl, Ben Y. Reis and John S. Brownstein

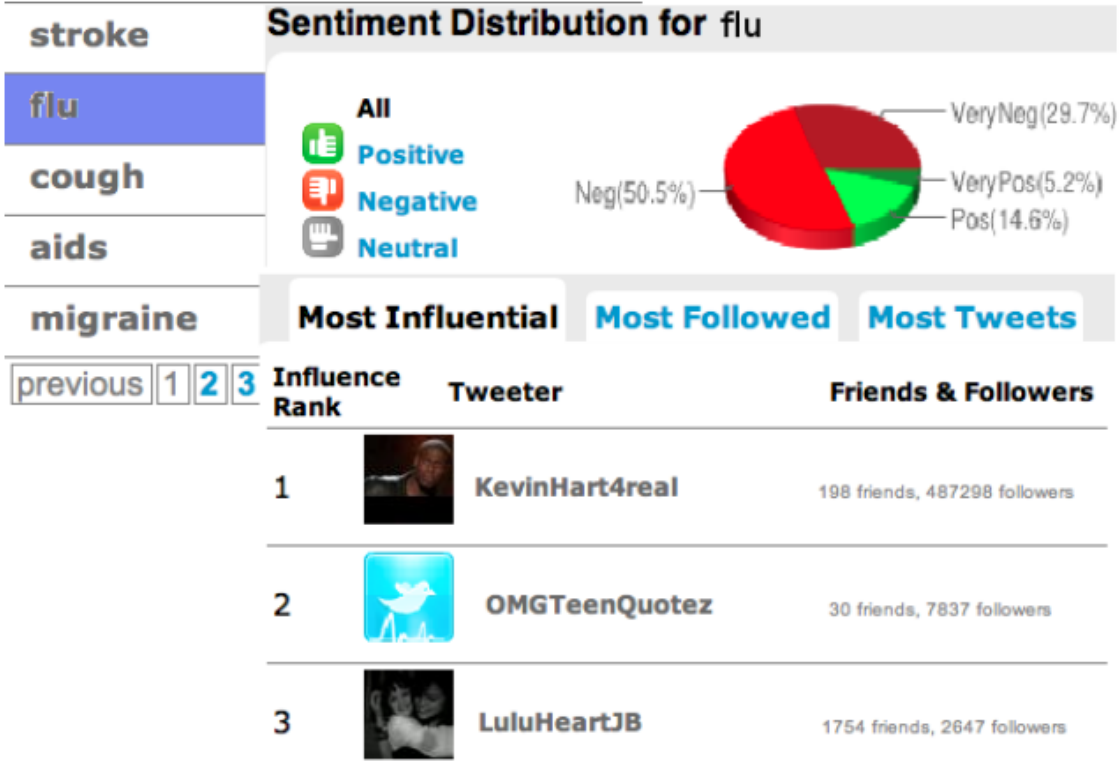
It shows an article list related to Dengue on the country map

No predictions, no spreading reports, gives only an abstract idea.



3. Detecting and Tracking Disease Outbreaks by Mining Social Media Data

Authors : Yusheng Xie, Zhengzhang Chen, Yu Cheng Kunpeng Zhang Kathy Lee Ankit Agrawal Wei-keng Liao
Alok Choudhary , Northwestern University, Evanston, IL USA



