Visualization Tool for Islamic Radical and Counter Radical Movements and their online followers in South East Asia

by

Sravan Kumar Garipalli

A Thesis Presented in Partial Fulfillment of the Requirements for the Degree Master of Science

Approved April 2015 by the Graduate Supervisory Committee:

Hasan Davulcu, Chair
Paulo Shakarian
Ihan Hsiao

ARIZONA STATE UNIVERSITY
May 2015
ABSTRACT

With the advent of social media and micro-blogging sites, people have become active in sharing their thoughts, opinions, ideologies and furthermore enforcing them on others. Users have become the source for the production and dissemination of real time information. The content posted by the users can be used to understand them and track their behavior. Using this content of the user, data analysis can be performed to understand their social ideology and affinity towards Radical and Counter-Radical Movements. During the process of expressing their opinions people use hashtags in their messages in Twitter. These hashtags are a rich source of information in understanding the content based relationship between the online users apart from the existing context based follower and friend relationship.

An intelligent visual dash-board system is necessary which can track the activities of the users and diffusion of the online social movements, identify the hot-spots in the users’ network, show the geographic footprint of the users and to understand the socio-cultural, economic and political drivers for the relationship among different groups of the users.
To all who supported me
ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my Professor Dr. Hasan Davulcu, for his support, without which my thesis would not have been possible. It is because of his guidance and support that I was able to complete my research. I would also like to thank Professors Paulo Shakarian and Sharon Hsiao for their support. I would also like to thank Ph.D student Nyunsu Kim for his support and guidance. I am very grateful for the unconditional support of my roommates and friends. Finally, I would like to thank my parents, for their motivation and encouragement to pursue this degree.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
</tbody>
</table>

