

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

The screenshot shows the IFE-BIO application window. At the top, there's a 'Sort By...' dropdown set to 'norm_time' and a 'Keyword search...' section with 'Add Keyword' and 'Remove Keyword' buttons. The search criteria are 'disease_name: dengue|dhf' and 'country: united states|mexico|cuba'. Below this is a large table of search results with columns: docno, doc_date, disease_name, time, norm_time, location, country, case., case_s., and case_descriptor. The table lists various dengue fever reports from 2001 to 2002 across different locations like Mexico, Cuba, and Hawaii. At the bottom, there are two panes. The left pane shows the details for document 'ProMed.20020322.11', including its date, disease name, time, and location. The right pane shows the full text of the document, which is a 'Hawaii: Dengue Fever Update' dated March 15, 2002, reporting an additional case and six more cases from the previous year.

docno	doc_date	disease_name	time	norm_time	location	country	case.	case_s.	case_descriptor
NL Agencia.20...	2002.08.04	dengue fever	last week	2002.07.28	border areas	Mexico	--	SICK	--
NL Agencia.20...	2002.06.11	hemorrhagic den...	Monday	2002.06.10	Benjamin ...	Cuba	ONE	SICK	one person
ProMed.20020...	2002.03.22	dengue fever	15 Mar 2...	2002.03.15	Hawaii	United States	ONE	SICK	one additiona...
ProMed.20020...	2002.03.22	dengue fever	15 Mar 2...	2002.03.15	Hawaii	United States	6	DEAD	6 cases
ProMed.20020...	2002.03.09	Dengue	5 Mar 2002	2002.03.05	Cuba	Cuba	--	SICK	--
ProMed.20020...	2002.03.09	the dengue epide...	5 Mar 2002	2002.03.05	Cuba	Cuba	87	SICK	87 severe cases
ProMed.20020...	2002.02.04	dengue	this week	2002.02.03	Cuba	Cuba	2	DEAD	2 adults
ProMed.20020...	2002.01.28	dengue	Yesterday	2002.01.30	IPK	Cuba	4	SICK	4 scientists
ProMed.20020...	2002.01.26	the Dengue Fever	14 Jan 20...	2002.01.14	Cuba Report	Cuba	--	SICK	--
ProMed.20020...	2002.08.12	dengue virus	January	2002.01	the Galapa...	Mexico	4	SICK	4 cases
ProMed.20020...	2002.08.12	dengue fever	January	2002.01	the Galapa...	Mexico	--	SICK	--
ProMed.20020...	2002.08.12	dengue fever	this year	2002	Mexico	Mexico	3	DEAD	3 people
NL Agencia.20...	2002.08.15	dengue fever	this year	2002	Mexico City	Mexico	2300	SICK	more than 2,...
ProMed.20020...	2002.06.17	dengue	2002	2002	Mexico	Mexico	38	SICK	38 cases
NL Agencia.20...	2002.06.11	dengue fever	this year	2002	Cuba	Cuba	4	DEAD	four lives
ProMed.20020...	2002.02.04	the dengue outbr...	3 Dec 2001	2001.12.03	the capital...	Cuba	--	SICK	--
ProMed.20020...	2002.08.12	dengue fever	Jul 2001	2001.07	Mexico	Mexico	245	SICK	245 cases
ProMed.20020...	2002.03.22	dengue fever	10 Jun 20...	2001.06.10	New Mexico	United States	118	SICK	118 cases
ProMed.20020...	2002.03.22	dengue	10 Jun 20...	2001.06.10	Hawaii	United States	MOST	SICK	most visitors
ProMed.20020...	2002.05.03	dengue hemorrha...	2001	2001	Cuba, Mex...	Mexico	58	SICK	58 cases
ProMed.20020...	2002.05.03	dengue hemorrha...	2001	2001	Cuba, Mex...	Mexico	2	DEAD	2 deaths
MED2001020...	2001.02.02	Dengue	23 Jul	2000.07.23	tropical a...	Mexico	--	SICK	--

ARTEFACT: IFE-BIO -- Record

docno: ProMed.20020322.11
doc_date: 2002.03.22
disease_name: dengue fever
time: 15 Mar 2002
norm_time: 2002.03.15
norm_etim: 2002.03.15
victim_types: --
location: Hawaii

ARTEFACT: IFE-BIO -- Document

Source: Centers for Disease Control and Prevention, Travelers' Health;
Released 2 Oct 2001; updated 4 Mar 2002 [edited]