- Can not obtain specific details about Sri Lanka

4. Biocaster : Detecting public health rumors with a Web-based text mining system

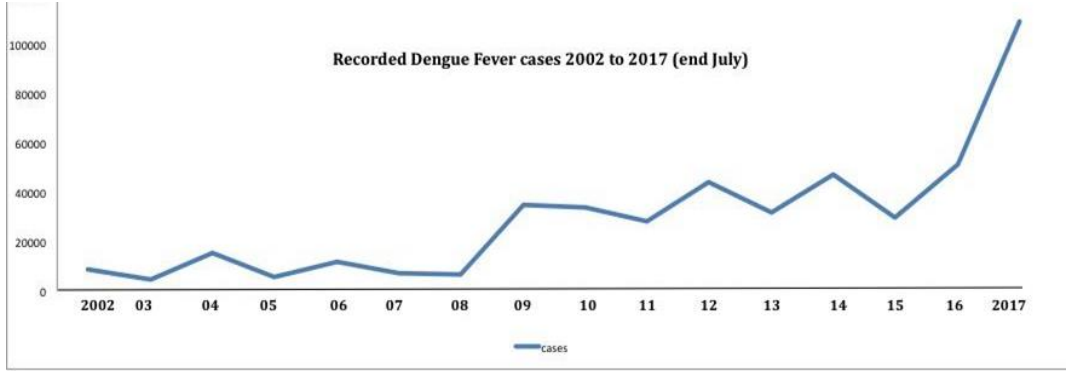
Authors : Nigel Collier,Son Doan,Ai Kawazoe,Reiko Matsuda Goodwin, Mike Conway,Yoshio Tateno,Quoc-Hung Ngo,Dinh Dien,Asanee Kawtrakul,Koichi Takeuchi,Mika Shigematsu,and Kiyosu Taniguchi

5. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health

Authors : Mykhalovskiy E, Weir L.,Department of Sociology, York University, Toronto, Ontario, Canada.

Problem Definition

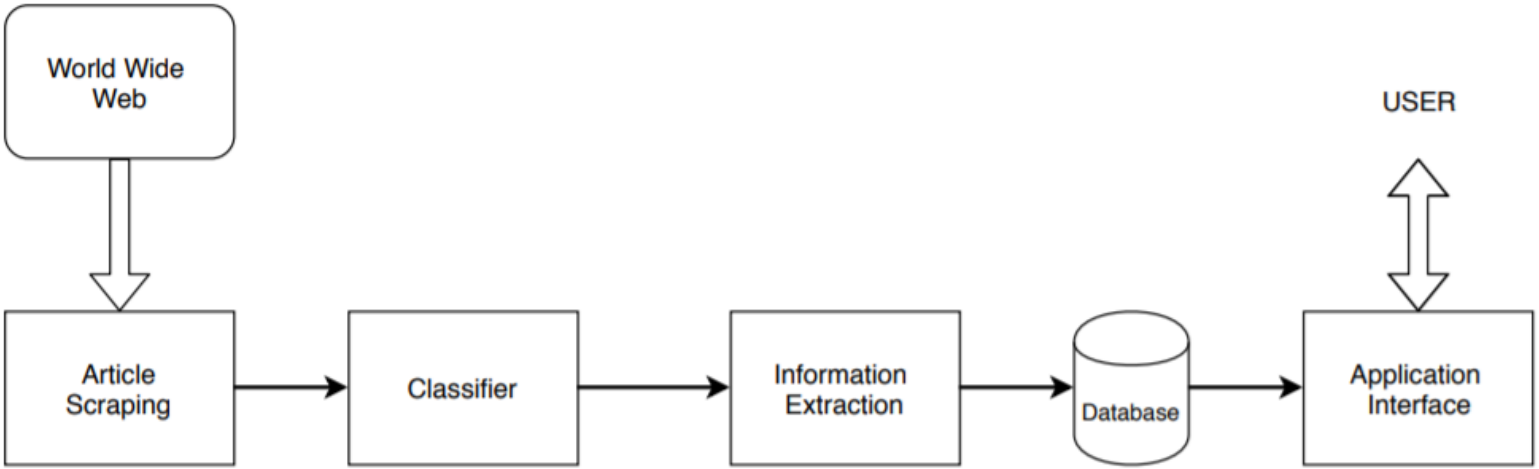
- ★ The existing systems for detecting dengue spreading in Sri Lanka are a paper-based system and a web-based system.
- ★ Dengue is still a deadly threat and a major problem in Sri Lanka
- ★ Important to find another approach for dengue surveillance, in order to give a better result
- ★ An Event-based disease monitoring system is not experimented yet in Sri Lanka.