## CHAPTER

1. **INTRODUCTION** .................................................. 1

1.1 Methodology ................................................... 1

1.2 Document Outline .............................................. 2

2. **BACKGROUND** .................................................... 3

2.1 Twitter ........................................................ 3

2.1.1 Hashtags ................................................ 3

2.1.2 User-User Relationship ................................... 3

2.1.3 Geographical Information in Tweets ....................... 4

2.2 Facebook Graph API ........................................... 4

2.3 HTML Parser Library-jsoup ..................................... 5

2.4 Data Indexing-Apache Solr ...................................... 5

2.5 JavaScript Visualization Library-d3.js ............................ 5

2.5.1 Chord Diagram .......................................... 6

2.5.2 Force Directed Graph .................................... 6

2.6 Visualizations from Google API .............................. 6

2.6.1 Google Annotated Timeline ............................... 7

2.6.2 Google Map Chart ....................................... 7

3. **RELATED WORK** .................................................. 8

4. **SYSTEM ARCHITECTURE** ......................................... 9

4.1 Organization Data Collection ................................... 10

4.2 Data Cleaning .................................................. 11
<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3  Data Analysis and Model Generation</td>
<td>11</td>
</tr>
<tr>
<td>4.4  Social Media Crawling and Analysis</td>
<td>15</td>
</tr>
<tr>
<td>4.5  Breaking Hashtag Analysis</td>
<td>16</td>
</tr>
<tr>
<td>5   VISUALIZATION TOOL AND COMPONENTS</td>
<td>17</td>
</tr>
<tr>
<td>5.1  Timeline-Hashtags Widget</td>
<td>17</td>
</tr>
<tr>
<td>5.2  Chord Diagram</td>
<td>20</td>
</tr>
<tr>
<td>5.3  Heat Map Widget</td>
<td>21</td>
</tr>
<tr>
<td>5.4  User-User Network</td>
<td>21</td>
</tr>
<tr>
<td>5.5  URL Information Widget</td>
<td>22</td>
</tr>
<tr>
<td>5.6  Tags-Users-Keywords Widget</td>
<td>23</td>
</tr>
<tr>
<td>6   EXPERIMENTAL RESULTS</td>
<td>24</td>
</tr>
<tr>
<td>6.1  Results for Malaysia Data sets</td>
<td>24</td>
</tr>
<tr>
<td>6.1.1 First Level of Classification</td>
<td>25</td>
</tr>
<tr>
<td>6.1.2 Second Level of Classification</td>
<td>26</td>
</tr>
<tr>
<td>6.1.3 Organization Level Classification</td>
<td>27</td>
</tr>
<tr>
<td>6.2  Results for UK Data sets</td>
<td>27</td>
</tr>
<tr>
<td>6.3  Results for Indonesia Data sets</td>
<td>28</td>
</tr>
<tr>
<td>7   SCENARIOS</td>
<td>32</td>
</tr>
<tr>
<td>7.1  Polarization between different Groups</td>
<td>32</td>
</tr>
<tr>
<td>7.2  Breaking Hashtags and Extremist Concepts</td>
<td>33</td>
</tr>
<tr>
<td>8   SUMMARY</td>
<td>34</td>
</tr>
<tr>
<td>9   FUTURE WORK</td>
<td>35</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>36</td>
</tr>
<tr>
<td>Table</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.1 Organizations Data-Malaysia</td>
<td>25</td>
</tr>
<tr>
<td>6.2 Results of Linear Regression-Malaysia</td>
<td>26</td>
</tr>
<tr>
<td>6.3 Accuracy of Classification Malaysia</td>
<td>28</td>
</tr>
<tr>
<td>6.4 Organizations Data-UK</td>
<td>28</td>
</tr>
<tr>
<td>6.5 Accuracy of Classification UK</td>
<td>29</td>
</tr>
<tr>
<td>6.6 Comparison of Accuracies for UK dataset using different Classifiers</td>
<td>31</td>
</tr>
<tr>
<td>6.7 Location Information of the Tweets UK dataset</td>
<td>31</td>
</tr>
<tr>
<td>6.8 Organizations Data-Indonesia</td>
<td>31</td>
</tr>
<tr>
<td>6.9 Accuracy of Classification Indonesia</td>
<td>31</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 The Overall System Architecture</td>
<td>9</td>
</tr>
<tr>
<td>4.2 Process of Collection of Organization Data</td>
<td>10</td>
</tr>
<tr>
<td>4.3 Linear Regression</td>
<td>12</td>
</tr>
<tr>
<td>4.4 Logistic Regression</td>
<td>12</td>
</tr>
<tr>
<td>4.5 Example showing the Data Analysis on A and y Matrices</td>
<td>13</td>
</tr>
<tr>
<td>4.6 Multi Class Linear Regression</td>
<td>14</td>
</tr>
<tr>
<td>4.7 Multi Class Logistic Regression</td>
<td>14</td>
</tr>
<tr>
<td>4.8 Example showing the Data Analysis on A and Y Matrices</td>
<td>14</td>
</tr>
<tr>
<td>5.1 Figure showing the Timeline, Chord Diagram and Heat Map Widgets</td>
<td>18</td>
</tr>
<tr>
<td>5.2 Figure showing the User-User Network, URL Information and Tags-Users-Keywords Widgets</td>
<td>19</td>
</tr>
<tr>
<td>5.3 Figure showing distributions of the Selected Hashtags in the Timeline</td>
<td>20</td>
</tr>
<tr>
<td>5.4 Figure showing distributions of the Radical and Counter-Radical Users</td>
<td>21</td>
</tr>
<tr>
<td>5.5 Figure showing an area selected using the Markers in Heat Map Widget</td>
<td>22</td>
</tr>
<tr>
<td>6.1 Figure showing the Hierarchy of the Users Classification in Malaysia</td>
<td>24</td>
</tr>
<tr>
<td>6.2 Figure showing the Hierarchy of the Users Classification in UK</td>
<td>29</td>
</tr>
<tr>
<td>6.3 Figure showing the Hierarchy of the Users Classification in Indonesia</td>
<td>30</td>
</tr>
<tr>
<td>7.1 Figure showing the Polarization between two different groups in UK</td>
<td>32</td>
</tr>
<tr>
<td>7.2 Figure showing User of a Radical Group and his Twitter page</td>
<td>33</td>
</tr>
</tbody>
</table>
Chapter 1

INTRODUCTION

With the advancements in science and technology there is a surge in the concept of online social networks. These social networks form various communities of millions of users in constant evolution. With the increased access to Internet, the online social network communication has taken new change. With this change, the way people are sharing, expressing their information resulting in the profound transformation on the society. The resultant data records from these online social media can be analyzed and studied to understand the users’ behavior.

Twitter has become one of the prominent online micro-blogging site where people express their opinions, messages in the form of short-messages called tweets. Users also add hashtags in their tweets. These hashtags are rich source of information in understanding the contents of the tweets. Twitter has about 500 million registered users in which about 280 million are active users. The daily volume of tweets in Twitter is about 400 million.

1.1 Methodology

The work proposed in this thesis is developed using the typical Software Development Life Cycle (SDLC) Model in incremental and evolutionary approach. The entire project is divided into tasks and sub-tasks and each one is carried out using the phases of SDLC which includes Requirements, Analysis, Design, Development and Deployment. The outputs and results in each sub-task are aggregated and used in the next sub-tasks and the final Visualization tool is built at the end.
1.2 Document Outline

The rest of the document is organized as explained below. Chapter 2 gives the overview of background information related to the project.

Chapter 3 gives the related work in the area of visualization tools with respect to the different Extremist groups and their associations.

Chapter 4 explains the system architecture and various components of the system and their implementations.

