Text Analytics in Finance

Stephen Pulman

Dept. of Computer Science, Oxford University stephen.pulman@cs.ox.ac.uk
and TheySay Ltd, www.theysay.io

@sgpulman

SAP Central Bank Executive Summit
Overview

- What is text analytics and who uses it?
- How does it work?
- Using sentiment analysis for financial prediction: some case studies.
- Predicting risk in the markets and in banking.
- Detecting deception: can we trust what CEOs are saying?
- Text properties of some central bank governors’ speeches.
What is text analytics?

Content analysis
- mining for significant correlations: who is interested in what?
- topic detection and classification: what are these texts about?

Analysis of author’s attitude
- sentiment analysis: positive and negative attitudes towards people, policies, products etc.
- emotion classification: fear, joy, surprise...
- detecting signals like risk, intent, gender, age, political orientation...
- deception and obfuscation: what is this writer trying to hide?
Who uses it?

<table>
<thead>
<tr>
<th>Many different sectors:</th>
</tr>
</thead>
<tbody>
<tr>
<td>retail sector: for brand management, CRM, tracking trends etc.</td>
</tr>
<tr>
<td>health care monitoring: patient satisfaction with hospitals</td>
</tr>
<tr>
<td>equities trading/hedge funds: news, opinions, and emotion affect price</td>
</tr>
<tr>
<td>regulatory agencies of all types; the security sector</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What kind of texts?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anything: news (FT, WSJ), blogs, social media, company reports, CEO earnings transcripts.</td>
</tr>
<tr>
<td>How well does it work? Not perfectly, but good enough to provide useful information, and on a scale impossible for human analysts.</td>
</tr>
</tbody>
</table>
How does it work?

Sentiment is compositional and requires linguistic analysis

Sentence

Noun

verb phrase

Adverb

Verb phrase

Verb

Verb phrase

Verb

Noun

Our product

never

fails

to kill

bacteria

neutral  negative  positive
Many signals can be found via machine learning

Assemble a corpus of texts

- Annotate with the labels you want to predict (e.g. fear, surprise, anger, risk)
- Represent text as a feature vector
  
  Text = “Heidelberg is a city on the river Neckar…”

  Features = words, pairs of words, ...

  Heidelberg, is, a, city, on, ...

  Heidelberg_is, is_a, a_city, city_on, ...

- Vector = counts of these features in the text:

  | 0 | 2 | 1 | 3 | 0 | 0 | 6 | 1 | ...

- Learn weights for each feature from training data.

- Combination of weights $w$ and features $f$ gives a number representing the right label: e.g. $\sum_{i=1}^{n} w_i f_i$ (many variations on this theme)
Applications in investment

Markets are driven partly by sentiment and emotion.

**Sentiment analysis can have predictive value:**

- Tetlock et al. 2008 “More than words” demonstrates connection between negative news sentiment and earnings.
- Zhang and Skiena 2010: first paper to convincingly show that sentiment data can improve a trading strategy.
- Bollen 2012 claimed that ‘calmness’ predicted Dow Jones Industrial Index (much disputed!)
- Levenberg et al. 2014 showed that a combination of sentiment and numerical data could accurately predict whether the US Non Farm Payroll figures would go up or down.
Predicting the US non-farm payroll

Non-Farm Payroll: a monthly economic index that measures job growth or decay - a ‘market mover’.

Questions:
- Can we predict the direction of the NFP from financial indicators?
- Can we predict the direction of the NFP from sentiment in text?
- Does a combination of signals give better accuracy?
- We use a novel ‘Independent Bayesian Classifier Combination’ (IBCC) method to get best combination of individual classifiers.

\(^1\) Joint work with Oxford Man Institute of Quantitative Finance
Back tested over data from 2000-2012

Almost 10m words of text containing relevant keys:

<table>
<thead>
<tr>
<th>Source</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated Press</td>
<td>54K</td>
</tr>
<tr>
<td>Dow Jones</td>
<td>236K</td>
</tr>
<tr>
<td>Reuters</td>
<td>169K</td>
</tr>
<tr>
<td>Market News</td>
<td>385K</td>
</tr>
<tr>
<td>Wall Street Journal</td>
<td>76K</td>
</tr>
</tbody>
</table>

- and numerical data from many different sources, including:

  * Consumer Price Index (CPI)
  * Institute of Supply Management manufacturing index (ISM)
  * Job Openings and Labor Turnover Survey (JOLTS)
“The Governor noted that despite jobs being down, there was a surprising bright spot: construction added 1,900 jobs in November - its largest gain in 22 months.”

pos: 0.925, neg: 0.0, neut: 0.075, conf: 0.69

“When I drive down the main street of my little Kansas City suburb I see several dark empty storefronts that didn’t used to be that way.”

pos: 0.0, neg: 0.973, neut: 0.027, conf: 0.674

“We continue to fare better than the nation - our rate has been at or below the national rate for 82 out of the past 83 months - but we must also recognize that there were 10200 jobs lost at the same time.”

pos: 0.372, neg: 0.591, neut: 0.037, conf: 0.723
### Compositional sentiment features

<table>
<thead>
<tr>
<th>Stream</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.59</td>
<td>0.69</td>
<td>0.37</td>
</tr>
<tr>
<td>Dow Jones</td>
<td>0.45</td>
<td>0.44</td>
<td>0.25</td>
</tr>
<tr>
<td>Reuters</td>
<td>0.50</td>
<td>0.46</td>
<td>0.36</td>
</tr>
<tr>
<td>Market News</td>
<td>0.66</td>
<td>0.70</td>
<td>0.23</td>
</tr>
<tr>
<td>Other Sources</td>
<td>0.58</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>WSJ</td>
<td>0.44</td>
<td>0.63</td>
<td>0.53</td>
</tr>
<tr>
<td>IBCC</td>
<td>0.67</td>
<td>0.81</td>
<td>0.85</td>
</tr>
</tbody>
</table>

**A** = Majority sentiment label over all sentences per time slice.

**B** = Average of pos and neg percentage score over all sentences per time slice.

**C** = Trends, i.e. differences between B percentages in successive time slices.

These are “area under the receiver operating characteristic curve” scores: the probability of making a correct prediction of the NFP direction.
Text combined with numerical streams

Numerical data alone does a good job:

<table>
<thead>
<tr>
<th>Source</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPI</td>
<td>0.70</td>
</tr>
<tr>
<td>ISM</td>
<td>0.85</td>
</tr>
<tr>
<td>JOLTS</td>
<td>0.66</td>
</tr>
<tr>
<td>LFL</td>
<td>0.71</td>
</tr>
<tr>
<td>Combined</td>
<td>0.90</td>
</tr>
</tbody>
</table>

But a combination of text and numerical data is best:

<table>
<thead>
<tr>
<th>Source</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Series + Text Trends</td>
<td>0.91</td>
</tr>
<tr>
<td>Time Series + Text Averages</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Predicting risk: volatility in equity prices

Kogan et al 2009 “text regression”

- Used “10-K” annual reports (1996-2006) for 10,000+ US companies to predict volatility.
- Section 7 of these reports contains “forward looking content about market risk”.
- Training data: for each company, items of form \( \langle \text{data}, \text{label} \rangle \), where \text{data} = \text{report for year N} (represented as feature vector) + \log \text{volatility for year N}, and \text{label} = \text{volatility for year N+1}.
- Learn weights using “support vector regression”.
- Train on 5 years, test on 6th, where test means given data, predict label.
Effect of Sarbanes-Oxley act 2002:

Good results post 2002:

- Baseline: historical \( \text{volatility } N+1 = \text{volatility } N \) or GARCH(1,1) (same on this data)
- Measure mean squared error between predicted and true log volatilities: small is good. $p < 0.5$ not significant

<table>
<thead>
<tr>
<th>Year</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline:</td>
<td>0.1747</td>
<td>0.1600</td>
<td>0.1873</td>
<td>0.1442</td>
<td>0.1365</td>
<td>0.1463</td>
</tr>
<tr>
<td>Predicted:</td>
<td>0.1852</td>
<td>0.1792</td>
<td>0.1599</td>
<td>0.1352</td>
<td>0.1307</td>
<td>0.1448</td>
</tr>
</tbody>
</table>

Sarbanes-Oxley (post Enron) seems to have had the effect of making reports more truthful.