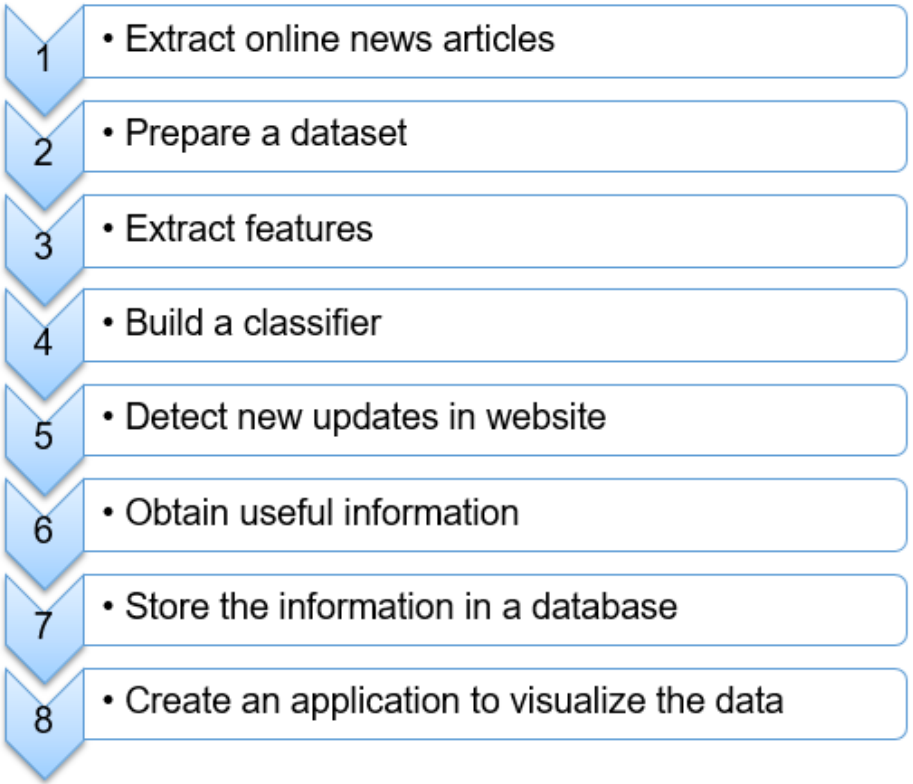
Design Justification

- This approach is experimented(and implemented) in other countries and they got effective results.
- Make use of the already available resources (news articles).

System Design



Methodology



Implementation Justification

Python language was chosen for implementation

- Provides many libraries supported to data mining
- Packages like NumPy, SciPy, and pandas produce good results for data analysis
- Scikit-learn provide many functionalities for machine learning tasks

Implementation Justification (Contd..)

- Data Collecting

Octoparse - modern visual web data extraction software used for extracting online news articles.

- ❖ Dengue
- ❖ Non-dengue

Octoparse :

- ➔ Facilitate to extract bulk information from website
- ➔ Extracted information can be exported in CSV format
- ➔ Other libraries - News-please, beautifulSoup ,newspaper (Unable to provide expe results)

Implementation Justification (Contd..)

- Preparing the dataset
  - ➢ Combine two data sets (Dengue and Non-dengue)
  - ➢ Labelled manually
  - ➢ Accurate labelling is important as the dataset is used for training.

- Extracting features from the articles
  - ➢ The text data should be converted into numeric values in order to apply data mining algorithms.
  - ➢ Bag of words model - Document classification method where the occurrence of each word is used as a feature for training a classifier

Implementation Justification (Contd..)

- Building a classifier
  - ➢ Accuracy was tested for multiple classification algorithms to find the best algorithm
  - ➢ 10-fold cross validation was used to measure the accuracy
  - ➢ Algorithms were tested using hold-out data set
  - ➢ To get a better accuracy,
    - ○ Tuning the vectorizer
      - ■ Stemming
      - ■ Removing stop words
    - ○ Tuning the parameters
      - ■ Ngram_range - Unigrams , bigrams
      - ■ Use_idf
      - ■ Smoothing parameter (alpha)
    - ○ Feature Selection

Implementation Justification (Contd..)