Chapter 5 deals with the Visualization tool and its components.
Chapter 6 gives the experimental results of the project on different data sets.
Chapter 7 captures the scenarios which are tracked using the Visualization tool.
Chapter 8 encompasses the summary of the thesis.
Chapter 9 gives the directions to future research activities that can be carried out.
2.1 Twitter

Twitter is a popular social media and micro-blogging site. It provides the user to post 140 character limited messages called tweets. Twitter provides a streaming API[11] which can be used to access the global stream of tweet data. We can also access the required tweet information using filters.

2.1.1 Hashtags

The hashtag is a keyword in the tweet with a prefix symbol # which is generally used to categorize the message or express some opinion. A tweet can have many hashtags associated with it and also the hashtag can be present anywhere in the tweet message. These hashtags can be used for mass media broadcast, event promotions, sentiment analysis, consumer complaints etc.

2.1.2 User-User Relationship

User-User relationship can be established in twitter in two ways. First, the Context based relationship and Content based relationship.

Consider two twitter users A and B with twitter screen names @a and @b respectively.

Context based relationship is based on followers and friends. If user A subscribes to the tweets of user B, then user A is a follower of user B. If there is two directional
follow relationship i.e. user A is follower of user B and user B is follower of A, then users A and B are friends.

Content based relationship is based on the user-mentions and retweets. If user A mentions user B (using @b) in his/her tweets then it is called as user-mention. If user A shares the tweet posted by user B to all his/her followers then it is called as retweet (usually the tweet message is prefixed with RT).

2.1.3 Geographical Information in Tweets

Location information is added to the tweets to give the location context to the tweet message. This information can be given using general location labels (like a city name) and also more specific location (specific landmarks, business etc) labels. When twitter is used on mobile phone platform more precise location information is added to the tweets in the form of latitude and longitude co-ordinates of the place you are tweeting from.

Apart from location information in the tweets, the user profile and description also has the location information.

2.2 Facebook Graph API

Graph API[7] is part of high level Facebook Platform component. It is used to read data from Facebook and write data to Facebook. The social graph of Facebook is presented in the form of nodes (user/page/comment), edges (connection between nodes like page’s photos or photo’s comments) and fields (information about nodes like name of page). Every node is represented by a unique ID, the information about particular node can be accessed using this unique ID.
2.3 HTML Parser Library-jsoup

jsoup\[13\] is a Java library developed as an open-source project, which is used for extraction and manipulation of HTML documents. It provides functionalities to scrape and parse the HTML documents, traverse and extract the data from HTML DOM traversals. Individual HTML elements like div, span, para etc. can be filtered from the entire HTML document. It also provides functions to get the text contents from these filtered elements.

2.4 Data Indexing-Apache Solr

Apache Solr\[12\] is a standalone search server where data in any format like JSON, XML, CSV, binary documents etc. can be indexed into it and also the indexed data can be queried over the HTTP.

It provides the features to handle large data set storage and effective retrieval. It provides a web administrative portal for easy control and configuration of accessing different Solr instances. It also provides full text search and matching capabilities using phrases, wildcards, joins and grouping of data.

2.5 JavaScript Visualization Library-d3.js

D3.js\[3\] is a popular JavaScript library used for producing interactive data visualizations. It provides hundreds of visualization diagrams which include Bubble Charts, Chord Diagrams, Force Directed Graphs, Treemaps, Collapsible Force Graphs and many others. It provides different functions for transitions, changing colors, doing Math, changing geometry of objects, changing Layouts etc.
2.5.1 Chord Diagram

The Chord Diagram[1] of d3.js is used to visualize the relationships between different entities. It consists of two charts, the chords and the arcs. The arc signifies the entity, the chord signifies the relationship between the two entities it is connected with (source arc and target arc).

2.5.2 Force Directed Graph

The Force Directed layout[2] of d3.js provides the ability to position the nodes and edges of a graph with minimum edge crossings. There are two types of forces involved in this layout first being the attractive force based on Hooke’s Law and second being the repulsive force based on Couloumb’s Law. The attraction force exists between the edges and the repulsive force exists between the nodes. This layout can be used to view the nodes which are similar to each other in form of a single cluster and dissimilar nodes are viewed as different clusters.

2.6 Visualizations from Google API

Google Visualization API [9] provides various types of visualization charts. These include Annotation Charts, Bar Charts, Bubble Charts, Calendar Charts, Maps, Timelines and many more.

The API provides:

- Functionalities to customize the charts
- Interactivity with the charts
- Different methods of loading the data to the charts
2.6.1 Google Annotated Timeline

The Google Annotated Timeline is an interactive time series chart line with optional annotations. The Timeline provides different configuration options to draw, redraw, color, configure date formats, display annotations etc. Timeline provides different methods to draw, to get selection of the range, to set the visible range. Timeline also provides event handlers to handle the events of range change, ready and selection.

