Classifying News Stories to Estimate the Direction of a Stock Market Index (Brett Drury, Luis Torgo and J. J. Almeida)[1]

Hao Fu, Jiatong Ruan
Introduction
Background

- **Timely** information from news -> **Prediction** of the prospects of economic actors
- News Information: the **past** or the **future** VS. Numeric data: the **past**
- Some published methods exist:
  - manually created rules
  - models learnt from **manually** selected data and **manually** constructed dictionaries
- Disadvantage: Rely on human annotator
**Related Work**

Manually organize news stories

- 19 categories with different levels. [2]
- Using machine readable news to automatically classify stories. [3]
- Increase to 39 categories. [4]
- Dictionary contains 423 features. [5]

Alignment of news stories to market movement [6]

- limited in single companies and where the company names are on headlines. [6]
News Story Classification

Manual constructed rules with automatically constructed dictionaries + Alignment of stories with sharp market movement + Self-training to construct a model to classify news stories

Fig 1: Proposed Classification
Data

Amount: **News stories** (>300,000)

News Source: **Really Simple Syndication** (RSS) feeds

Time Period: **Oct. 2008 - Jun. 2010**, crawler ran at the same time each day

Database: **RDBMS**: headline, description, published data and story text

Stock Data: **Yahoo Finance**
Data

Data pre-process:

- Remove duplicate stories and non-finance stories
- Remove sentences that did not contain the named entities: companies, organizations, market indexes and company employees.
- The sentence set was parsed with the ANNIE Part of Speech Tagger[8].
Model from Rule Selected Data[7]

- **Economic Actor** (company, organization, market, etc.)
- **Verb/Adj.**
- **Object** (profits, unemployment, etc.)

- **Classified** as positive or negative
- **Unclassified**

Fig 2: Rule Classifier Model
Alignment of Market Data

- Assumption: If the market moves *sharply* then this movement will be reflected in the published news stories.
- This strategy selected data by labelling news stories by their *co-occurrence* with a single *market movement*.
- A *positive* day is assumed to be when the market move by more than 1.7% and a *negative* day when the market lose more than 2.11%. 
Hybrid of Rules and Alignment

- This strategy attempts to mitigate the flaws of a rule classifier and alignment with a simple voting strategy.

Fig 3: Hybrid Strategy for equal labels
Hybrid of Rules and Alignment

The strategy ensured that stories which were **contrary** to market trend were not included in the training set.

Fig 4: Hybrid Strategy for contradictory labels
Proposed Algorithm

Initial training set by "Hybrid" strategy

Induce three models by headline, description and story text from training set

Any unlabelled story not processed?

Classify instance by three models

Check if their confidences all greater than a given constant?

Check if their classification all same?

Add this document to a temporary set Id and remove it from unlabelled stories

Merge training set and temporary set Id

If temporary set Id empty?

Return training set and models

Fig 5: Flow Diagram for Proposed Algorithm
Evaluation

- The evaluation methodology is based on estimated F-Measure.
- The F-Measure is estimated for models generated from: headline, description and story text information.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Headline</th>
<th>Text</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rules</td>
<td>0.77</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>Alignment</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.66</td>
<td>0.57</td>
<td>0.58</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.84</td>
<td>0.71</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Fig 6: Estimated F-Measure for competing strategies
Conclusion

- This paper presents a proposed method for categorizing news stories into positive or negative categories.
- By combining a rule classifier and alignment with market movement the chance of identifying events which may influence the market is increased. The proposed method adds further documents with a self-training method.
- The proposed method has a clear advantage over the competing methods by F-Measure.
Contribution

- Designed a **hybrid strategy** that can mitigate the flaws of rule classifier and alignment of market data.

- Proposed a new algorithm by introducing **self-training** to utilize **unlabelled** training data for training more robust model.
Limitations

- How models are induced from headline, description and story text, which is really important for us to evaluate, is not clearly presented in paper.

- Market movement depends on many factors, some of which might be contradictory, it's probably not a good idea to ignore data that contrary to market trend.
Future Work

- Evaluate techniques with news published when the market is closed.

- Assign a relevance measure to news story.

- Utilize news volume.
Q & A
Reference


Reference