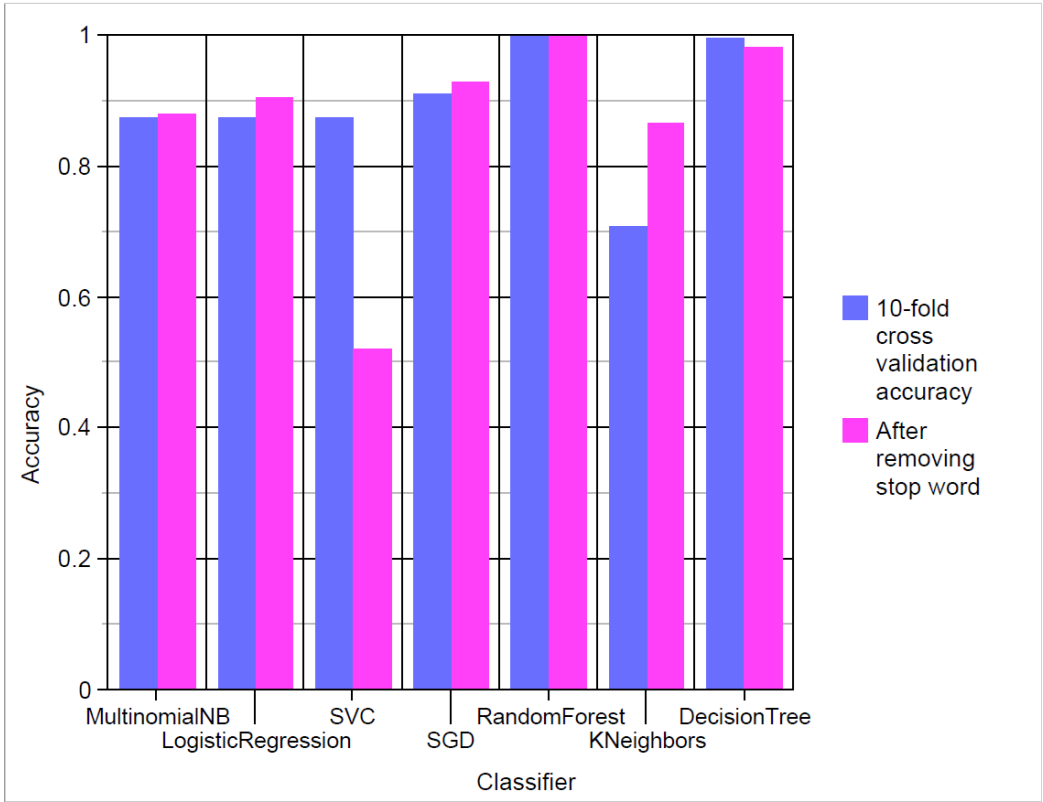
- Detecting new updates in the website
  - ➢ RSS (Rich site summary) feeds were used
  - ➢ Python script was developed to export new articles in CSV format

Results and Evaluation

- The behavior of the accuracy values of classifiers

Classification algorithm	Accuracy (10-fold cross validation)	Accuracy (After removing stop words)
MultinomialNB	0.873365785	0.879551139
LogisticRegression	0.873365785	0.90412573
SVC (Support Vector Machine)	0.873365785	0.518852739
SGD (Stochastic Gradient Descent)	0.91022799	0.927639109
RandomForest	0.997938144	0.997938144
KNeighbors	0.707141795	0.864202088
Decision Tree	0.99485588	0.982682708

Results and Evaluation(Contd..)



Results and Evaluation (Contd..)

- Results for the classifiers

Expected output for holdout data set ----->[0 0 0 0 0 1 1 1 1 1]

Classification Algorithm	Classification Result for theTest_dataset
MultinomialNB	[0 0 1 0 0 1 1 1 1 1]
LogisticRegression	[0 0 1 0 0 1 1 1 1 1]
SVC (Support Vector Machine)	[1 1 1 1 1 1 1 1 1 1]
SGD (Stochastic Gradient Descent)	[0 0 1 0 0 1 1 1 1 1]
RandomForestClassifier	[0 0 0 0 0 1 1 1 1 1]
KNeighborsClassifier	[0 0 1 0 0 1 1 1 1 1]
DecisionTreeClassifier	[0 0 0 0 0 1 1 1 1 1]

Conclusion

- According to the results, Random Forest and Decision Tree classifiers give the highest accuracy values
- They predict the classes of hold-out data set correctly
- Random Forest classifier gives a slightly higher accuracy
- Therefore Random Forest classifier can be used as the classifier in the system