The options to zoom in or zoom out of the timeline can be done using the range selection present in the bottom of the timeline.

2.6.2 Google Map Chart

The Google Map chart presents a map using the Google Maps API. The data displayed on the map can be viewed using the pairs of latitude-longitudes or using string address.

The Map chart provides configuration options for zoom level, map type(normal/terrain/satellite). It also provides option to select a particular area of the map, which is formed by the connected lines using the markers.
Chapter 3

RELATED WORK

Many studies are done to understand the social network of users and the relationship between the users using concepts in Statistics and Visualizations. Peter and Ben developed SocialAction[18] which is used for exploratory data analysis where they integrated both statistics and visualizations. It enabled the healthcare consultants, political researchers, counter terrorism researchers to visualize the network of users and to detect the important individuals, their relationships and different clusters of users. Some use cases they presented were understanding Global Jihad Terrorist Network and understanding Voting Patterns Among United States Senators.

LookingGlass [14], an intelligent visual platform is developed to understand and track the online social movements based on the users’ tweet activity. These social movements include organizations of Radical and Counter-Radical. It connects the relationship between the socio-cultural, political and economic drivers which lead to the radicalism and extremist ideas.

Studies [4], [16] are done to understand the breaking out or bursting of hashtags and various network and content measures responsible for the break out of a hashtag. These systems predict the break out of a hashtag.

Current thesis work integrates the concept of breaking hashtags from [4] into LookingGlass[14] to understand and track the social movements for the geographic regions Malaysia, UK, Indonesia their topics, perspectives and rhetoric emanating from the social and political disruptions.
This chapter gives the complete overview of the system architecture of the project and explains the specific components and different tasks carried out.

The major tasks in the system are as follows:

- Data Collection and Modeling from Organization data
- Data Collection from Social Media
- Indexing the data
- Visualization Tool Development

![Figure 4.1: The Overall System Architecture](image-url)

The list of organizations for particular geographical region are identified first and then the organizations are characterized as Radical-Islamist or Counter-Radical Islamist using scaling of different dimensions like Epistemology, Religious Diversity
Tolerance, Change Orientation, Violence, Violence Ideology, and Violence Engagement based on inputs from social scientists and domain experts[15].

4.1 Organization Data Collection

The web articles of the official websites, blogs and Facebook pages of these organizations are collected.

![Diagram: Process of Collection of Organization Data](image)

**Figure 4.2:** Process of Collection of Organization Data

The crawlers developed for this data collection are as follows:

- Web crawler for collecting documents from websites of organizations
- Facebook crawler for collecting posts from the public Facebook pages of the organizations
The web crawler was developed using the Jsoup [13] JAVA library. First step was to collect all the unique URLs of particular organization. Second step was to extract the web HTML documents of these URLs and clean the HTML documents to extract the text contents accordingly. The web crawler is customized for each organization as the HTML document structure to each organization is unique.

The Facebook crawler was developed using the Graph API[7] of the Facebook and JAVA.

4.2 Data Cleaning

The text documents crawled by the crawlers are cleaned by removing stop words, URLs, numeric data and other special characters. A JAVA based program is written to perform this task, which tokenizes the text and removes special characters, numbers, stop words, and URLs. Only plain keywords of each text document are retained.

After this task each document has the list of keywords and their frequency of occurrence in the document.

Also a total bag of words i.e, the vocabulary for the text corpus for all the documents of all the organizations is constructed. These words are considered as the features for the text corpus.

4.3 Data Analysis and Model Generation

It is assumed that documents and posts crawled from the Radical organizations are radical and similarly the documents and posts crawled from Counter-Radical organizations are counter-radical.

The data analysis and model building is done based on the above assumption. A document term matrix(A) is constructed from all the documents of all organizations
where each cell value is the frequency of occurrence of the particular term in the particular document. The documents from Radical organizations are labeled as -1 and similarly the documents from Counter Radical organizations are labeled as 1. These are used as labels for building the model.

The data analysis is a two step process: first build a model for finding discriminative features (keywords) for both groups of documents Radical and Counter-Radical. Second within each group (Radical or Counter Radical) find the features with respect to each organization in the particular group.

For the first step, we use the linear and logistic regression, the inputs for this analysis is document-term matrix $A$, labels of each document $y(1/-1)$ and regularization parameter $\lambda$.

Consider $A$ as $m \times n$ matrix where there are $m$ documents from organizations and total vocabulary is $n$ terms. $y$ will be of dimensions $m \times 1$.

