BEYOND WORD-LEVEL TO SENTENCE-LEVEL SENTIMENT ANALYSIS FOR FINANCIAL REPORTS

Chi-Han Du,† Ming-Feng Tsai,‡ Chuan-Ju Wang†
†Academia Sinica, Taiwan
‡National Chengchi University, Taiwan

What is sentiment analysis for financial reports?

Labeled in word-level by financial sentiment word lexicon (Loughran, 2011)

In addition, revenues increased due to fee income on growing variable COLI account values, partially offset by declines in fees on leveraged COLI as that block of business continues to decline due to the HIPA Act of 1996. Benefits, claims and expenses increased $593, or 63%, to $1.5 billion in 1998 from $938 in 1997 due primarily to the MBL Recapture discussed previously.

Labeled in sentence-level by multiple financial experts (high risk)

In addition, revenues increased due to fee income on growing variable COLI account values, partially offset by declines in fees on leveraged COLI as that block of business continues to decline due to the HIPA Act of 1996. Benefits, claims and expenses increased $593, or 63%, to $1.5 billion in 1998 from $938 in 1997 due primarily to the MBL Recapture discussed previously.

Motivation

→ Use existing knowledge (financial sentiment lexicon) to improve sentence-level classification performance of deep learning models.
→ Extend boundary of financial sentiment out of word range by semantics, for each sentiment (positive, negative, litigious, and uncertain) shown in sentence.

Sub-phrase Algorithm

1. function Sub-Phrase \((T_M, k, \ell)\):
2. Input : A frequency table \(T_M\) including the top \(k\) most frequent sentiment \(n\)-grams and their frequencies, for \(n = 2, \ldots, M\); the number of iterations, \(\ell\)
3. Output: A reference table, \(W\)
4. \(W \leftarrow \{\}\)
5. for \(e = 1\) to \(\ell\) do
6. Find the most frequent word pair \(w_i, w_j\) in \(T_M\);
7. Find all \(n\)-grams containing \(w_i\) and \(w_j\) within \(T_M\);
8. Merge these two words into a new “word”;
9. Add the merged new “word” \(w_{ij}\) to the reference table \(W\);
10. Delete the most frequent word pair \(w_i, w_j\) in \(T_M\);
11. Update the frequency table \(T_M\) by replacing \((w_i, w_j)\) as \((w_{ij})\);
12. end
13. return \(W\);

Main Results

<table>
<thead>
<tr>
<th></th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>tf-idf</td>
<td>88.27</td>
</tr>
<tr>
<td>tf-idf+senti-phrases</td>
<td>87.15</td>
</tr>
<tr>
<td>LSTM [8]</td>
<td>86.96</td>
</tr>
<tr>
<td>LSTM+senti-phrases</td>
<td>87.14</td>
</tr>
<tr>
<td>CNN [9]</td>
<td>86.33</td>
</tr>
<tr>
<td>CNN+senti-phrases</td>
<td>86.35</td>
</tr>
<tr>
<td>fastText [10]</td>
<td>87.76</td>
</tr>
<tr>
<td>fastText+senti-phrases</td>
<td>88.03</td>
</tr>
<tr>
<td>SiameseCBOW [11]</td>
<td>87.92</td>
</tr>
<tr>
<td>SiameseCBOW+senti-phrases</td>
<td>88.79</td>
</tr>
</tbody>
</table>

→ Combining words to generate senti-phrases is not beneficial to the traditional bag-of-word model.
→ Complicated DL models achieve better performance than naive models, but all DL models perform better when using senti-phrases.

An application

→ Our new developed tool: Financial Risk Information Detecting and analyzerYing System (FRIDAYS) (AAAI’19)

→ The proposed algorithm is fast to compress data and even improve the semantics of NLP models for financial texts.
→ As a result, in the future it could be applied for summarization of financial corpus, or even automatic generation (NLU) for financial reports.