Hawaii: Dengue Fever Update

As of 15 Mar 2002, Hawaii state health officials reported one additional recent case of dengue fever and 6 cases that occurred last year but were not confirmed by laboratory testing until 2002. The single recent case occurred in February 2002 in Haiku, Maui, and 5 of the cases from last

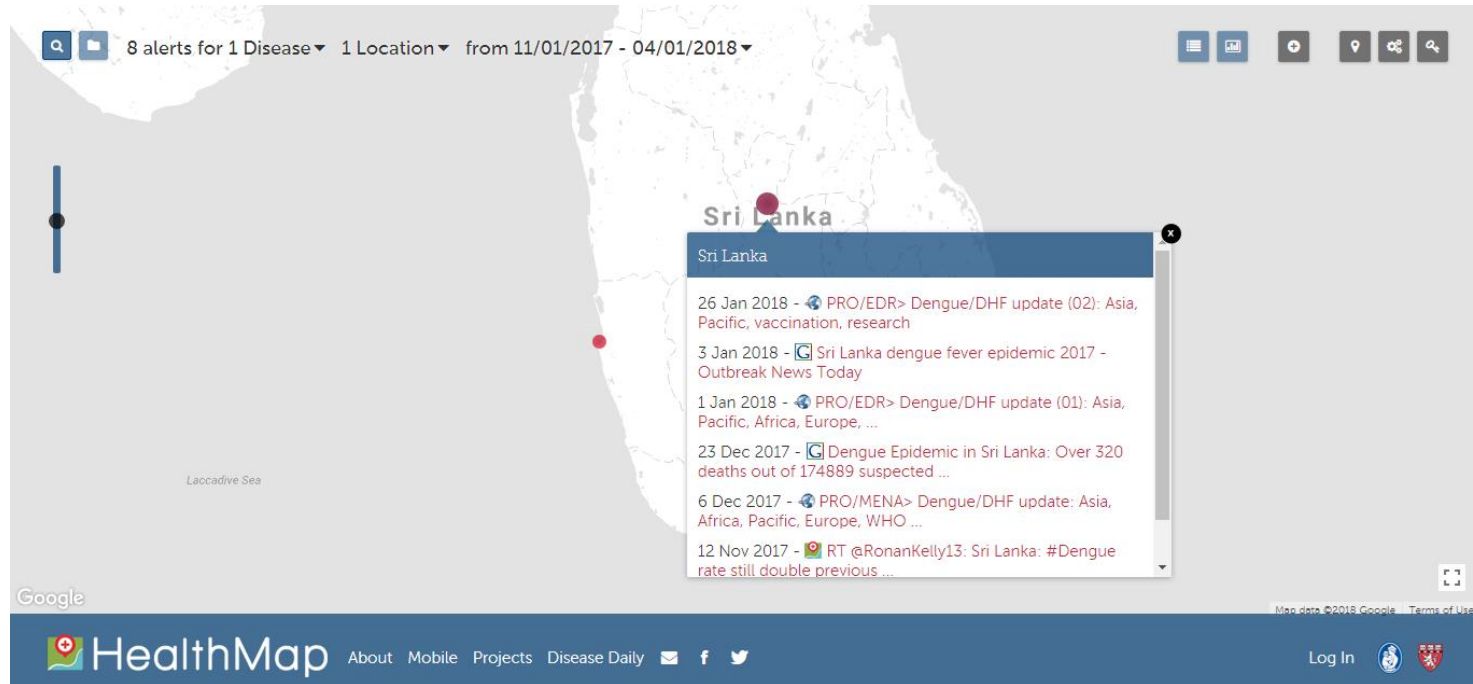
- This system gathers Web pages, extracts information about outbreaks, and presents the extracted information in a tabular form with links to the document.