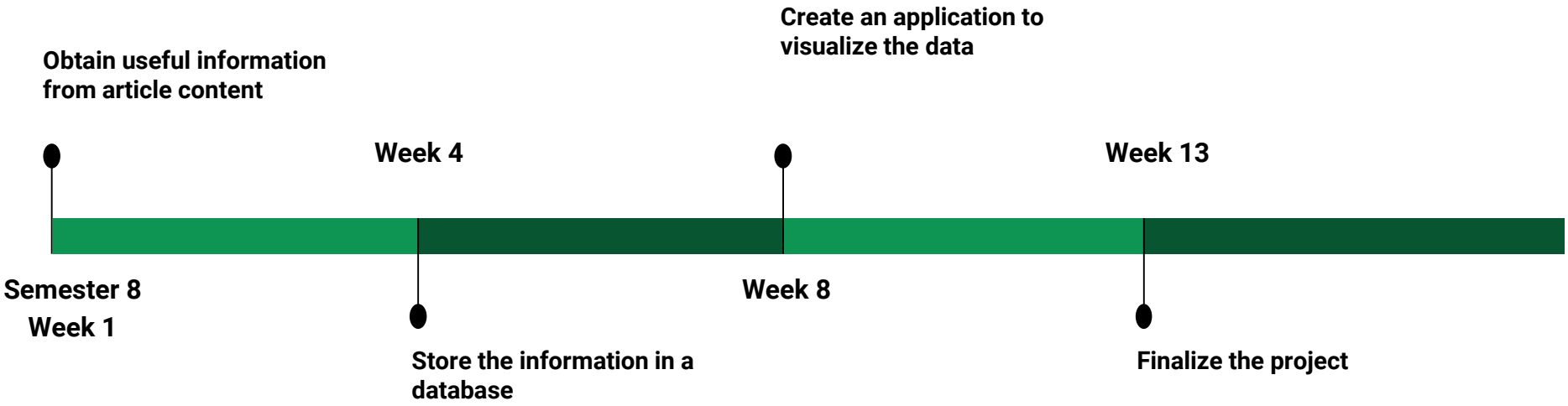
Future Directions

- Improving this system to monitor other diseases also
- Broadening the range of resources used.

Milestones

	Milestones	week 3	week 4	week 5	week 6	week 7	week 8	week 9	week 10	week 11	Semester 8
1	Collecting News articles about dengue in Sri Lanka from Web							Dead week	End Exam	Presentation	
2	Extract relevant information from collected articles										
3	Preparing a dataset										
4	Data preprocessing										
5	Determining a classification algorithm										
6	Obtain useful information										
7	Store the information in a database										
8	Visualizing the results										
9	Finalizing the project										

Milestone Plan for the Second Phase



Thank You !

Q & A

Results and Evaluation (Contd..)

- Results after feature selection for Random Forest Classifier

Feature Selection Methods for Random forest Classifier	Number of features(from 16327)	Train-test split Accuracy
Removing features with low variance	818	0.815709969789
Select from mode	8	0.993957703927
Tree based feature selection	2943	0.827794561934
Univariant selection	50	0.948640483384

Results and Evaluation (Contd..)

- Parameter tuning - results for Multinomial Naive Bayes

	No. of features used	**10 fold Cross Validation Accuracy**	**10 fold Cross Validation Accuracy after parameter tuning**
Using original features	16327	0.871774891774891 76	0.898082744702
After removing stop words (English)	16037	0.879875901875901 81	0.898082744702
After removing stop words (English) and using stemming (Snowball Stemmer)	10845	0.872815295815295 93	0.91019172553

Dialog Veta App

Veta App

Prevent Dengue the smart way with the Veta App powered by Dialog

The Veta App, powered by Dialog and supported by the Ministry of Health & and the National Dengue Control Unit, is a technology platform for communities to prevent the spread of Dengue.

By downloading the Veta App, you will be able to,

- Report suspected or confirmed Dengue cases & Dengue breeding sites
- Receive alerts on reported Dengue cases in the locations that you frequently visit
- Assist National Dengue Control Units to mobilize the resources effectively and execute preventive measures in the required areas
  AND
- Alert your neighborhood and take preventive measures against Dengue.

What makes us different ?

- Using enough/available sources of information (news articles)
- Focused on giving detailed information of dengue spreading in Sri Lanka
- This approach is to experiment an event-based surveillance system for Sri Lanka