Sentiment Analysis in Twitter: A Graph based Hashtag Sentiment Classification Approach

Santosh Tankala (2642576)
VishnuTeja (2671322)
Akhil Reddy (2641570)
What is Belief Network?

- Dynamic Programming
- Directed acyclic Graph
- Conditional Probability Tables
Hashtag Level Sentiment Classification

• **Focus**: Hashtag Level Sentiment Classification

• **Goal**: Overall sentiment polarity for a given hashtag in a specified time period.

• 3 Types of information to address the task
  – Sentiment polarity of tweets containing the hashtag.
  – Hashtag co-occurrence relationship.
  – Literal meaning of hashtags.
Hashtag Graph Model

• Hashtag Graph HG =\{H, E\};
• Hashtags H_i = \{h_1, h_2, h_3, \ldots h_m\};
• Tweet T_i = \{t_1, t_2, t_3, \ldots t_n\};
• Sentiment Polarity Y_i = \{y_1, y_2, y_3, \ldots y_n\} where y_i = \{pos, neg\}
Figure 1: An example of a Hashtag Graph Model
i. Potential function for Tweet based factor.

ii. Polarity probability for each tweet.

iii. Hashtag-hashtag Information – Gives link info to allow neighbouring hashtags to influence classification result.

iv. Regularization function.

\[
\log (\Pr(y|HG)) = \sum_{h \in H} \log (\phi(y|h)) \\
+ \sum_{(h_j, h_k) \in E} \log (\psi_{j,k}(y_j, y_k|h_j, h_k)) - \log Z
\]
Appropriate Collective Classification Algorithms

• A structured logistic regression based algorithm was investigated as inference method in a link based text classification framework.

• **Algorithms:**
  1. Loopy Belief propagation (LBP)
  2. Relaxation Labelling (RL)
  3. Iteration Classification Algorithm (ICA)
Loopy Belief Propagation (LBP)

\[
\phi_i(y_i | h_i) = \sum_{\tau \in \mathcal{T}_i} \Pr_{y_i}(\tau)
\]

\[
\psi_{j,i}(y_j, y_{i_0} | h_j, h_{i_0}) = \frac{\#(h_j, h_{i_0})}{\#(h_j) + \#(h_{i_0}) \cdot \mathbb{1}_{y_j = y_{i_0}}}
\]
Relaxation Labelling (RL)

\[ r(y_i, y_j) = \frac{p(y_i, y_j) - p(y_i)p(y_j)}{(p(y_i) - p(y_i)^2)^2(p(y_j) - p(y_j)^2)^2} \]  \quad (5)

Naturally, it is expected that the hashtags that are more likely to co-occur in tweets to have more mutual influence. We hereby have:

\[ d_{i,j} = \frac{\#(h_i, h_j)}{\#(h_i)} \]  \quad (6)
Iterative Classification Algorithm (ICA)

\[
\phi_{\delta}(y_{\delta} | k_{\delta}) = \sum_{\tau \in T_{\delta}} I_{y_{k_{\delta}} = y_{\delta}}
\]

\[
\psi_{\delta, j}(y_{\delta}, y_{\gamma} | k_{\delta}, k_{\gamma}) = \frac{\#(k_{\delta}, k_{\gamma})}{\#(k_{\delta}) + \#(k_{\gamma})} \cdot I_{y_{k_{\delta}} = y_{\delta} \land y_{k_{\gamma}} = y_{\gamma}}
\]
Experimental Study

• **Data Collection & Evaluation**

  – Challenge: Sentiment Polarity tweets cannot always be determined with confidence.
  – Select Strong Positive & Strong Negative Lexicons.
  – Calculate accuracy, precision, recall & F1 metrics.
• Data Collection & Evaluation Cont..
  – Raw Data: 0.6 Million tweets
  – After Data Cleaning & Pre-processing:
    Hash tags: 2181  
    Tweets: 29195  
    Size of Edge set: 27430  
    Strong Positive words: 595
    Strong Negative words: 352
  Total Dataset is then split into Training set & Test set for classification results.
An Example of Enhanced Boosting Classification
Hash Tag level Sentiment Classification