Linear or Regression analysis results in matrix $X$ of size $n \times 1$. The values of $X$ being the weightages of each $m$ terms. Non-zero values of $X$ are considered and values which are greater than 0 correspond to features of Counter-Radical Group and negative values correspond to Radical Group.

$$
\min_x \frac{1}{2} \| A x - y \|_2^2 + \frac{\rho}{2} \| x \|_2^2 + \lambda \| x \|_1
$$

**Figure 4.3:** Linear Regression

$$
\min_x \sum_{i=1}^m w_i \log(1 + \exp(-y_i (x^T a_i + c))) + \frac{\rho}{2} \| x \|_2^2 + \lambda \| x \|_1
$$

**Figure 4.4:** Logistic Regression
The example shown in the figure 4.5 consists of 10 documents and total vocabulary of 10 terms. Matrix A has the frequency of occurrence of these 10 terms in the 10 documents. There are 5 documents from each polarity. In total there are 4 organizations. Matrix y contains the polarity of each document represented by values 1/-1. Matrix X is the resultant Keywords weightages matrix containing zero/positive/negative values. The Discriminating features are obtained using the non-zero values of X.

The second step includes finding the keywords which are specific to each organization. The inputs are the document term matrix A size m x n, multi-class labels matrix Y of size m x k, where k being the total number of organizations and regularization parameter $\lambda$.

Linear or Regression analysis results in matrix X of size n x k. The values of X being the weightages of each m terms in each organization k. Only positive values of X are considered as the terms specific to each organization.
Figure 4.6: Multi Class Linear Regression

\[
\min_{x} \frac{1}{2} \|Ax - y\|^2 + \lambda \|x\|_{\ell_1/\ell_q},
\]

Figure 4.7: Multi Class Logistic Regression

\[
\min_{x} \sum_{i=1}^{k} \sum_{l=1}^{m} w_{il} \log(1 + \exp(-y_{il}(x_l^T a_{il} + c_l))) + \lambda \|x\|_{\ell_1/\ell_q},
\]

The example shown in figure 4.8 explains the multi-class linear and logistic regression analysis. Y is of size 10 x 4. The resultant keywords weightages matrix is of size 10 x 4.

Both the discriminating and organization specific are combined to form a list of candidate features from the organizational data.

Figure 4.8: Example showing the Data Analysis on A and Y Matrices
4.4 Social Media Crawling and Analysis

The resultant keywords from the organizational data are reviewed by social scientists and domain experts to get finer list of keywords. Along with the relevant keywords, the names of the organizations, abbreviations of the organizations, leader names of the organizations, top bi-grams and top frequent keywords in the entire corpus are accumulated to form the total set of candidate keywords for filtering the social media.

This new set of keywords and the geographical information of the country under consideration, both are used as filters for the streaming API and the tweets and user information are stored in the database.

The semantics of user-user relationship are also captured and saved in the database. The hashtags present in the tweets are also saved in the database.

To identify the polarity of each user, we consider all the tweets of a particular user aggregated over a period of time (say week or month) and also the contents of media articles used by that user in the time frame and make a single user document. Now the text in this document is cleaned in similar fashion to the organization document. This process is repeated for all the users.

For selected time frame, now we obtain the document term matrix similar to A and we use the matrix X i.e., the keyword weightages matrix, using the equation $AX = y$ we get the labels of each user in the matrix $y$. Thus we obtain the each polarity of the user for the time frame. This process is repeated for all time frames and user’s polarity (R/CR) is determined.

To determine the organization of the user, we use the equation $AX = Y$, where $X$ is the resultant matrix in the multi-class linear and logistic regression model developed using the organizational documents. For each user now we get the scores in $Y$, which
corresponds to the affinity towards an organization. The user is assigned to the organization based on the higher value in his particular row.

4.5 Breaking Hashtag Analysis

The concept of breaking hashtags are introduced in [4] using the three sigma rule from Statistics which states that nearly all values lie within three standard deviations ($\sigma$) of the mean ($\mu$) in a normal distribution the hashtag volume breakouts are defined. The daily volumes are considered for a hashtag and a fixed sized sliding window (of length 20 daily intervals), is used to compute a running average and standard deviation for each hashtag’s volume distribution. This period is called as accumulation period. Then, non-overlapping episodes are considered within a time-series of daily volumes for each hashtag whenever its daily volume exceeds ($\mu + 1\sigma$) of the previous 20 day periods. If the daily volume exceeds the ($\mu + 2\sigma$) without falling below minimum (0, ($\mu - 2\sigma$)) then the episode of the hashtag is defined as breaking, or otherwise non-braking. These breaking hashtags can be used to understand the behavior of the users during the online discussions.
Chapter 5

VISUALIZATION TOOL AND COMPONENTS

The Visualization Tool is in the form of a User Interface (UI). The UI is the collection of different information widgets combined in a single web HTML page. The UI is developed using the d3.js [3] and Google Annotated Timeline [8].

