Detecting Illicit Drug Usage on Twitter

Darrel A. Pollard dap2525@rit.edu
Advisor: Christopher M. Homan cmh@cs.rit.edu

CONTRIBUTIONS
Using Twitter data we build a bootstrapping, humans-in-the-loop classifier to identify tweets about drugs.

TOOLS & DATA
Tweets were JSON objects obtained from twitter that contain fields containing information like the message and date tweeted. The JSON objects were updated with label “drug_relation” which contain value true or false.

Example Data: [{"drug_related": true, "text": "I swear weed make my :( turn into :) every time"}]

Tweets were preprocessed and feature extracted using:
• ScikitLearn – Vectorizer & SVM
• NLTK – Natural Language Processing
• Tweepy – Python Twitter API gathering tweets

Data
• Set 1 - Contained Drug Keyword data and U.S. general tweets 1,500 tweets
• Set 2 – contain another set of drug filtered data, Rochester Tweet data, and Detroit Tweets - 1,604 tweets

METHOD
Conventional methods such as keyword checking often yield false positives. This is the case when using the Streaming API to pull tweets based on drug keywords.

“This is Dope”

Support Vector Machine
Using SVM we trained two classifiers to predict if a tweet was drug related or not drug related.

• Extracted Features based on Tri-grams of each tweet.
• Due to the imbalance of the classes in both data sets, we used variable weight set passed to the SVM to train for the best out come.

RESULTS
Recall and Precision of Classifier A and B are bellow. These result are from 20 percent test data.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Not Drug Related</th>
<th>Drug Related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.97</td>
<td>0.93</td>
</tr>
<tr>
<td>Recall</td>
<td>0.99</td>
<td>0.62</td>
</tr>
<tr>
<td>F1-Score</td>
<td>0.98</td>
<td>0.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Not Drug Related</th>
<th>Drug Related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>Recall</td>
<td>1.00</td>
<td>0.85</td>
</tr>
<tr>
<td>F1-Score</td>
<td>0.99</td>
<td>0.92</td>
</tr>
</tbody>
</table>

CONCLUSION
Ideally Bootstrapping should be done in several iterations and more data to balance out the class imbalance would also improve performance. In the future we would also like to add sentiment analysis as a way to improve and detect drug related topics.