Detecting Crime Types using Twitter
Deepak Venugopal
University of Memphis

**Motivation**
Predict fine-grained crime occurrences that could help in efficient allocation of resources

**Approach**
Use Twitter to provide real-time predictions of crime events specific to a Geo-Location

**Data**
- Collect tweets from Twitter API for a specific geolocation
- Filter tweet text with a vocabulary of common words suggesting crimes
- Annotate each tweet with one of following categories of crime (Violent Crime, Narcotics, Racism, Fraud)
- Data collected from co-ordinates corresponding to TN
- Challenging to obtain sufficient data from specific geolocations (e.g. Memphis)

**Twitter-based real-time prediction**
- Built on top of Amazon Cloud Services
- Collect and extract language features from tweets using Spark Streaming (cluster computing framework)
- Apply machine learning methods to categorize tweets based on language features
- Rich visual interface to interpret prediction results through Amazon cloud-based Kibana

**HMM model**
- Twitter data has the advantage that it is real-time
- However, twitter data can be very noisy and it is quite difficult to get reliable signals for crime prediction from twitter text alone
- Augment with real crime data from BlueCrush
- BlueCrush stores real crime events in Memphis, location and type of crime
- Develop a Hidden Markov Model (HMM) to predict future crime types based on recorded crimes
- Model latent crime states
- From observed incidents learn to transition between crime states

**Future Work**
- Combine predictors from Twitter and HMM in a principled manner
- Derive advanced linguistic features using neural embeddings by taking advantage of unlabeled data
- Come up with detailed evaluation measures

**Results**
- 5000 Tweet Corpus Manually annotated for training

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Fold-1</th>
<th>Fold-2</th>
<th>Fold-3</th>
<th>Fold-4</th>
<th>Fold-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultinomialNB</td>
<td>0.79</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.83</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.83</td>
<td>0.86</td>
<td>0.83</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>SVM (RBF)</td>
<td>0.5</td>
<td>0.6</td>
<td>0.48</td>
<td>0.51</td>
<td>0.5</td>
</tr>
<tr>
<td>SVM (linear)</td>
<td>0.86</td>
<td>0.88</td>
<td>0.88</td>
<td>0.86</td>
<td>0.86</td>
</tr>
</tbody>
</table>

HMM predictions using a corpus of incidents from 12/1/2018 – 2/16/2018 (Error for the predicted values in the last 10 days in the corpus)

<table>
<thead>
<tr>
<th>States=3</th>
<th>States=5</th>
<th>States=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.12</td>
<td>0.17</td>
</tr>
</tbody>
</table>

5-fold Cross Validation on annotated twitter corpus (weighted F1-scores)