**Abstract**

This paper presents a web-based information system, RiskFinder, for facilitating the analyses of soft (textual information) and hard (numerical measures) in financial reports and broadens the analyses from word level to sentence level.

- Collecting financial reports from SEC and data preprocessing
- Collecting metadata from both website and parsing from reports
- Construct risk-sentiment dataset
- Adopt different models on risk-sentiment dataset to obtain classifier
- Use above classifier and 10-K corpus to build RiskFinder

Figure. System Construction Flow

* RiskFinder is now online: https://cfda.csie.org/RiskFinder/

**System Description**

- Metadata, including basic information on each company as well as several relevant financial measures.
  1. Company name
  2. Company’s CUSIP number
  3. Report release date
  4. Annualized post-event return volatilities
  5. Number and percentage of highlighted risk-related sentences in selected reports

- Visualization
  
  
  ![Visualization](image)

  Our system displays time-aligned quantitative information (e.g. **historical prices and trading volumes**) to facilitate the analysis of textual and numeric information and capture more insight in the financial risk.

  - Red Vertical Line: release date of the report
  - Users can adjust the window (6) to show the corresponding quantitative information for a certain period.

**Risky Sentences Detector**

RiskFinder focuses on the sentence-level investigation. To obtain better sentence representations, we adopt two sentence embedding techniques, FastText and Siamese-CBOW, to conduct sentence-level textual analyses. Due to the significant superior performance of FastText, we adopt it to highlight high-risk sentences in the reports in the proposed system.

<table>
<thead>
<tr>
<th>Siamese-CBOW</th>
<th>FastText</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.656</td>
</tr>
<tr>
<td>Precision</td>
<td>0.776</td>
</tr>
<tr>
<td>Recall</td>
<td>0.438</td>
</tr>
<tr>
<td>F1-measure</td>
<td>0.558</td>
</tr>
</tbody>
</table>

* 4-fold cross-validation

**Risk-Sentiment Dataset**

Form 10-K risk-sentiment dataset provides the financial risk-sentiment dataset consisting of two types of data:
1. A set of risk-labeled financial sentences
2. The pre-trained sentence embeddings.

<table>
<thead>
<tr>
<th></th>
<th>High-Risk</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Sentence</td>
<td>138</td>
<td>432</td>
</tr>
<tr>
<td>Cronbach’s alpha</td>
<td>0.784</td>
<td></td>
</tr>
</tbody>
</table>