**Abstract**

STRAP is a system:
1. for capturing and analyzing social network data
2. for the purpose of sentiment analysis and prediction in real-time based on public opinions.
3. initially set to predict the New York Times Bestsellers’ List for the upcoming week
4. that is highly extensible and can be used in multiple problem domains with minimal modification.

**Motivation**

- **Book Retailers/Publisher:** Know in advance what books are more popular and more likely to run out of stock
- **General Retailers:** Find a ranking of popularity of different products (target commercials, etc)
- **Open Source:** Provide an application that can be reused to satisfy alternate objectives.

**Goals**

- Develop the software architecture to support our objectives that is abstracted, extensible, and scalable.
- Support both contextual and sentiment analysis in order to provide on-the-fly insight extraction.
- Demonstrate one particular application of our system using literary references.

**System Workflow**

**User queries the system**

Retrieves data

System predicts upward or downward trend

**Sample Tweets**

"I finished Life of Pi yesterday. Such a good book, so easy to read and stay interested with, and the ending has stayed with me a bit."

"My spiritual awakening was caused by both The Life of Pi (novel) and V for Vendetta. Don’t ask why the latter."

"Started reading "Fifty Shades of Grey" I'm very intrigued!!"

**Results**

- **Collector** has collected over 300 million tweets
- **Filter** has a 0.00003x compaction factor
- **Ranker** uses the following features for prediction:

<table>
<thead>
<tr>
<th>Features</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral cues</td>
<td>28.44%</td>
</tr>
<tr>
<td>Sadness cues</td>
<td>35.53%</td>
</tr>
<tr>
<td>Joy cues</td>
<td>11.87%</td>
</tr>
<tr>
<td>Avg. SentiWordNet</td>
<td>-19.58%</td>
</tr>
<tr>
<td>Num. of cues</td>
<td>30.49%</td>
</tr>
<tr>
<td>Negative cues</td>
<td>38.82%</td>
</tr>
<tr>
<td>Positive cues</td>
<td>22.12%</td>
</tr>
<tr>
<td>Pointwise Mutual Information</td>
<td>-22.85%</td>
</tr>
</tbody>
</table>

- Combine the above with a classifier yields a predictor that outperforms the odds.

**Future Work**

- **Extension:** Modify the pipeline to apply to other products, data sources, etc.
- **Filter:** Improve filtering methods using Machine Learning techniques (e.g. Naïve Bayes, SVM, etc)
- **Statistical Methods:** Use statistical modeling in addition to social data for better ranking results.

**Conclusion**

The system:
1. yielded statistically significant results when used to assess popular literary works.
2. was only limited by the paucity of literary mentions on twitter.
3. fully satisfied our secondary and more concrete objective of designing a system to filter, analyze and rank a sampled Twitter stream.
4. is abstractable to a variety of domains (e.g. products, movies, etc), is functionally extensible and operates speedily at scale.