The various information widgets present in the UI are as follows:

- Timeline-Hashtags Widget.
- Chord Diagram.
- Heat Map Widget.
- User-User Network.
- URL Information Widget.
- Tags-Users-Keywords Widget.

Filtering of data can be done in any widget, which in turn changes the data in all other widgets based on the selection of filters. The users can be visualized based on filtering of time period, hashtags, organization and geographical location or any combination of the filters.

5.1 Timeline-Hashtags Widget

This widget presents the timeline developed using Google Annotated Timeline [8] and also list of breaking hashtags. By default the timeline shows the weekly or monthly volumes of the total tweets. Whenever a particular hashtag is selected,
**Figure 5.1:** Figure showing the Timeline, Chord Diagram and Heat Map Widgets
Figure 5.2: Figure showing the User-User Network, URL Information and Tags-Users-Keywords Widgets
the timeline is modified to show the weekly/monthly frequency distribution of the hashtag. A single hashtag or a group of hashtags can be selected to see the distribution of them for the selected timeline. Figure 5.3 shows the timeline when some sample hashtags are selected.

![Figure showing distributions of the Selected Hashtags in the Timeline](image)

**Figure 5.3:** Figure showing distributions of the Selected Hashtags in the Timeline

5.2 Chord Diagram

The distribution of Radical and Counter-Radical users is shown in the form of chord diagram. It depicts the User-Group mapping and also User-Organization mapping. It shows the distributions of all users current organization of particular week/month to the organization of the user in the previous week/month. The Chord diagram is rendered using the d3.js.

A particular organization can be selected from chord diagram to view the specific users and their information belonging to that organization. Also a particular chord can be selected to view the information of specific users who are shifting from organization A to organization B.
5.3 Heat Map Widget

The HeatMap Widget shows the geographical distribution of the users plotted on a map. It is developed using the Google Charts map Visualization\cite{10}. Filtering of the users can be done using the markers on the map. Upon clicking the area formed by this markers, users pertaining only to that particular area are shown.

The Heat Map shows the geographic footprint of the users. It not only shows the hot spots of the region under consideration but also visualizes the trans-national connections in other geographic regions.

5.4 User-User Network

The User-User Network captures the semantics of user mention and re-tweet features present in the twitter. It identifies the top influential users and the followers (in terms of user mention and re-tweet features) of these users. Top 100 users are selected based on the number of times of their occurrence. The Betweenness Centrality measure of all users are calculated using Brandes Algorithm \cite{5} to identify the influential users.
The size of these influential users is in accordance to the z-scores \[19\] calculated using the values of users having non-zero betweenness centrality. Using these influential users and their followers a User-User Network is created using the Force Directed Graph of d3.js\[2\].

The z-score\[19\] of a particular user \(i\) is calculated by

\[ zscore_i = \frac{\text{BetweennessCentrality}_i - \mu}{\sigma} \]

where \(\mu\) is the mean and \(\sigma\) is the standard deviation of all non-zero betweenness centralities.

A click on a user in this network opens up the twitter profile of the user.

5.5 URL Information Widget

This widget shows all the URLs which are used in the tweets of the users. The following information of the URLs is shown in the widget:

- URL Title
- Domain of the URL
• Count, which mentions how many times this URL is mentioned in the tweets

• The publish date of the article

5.6 Tags-Users-Keywords Widget

This widget shows the top 20 breaking hashtags, top 20 users based on their scores and top 20 keywords used in the tweets of the users.

This information present in this widget gives the additional information which is related to the existing data based on the filtering done in other widgets.
Chapter 6

EXPERIMENTAL RESULTS

The data analysis and model development for classifications of the online social media users is done on data sets of Malaysia, UK and Indonesia.

6.1 Results for Malaysia Data sets

The organizations in Malaysia are grouped into two groups at higher level Radical and Counter Radical.

\textbf{Figure 6.1:} Figure showing the Hierarchy of the Users Classification in Malaysia
Within the Radical group of organizations there are two sub groups called Ethnonationalists and Change-Oriented.

### 6.1.1 First Level of Classification

A first level of classification model is developed to classify the documents of organizations as Radical (R) or Counter-Radical (CR). The document term Matrix A which is generated is used to build the model. Matrix A is split into two data sets called training data set and testing data set.

The linear and logistic regression is performed on this training data set to build a model. This model is reverse engineered to test the testing data. As the labels (R/CR) of this data is already known, the accuracy of the built model is calculated.

To ensure the model built is accurate to classify the given document, 10-fold cross validation is performed on the entire A matrix to get different training and testing data sets. At each level the accuracy of the model is calculated and an average accuracy of the model is obtained accordingly.