Changes in volatility signals

Phrases like “accounting policy” correlate with high volatility before 2002, low afterwards. “Mortgages” and “REIT” (real estate investment trust) change from low to high volatility indicators over this period.
Detecting risk in banking

The European Banking Union supervises 120+ banks, assessing risk using backward looking “key risk indicators” and survey data. Can text analytics add useful information?

Nopp and Hanbury 2015

- collect 500+ CEO letters and “outlook” annual statements from EU banks
- Get Tier 1 Capital Ratio (= Tier 1 Capital/Risk-weighted Assets) figures for end of period covered by reports.
- train a model to predict UP/DOWN movements in T1 on basis of positive/negative sentiment scores of reports.
- predictions quite accurate when aggregated, less accurate for individual banks
- The phrase “light touch regulation” changes status after 2007...
Results

Reasonable predictive value

<table>
<thead>
<tr>
<th></th>
<th>Correlation coefficient</th>
<th>Uncertainty</th>
<th>Negativity</th>
<th>Positivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier 1 + CEO letters</td>
<td>0.86</td>
<td>0.79</td>
<td>-0.69</td>
<td></td>
</tr>
<tr>
<td>Tier 1 + outlooks</td>
<td>0.85</td>
<td>0.89</td>
<td>0.12</td>
<td></td>
</tr>
</tbody>
</table>

Predicting the direction of Tier 1

- Train a support vector machine classifier on aggregated data.
- Predict whether banks will increase/decrease their Tier 1 ratio.
- Correct for 12 out of the 13 years tested (2002-2013).

Conclusion

These predictions are not very fine-grained, but the results demonstrate that there is a usable signal in such texts.
Linguistic characteristics of deceptive vs. truthful text

Several studies have found that:

**Liars**
- Liars tend to use more emotion words, fewer first person (“I”, “we”), more negation words, and more motion verbs (“lead”, “go”). (Why?)
- Possible that liars exaggerate certainty and positive/negative aspects, and do not associate themselves closely with the content.

**Truthtellers**
- Truthtellers use more self references, exclusive words (“except”, “but”), tentative words (“maybe”, “perhaps”) and time related words.
- Truthtellers are more cautious, accept alternatives, and do associate themselves with the content.
Lying CEOs

Larcker and Zakolyukina (2012) looked at the language used by CEOs and CFOs talking to analysts in conference calls about earnings announcements.

Looking at subsequent events:

- discovery of ‘accounting irregularities’
- restatements of earnings
- changes of accountants
- exit of CEO and/or CFO

- you can identify retrospectively who was telling the truth or not.

This gives us a corpus of transcripts which can be labelled as ‘true’ or ‘deceptive’: training data.
Detecting deception

**It can save your money**

- Training a classifier on features like those just described, Larcker and Zakolyukina were able to get up to 66% prediction accuracy.
- Building a portfolio of deceptive companies will lose you 4 to 11% per annum...

**Verisimilitude**

- We (TheySay) have build a general ‘verisimilitude’ classifier which looks out for linguistic indicators of deception...
- ... and also measures clarity and readability (vs. obfuscation, hedging, etc.)
- We tried this on speeches by the 4 UK party leaders prior to the 2015 election, and guess what?
Bankers’ speeches

Data
We gathered speeches from the last two years by:
- Mario Draghi (European Central Bank)
- Mark Carney (Bank of England)
- Jens Weidmann (German Federal Bank)
- Francois Villeroy de Galhau (Bank of France)
- Ignazio Visco (Bank of Italy)
- Yannis Stournaras (Bank of Greece)

Analysis
We processed them with TheySay’s PreCeive API (www.theysay.io) to detect sentiment, emotion, and verisimilitude, and visualised the results using SAP’s Lumira.
Carney and Draghi often move together (positivity)
Who is the most positive and negative?
Who is the most ashamed or surprised?
Which emotions are strongest?
Who can you trust least?
Conclusions

Text is an under-used resource

- Text is free - no costly surveys or data gathering needed.
- Text analytics can uncover useful signals.
- These signals may have predictive power...
- ... especially when combined with other numerical indicators.
- There are many potential applications in finance.
- Using text analytics for risk detection and prediction may be a useful tool in maintaining financial stability
References


Zhang, W., Skiena, S, 2010, *Trading strategies to exploit blog and news sentiment*, In Fourth Int. AAAI Conf. on Weblogs and Social Media, 375-378.