• Stage 1: Subjective Classifier (SVM Classifier)
  – To filter neutral tweets.
  – Score denoted by $s$

• Stage 2: Polarity Classifier (SVM Classifier)
  – To find positive/negative of the tweet.
  – Score denoted by $\xi$

Model Classification Used: 5-fold Cross Validation
Probability Calculation of a tweet

\[
Pr_{pos}(t) = \begin{cases} 
1 & s \geq \xi \\
0.5 + s/(2\xi) & s \in (-\xi, \xi) \\
0 & s \leq -\xi
\end{cases}
\]  

(10)

\[
Pr_{neg}(t) = 1 - Pr_{pos}(t)
\]  

(11)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>subjectivity(1)</td>
<td>83.13%</td>
<td>59.45%</td>
<td>36.59%</td>
<td>45.27%</td>
</tr>
<tr>
<td>polarity(2)</td>
<td>88.96%</td>
<td>90.49%</td>
<td>94.82%</td>
<td>92.60%</td>
</tr>
<tr>
<td>(1)+(2)</td>
<td>84.13%</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

Table 1: Performance of the tweet-level classifier
Probability of a Hash Tag: SVM-Voting Baseline

\[
Pr(y_{\hat{y}}|h_{\hat{y}}) = \frac{\sum_{\tau \in T_{\hat{y}}} Pr_{y_{\hat{y}}} (\tau)}{\sum_{\tau \in T_{\hat{y}}} Pr_{pos} (\tau) + \sum_{\tau \in T_{\hat{y}}} Pr_{neg} (\tau)}
\]

\[
y = \arg \max_{y_{\hat{y}} \in \{pos, neg\}} Pr(y_{\hat{y}}|h_{\hat{y}})
\]
Evaluation of Hashtag level Sentiment Classification

<table>
<thead>
<tr>
<th>Setup</th>
<th>Accuracy(%)</th>
<th>Pos-Precision(%)</th>
<th>Pos-Recall(%)</th>
<th>Pos-F1(%)</th>
<th>Neg-Precision(%)</th>
<th>Neg-Recall(%)</th>
<th>Neg-F1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-voting</td>
<td>55.96</td>
<td>64.03</td>
<td>68.23</td>
<td>66.06</td>
<td>39.61</td>
<td>35.22</td>
<td>37.29</td>
</tr>
<tr>
<td>LBP</td>
<td>56.28</td>
<td>73.26</td>
<td>47.89</td>
<td>57.92</td>
<td>44.44</td>
<td>70.45</td>
<td>54.50</td>
</tr>
<tr>
<td>RL</td>
<td>58.07</td>
<td>71.90</td>
<td>54.62</td>
<td>62.08</td>
<td>45.45</td>
<td>63.92</td>
<td>53.12</td>
</tr>
<tr>
<td>ICA</td>
<td>59.23</td>
<td>71.81</td>
<td>57.81</td>
<td>64.05</td>
<td>46.36</td>
<td>61.64</td>
<td>52.92</td>
</tr>
<tr>
<td>LBP-Boost</td>
<td>77.72</td>
<td>97.30</td>
<td>66.69</td>
<td>79.14</td>
<td>62.91</td>
<td>95.88</td>
<td>75.97</td>
</tr>
<tr>
<td>RL-Boost</td>
<td>72.97</td>
<td>98.33</td>
<td>57.80</td>
<td>72.81</td>
<td>57.99</td>
<td>98.33</td>
<td>72.95</td>
</tr>
<tr>
<td>ICA-Boost</td>
<td>77.40</td>
<td>95.57</td>
<td>67.05</td>
<td>78.88</td>
<td>63.02</td>
<td>96.04</td>
<td>76.11</td>
</tr>
</tbody>
</table>

Table 2: Performance of the hashtag-level classifiers.
Conclusion & Enhancements

• As the baseline approach is not encouraging a Graph based model is proposed to improve the hash tag level sentiment classification.
• The enhanced boosting framework has shown significant results.
• In addition to giving polarity, a snapshot information for the hashtag can be improved.