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

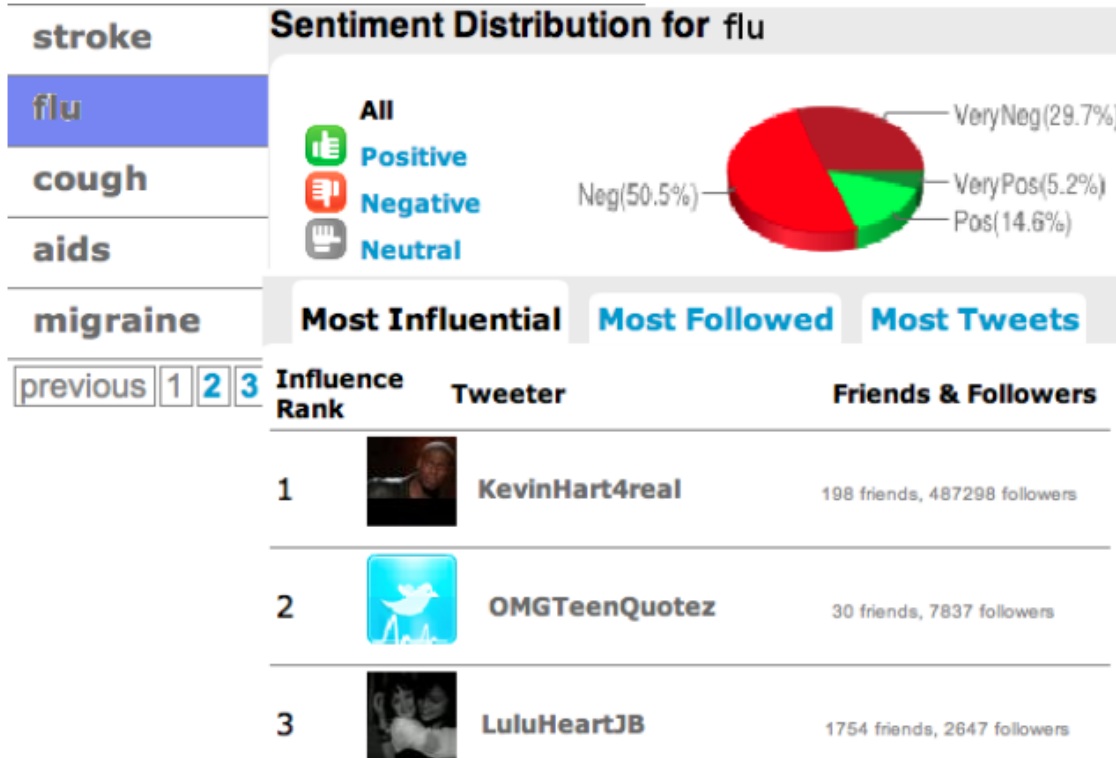
It shows an article list related to Dengue on the country map

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

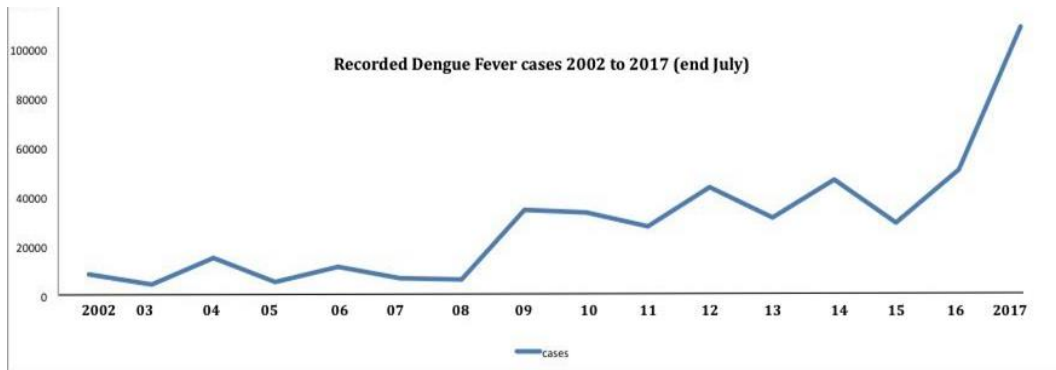
Authors : Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi

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

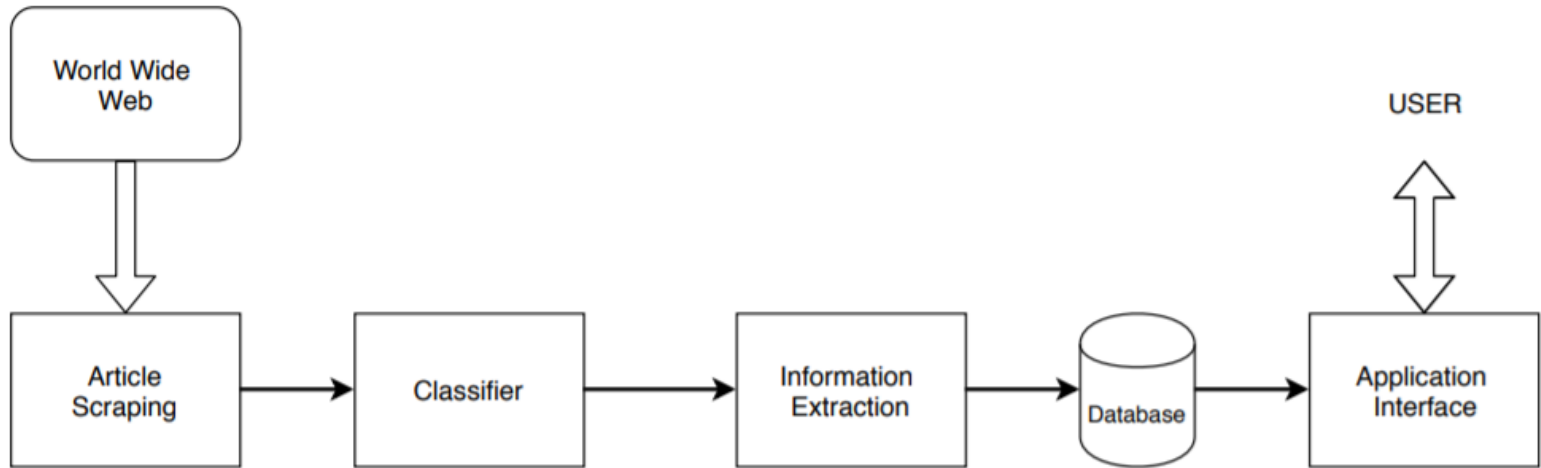
- ★ The existing systems for detecting dengue spreading in Sri Lanka are a paper-based system and a web-based system.
- ★ Dengue is still a deadly threat and a major problem in Sri Lanka
- ★ Important to find another approach for dengue surveillance, in order to give a better result
- ★ An Event-based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



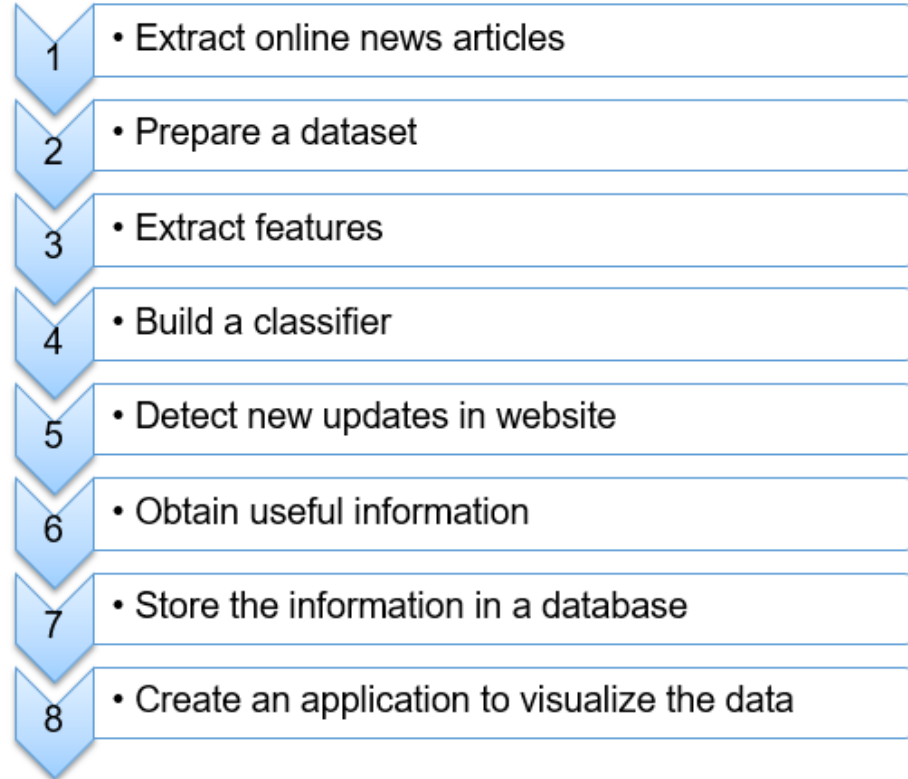
Design Justification

- 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 expected results)