This process is repeated for different values of the regularization parameter. Also at each level of testing of the model for different values of regularization parameter, the number of discriminating features (keywords) for the groups R and CR are also obtained.

---

**Table 6.1: Organizations Data-Malaysia**

<table>
<thead>
<tr>
<th>Organization Type</th>
<th>Number of Organizations</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counter-Radical</td>
<td>8</td>
<td>53276</td>
</tr>
<tr>
<td>Radical - Ethnonationalists</td>
<td>3</td>
<td>6041</td>
</tr>
<tr>
<td>Radical - Change Oriented</td>
<td>6</td>
<td>25700</td>
</tr>
<tr>
<td>Regularization Parameter</td>
<td>Accuracy</td>
<td>Number of features</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>0.01</td>
<td>0.905136776</td>
<td>600</td>
</tr>
<tr>
<td>0.02</td>
<td>0.89798527</td>
<td>252</td>
</tr>
<tr>
<td>0.03</td>
<td>0.89426821</td>
<td>150</td>
</tr>
<tr>
<td>0.04</td>
<td>0.886587513</td>
<td>97</td>
</tr>
<tr>
<td>0.05</td>
<td>0.882564662</td>
<td>84</td>
</tr>
<tr>
<td>0.06</td>
<td>0.87980045</td>
<td>72</td>
</tr>
<tr>
<td>0.07</td>
<td>0.877295005</td>
<td>63</td>
</tr>
<tr>
<td>0.08</td>
<td>0.872049033</td>
<td>52</td>
</tr>
<tr>
<td>0.09</td>
<td>0.87004942</td>
<td>46</td>
</tr>
<tr>
<td>0.1</td>
<td>0.870555272</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 6.2: Results of Linear Regression-Malaysia

Both linear and logistic regression models are evaluated with different values of regularization parameter. Linear regression model yielded good results. The results of the linear regression modeling with different regularization parameter values, the accuracy obtained and the number of discriminating keywords generated are listed in Table 6.1.

Linear Regression is considered and the regularization parameter value is set as 0.02 which resulted in the accuracy of the model around 90% with 10-fold cross validation. The number of discriminating features obtained are 253.

6.1.2 Second Level of Classification

The second level of classification is between the Ethnonationalists and Change-Oriented sub groups in the Radical group of organizations.
A similar process is followed as the first level of classification. Considering the Linear Regression with regularization parameter set to 0.02, the resultant accuracy of the model with 10-fold cross validation is 94%.

6.1.3 Organization Level Classification

After identification of the sub group as Counter-Radical or Ethnonationalist or Change-Oriented, a multi-class Linear and logistic regression analysis is performed on each sub group to find the organization of the document within each sub-group.

The entire data set is now made into three parts as mentioned above. On each part the multi class regression models are built using the training data and the model is evaluated using the testing data sets.

Also the features which are specific to each organization are also obtained.

Considering the multi-class linear regression model with regularization parameter set to 0.02, the accuracy for the Counter-Radical organizations group is 84.5% with 10-fold cross validation. Similarly the average accuracy between the organizations of the Radical groups is around 91% with 10-fold cross validation.

Organization specific keywords for entire data set obtained are 190.

Total unique features, i.e. the combination of discriminating features and the organization specific keywords for the Malaysia data set are 326.

6.2 Results for UK Data sets

The organizations in UK are grouped into two groups ah higher level Muslim and Non-Muslim.

The data classification model is build using the logistic regression framework for UK data set. At higher level the users are first classified as Muslim or Non-Muslim. Within the Muslim group of users, they are sub divided and classified as Radical
### Classification Accuracy

<table>
<thead>
<tr>
<th>Classification</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radical/Counter-Radical Groups Classification</td>
<td>0.8979</td>
</tr>
<tr>
<td>Ethno-Nationalist/Change Oriented Classification</td>
<td>0.94</td>
</tr>
<tr>
<td>Organizations within Counter-Radical Group Classification</td>
<td>0.845</td>
</tr>
<tr>
<td>Organizations within Radical Group Classification</td>
<td>0.9102</td>
</tr>
</tbody>
</table>

**Table 6.3:** Accuracy of Classification Malaysia

<table>
<thead>
<tr>
<th>Organization Type</th>
<th>Number of Organizations</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counter-Radical</td>
<td>6</td>
<td>612</td>
</tr>
<tr>
<td>Radical</td>
<td>18</td>
<td>4522</td>
</tr>
<tr>
<td>British Nationalists</td>
<td>3</td>
<td>8705</td>
</tr>
</tbody>
</table>

**Table 6.4:** Organizations Data-UK

and Counter-Radical. Similarly after finding each sub group of users as Radical or Counter-Radical and Non-Muslim, then they are classified to the particular organization with in the sub group.

Table 6.6 gives the different classification models used and their accuracies for the UK dataset.