Octoparse



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

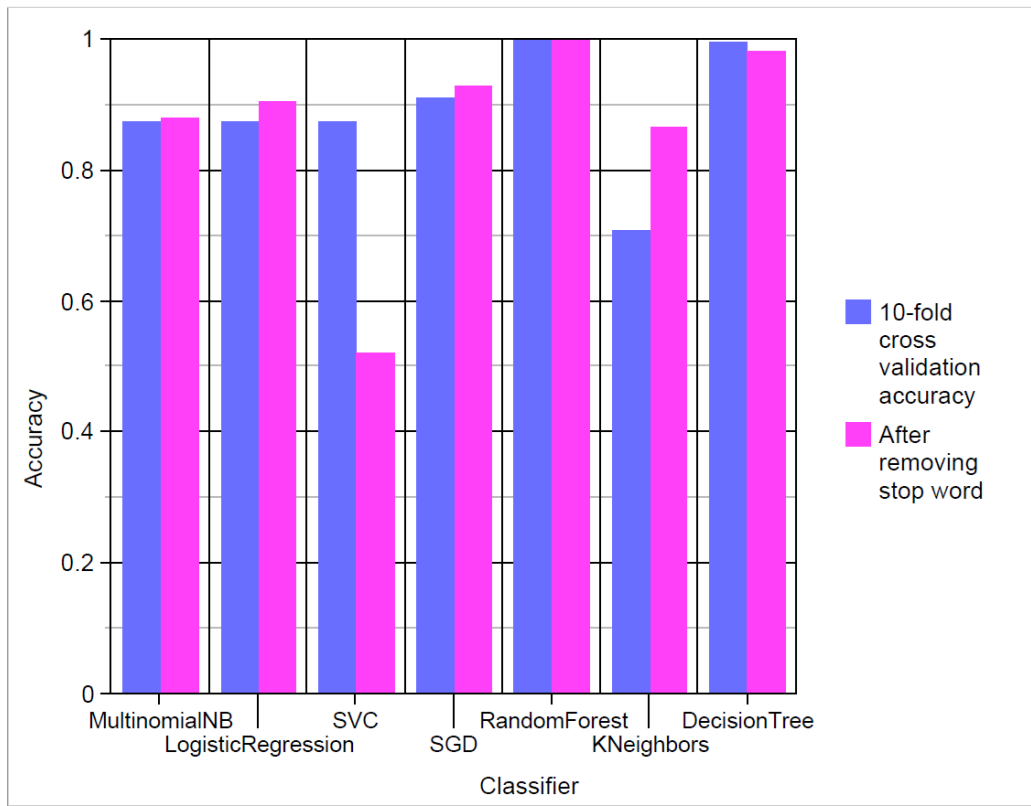


Results and Evaluation

- 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 the Test_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

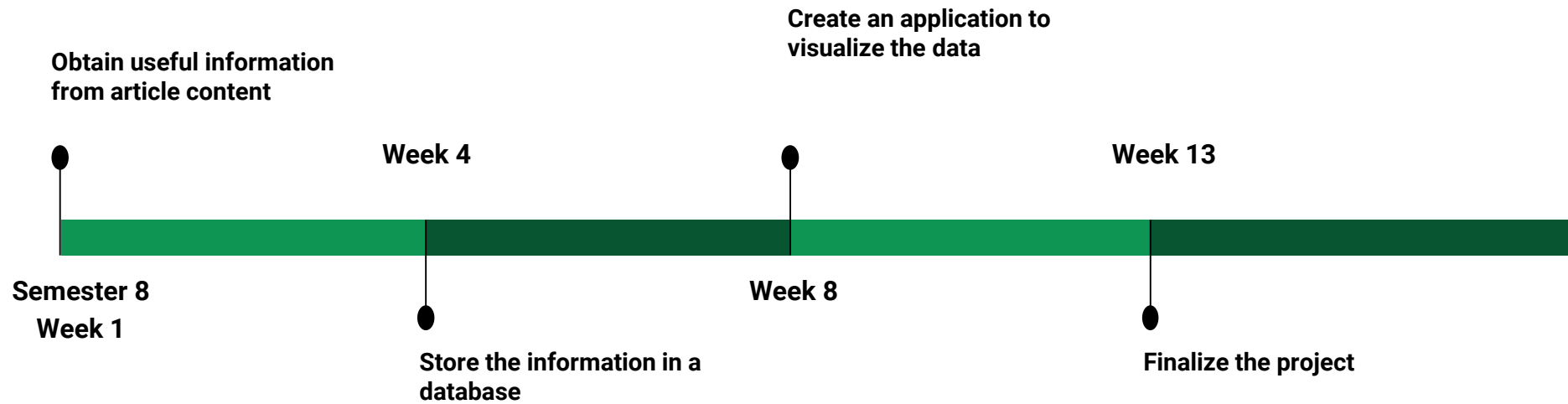
- Improving this system to monitor other diseases also
- Broadening the range of resources used.



Milestones

[illegible]

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	0.87177489177489176	0.898082744702
After removing stop words (English)	16037	0.87987590187590181	0.898082744702
After removing stop words (English) and using stemming (Snowball Stemmer)	10845	0.87281529581529593	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