The location information is collected from the twitter data using 2 methods. First, the tweet contains the latitude longitude information and the place information. Second being the location of the user from his profile.

### 6.3 Results for Indonesia Data sets

The organizations in Indonesia are grouped into Radical and Counter-Radical organizations at the higher level. Each sub-group has list of organizations in it.
The data classification model is built using the logistic regression framework for Indonesia data set. At higher levels, the users are first classified as Radical or Counter-Radical. After finding each sub-group of users as Radical or Counter-Radical and Non-Muslim, then they are classified to the particular organization within the sub-group.
Figure 6.3: Figure showing the Hierarchy of the Users Classification in Indonesia
<table>
<thead>
<tr>
<th>Logistic Regression</th>
<th>SVM</th>
<th>Random Forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muslim/Non-Muslim level</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>Organizations level</td>
<td>0.84</td>
<td>0.88</td>
</tr>
</tbody>
</table>

**Table 6.6:** Comparison of Accuracies for UK dataset using different Classifiers

<table>
<thead>
<tr>
<th>Information</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets with Latitude and Longitude information</td>
<td>5%</td>
</tr>
<tr>
<td>Tweets with location information obtained from User’s profile</td>
<td>94.9%</td>
</tr>
<tr>
<td>Tweets with no location information either ways</td>
<td>0.1 %</td>
</tr>
</tbody>
</table>

**Table 6.7:** Location Information of the Tweets UK dataset

<table>
<thead>
<tr>
<th>Organization Type</th>
<th>Number of Organizations</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counter-Radical</td>
<td>14</td>
<td>26524</td>
</tr>
<tr>
<td>Radical</td>
<td>13</td>
<td>11246</td>
</tr>
</tbody>
</table>

**Table 6.8:** Organizations Data-Indonesia

<table>
<thead>
<tr>
<th>Classification</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radical/Counter-Radical Groups Classification</td>
<td>0.98</td>
</tr>
<tr>
<td>Organizations level Classification</td>
<td>0.83</td>
</tr>
</tbody>
</table>

**Table 6.9:** Accuracy of Classification Indonesia
This chapter covers some of the scenarios captured by the LookingGlass from the different data sets. These scenarios are taken from the LookingGlass demo video \cite{6} prepared by Professor Hasan.

7.1 Polarization between different Groups

The LookingGlass of UK shows the polarization of the users between the British Muslims and the followers of Anti-Muslim English Defense League and British Nationalist parties.

The User-User Network captures the polarization between the two different groups in UK.

\textbf{Figure 7.1:} Figure showing the Polarization between two different groups in UK
7.2 Breaking Hashtags and Extremist Concepts

In LookingGlass Indonesia, on selection of different breaking hashtags related to extremist concepts in the timeline and selecting users of Unaffiliated_R group in the chord diagram reveals a Radical user and following his twitter page shows the images he is using.

The URL widget also shows the media articles used by the group of most influential users which are endorsing strict Syria enforcement and hateful articles related to Shiah minority.

Also zooming into the heat-map widget gives the exact location from where the tweets are coming from.

Figure 7.2: Figure showing User of a Radical Group and his Twitter page
Chapter 8

SUMMARY

This thesis work focuses on Data analysis and visualization tool development that tracks the activities of online followers of Islamic Radical and Counter-Radical groups and social movements. Similar data analysis and visualization tool is developed for different geographical regions Indonesia, United Kingdom(UK) and Malaysia(in progress). The visualization tool not only gives the real time data insights about the users and their activities but also helps to analyze and detect the hot-spots of geographical regions, the active topics of debates in online communities, the drivers for growth and shrinkage of the online followers of particular ideology or belief(organization).
Chapter 9

FUTURE WORK

The data collected and the research activity currently done can be extended to discover new trends. The activities that can be considered are as below:

- To integrate users of other social media like Facebook, current consideration is limited to the twitter social media only.

- To perform the Sentiment analysis and see the change in sentiments of the twitter users from their messages before and after the break out of hashtags.

- To find latent communities and groups present within each organization group and re-train the data analysis model to understand how these latent groups are expanded or shrunken.

- To consider only the ideology related web articles of the organizations rather than considering all the web-articles of it.

- To also consider the context based user-user relationship (friend and follower) along with the content based relationship.

- The approach can be extended to predict the outcomes of election results by understanding the percentage of online followers and users affiliated to each organization/political party.
REFERENCES


Shoubin Kong, Fei Ye, Ling Feng, and Zhe Zhao. Towards the prediction problems of bursting hashtags on twitter. Journal of the Association for Information Science and Technology, pages n/a–n/a, 2015. ISSN 2330-1643. doi: 10.1002/asi.23342. URL http://dx.doi.org/10.1002/asi.23342

