Prof. Dr. Alexander Hillert

Text Mining with Python

2018 Spring Conference of E-Finance Lab and IBM Deutschland
February 1, 2018, Goethe-University Frankfurt
• In the US, mutual fund companies send semiannual reports to their investors (Investment Company Act 1940).

• Besides quantitative information, the reports include a letter from the fund’s management. The letter can cover different topics, e.g.,
  o the fund’s performance relative to a benchmark
  o discussion of economic and market conditions
  o advertising the fund

• Research questions
  1. Does the writing style of letters have an impact on capital flows?
  2. Does it predict future fund performance?
“Economies across the globe were buffeted by the severe credit contraction that destabilized financial markets and led to bank closures, failures of financial services companies, and massive government bailouts. […] The Dow Jones Industrial Average was down 31.9 percent in 2008, its worst year since 1931. The Standard & Poor's 500 index dropped 22 percent in the fourth quarter alone, and 37 percent for the full year - its worst performance since 1937. The NASDAQ market had its worst year ever.”

Negativity (Loughran/McDonald word list): 4.33%

American Century Quantitative Equity Fund
(Dec. 2008; 6 month return -33.84%)

“As a company, American Century Investments® is well positioned to deal with market turmoil.[…] In addition, our actively managed, team-based approach allows our portfolio teams to identify attractive investment opportunities regardless of market conditions. Our seasoned investment professionals have substantial experience and have successfully navigated previous market crises.”

Negativity (Loughran/McDonald word list): 2.78%
Objective: analyze the writing style of shareholder letters for U.S. equity funds.

Key steps
1. Obtain shareholder reports
2. Identify the letter from the fund’s management
3. Determine tone (negativity)
4. Match with fund data and run the analysis
Obtaining shareholder reports (1)

The EDGAR system of the Securities and Exchange Commission (SEC)
https://www.sec.gov/search/search.htm

Search by fund name or Central Index Key (CIK) = unique investment management company ID
(Semi-annual) shareholder reports

- N-CSR (annual) and N-CSRS (semi-annual) filings
- Link to the first example: https://www.sec.gov/Archives/edgar/data/1003859/000093506909000687/0000935069-09-000687-index.htm
- Submissions consist of the main document, exhibits, and the complete submission file.
- Links to the complete submission files are reported in the SEC’s daily/quarterly submission overviews
  - Example: https://www.sec.gov/Archives/edgar/full-index/2008/QTR4/
  - select the “form.idx”
(Semi-annual) shareholder reports – cont.

- List of submissions from 2008 Q4
  - Column 1: filing type → select N-CSR and N-CSRS filings
  - Column 3: CIK → investment management company ID
  - Column 4: filing date → date shareholders receive the letter
  - Column 5: link → where to find the filing on the EDGAR server

<table>
<thead>
<tr>
<th>CIK</th>
<th>Filing Type</th>
<th>CIK</th>
<th>Filing Type</th>
<th>CIK</th>
<th>Filing Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-CSR</td>
<td>ADVISORS SERIES TRUST</td>
<td>1027506</td>
<td>2008-12-09</td>
<td>edgar/data/1027506/0001193097-08-000158.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>RESIGS VALUE FUND INC</td>
<td>1052165</td>
<td>2008-11-06</td>
<td>edgar/data/1052165/0001144204-08-001384.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>ATG Retirement CO II</td>
<td>1062374</td>
<td>2008-11-07</td>
<td>edgar/data/1062374/0001193097-08-022456.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>AIM COUNSELOR SERIES TRUST</td>
<td>1112996</td>
<td>2008-11-07</td>
<td>edgar/data/1112996/0000950129-08-005502.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>AIM Core Allocation Portfolio Series</td>
<td>1335263</td>
<td>2008-11-07</td>
<td>edgar/data/1335263/0000950129-08-005502.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>AIM INVESTMENT SECURITIES FUNDS</td>
<td>642790</td>
<td>2008-10-09</td>
<td>edgar/data/642790/0000950129-08-005502.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>AIM TREASURERS SERIES TRUST</td>
<td>828806</td>
<td>2008-11-07</td>
<td>edgar/data/828806/0000950129-08-005503.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>ALGER CHINA-US GROWTH FUND</td>
<td>1219064</td>
<td>2008-12-29</td>
<td>edgar/data/1219064/0001144659-08-079499.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>ALGER FUNDS</td>
<td>3521</td>
<td>2008-12-29</td>
<td>edgar/data/3521/0001193097-08-078405.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>ALGER FUNDS II</td>
<td>92751</td>
<td>2008-12-29</td>
<td>edgar/data/92751/0001193097-08-078494.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>ALGER INSTITUTIONAL FUNDS</td>
<td>921145</td>
<td>2008-12-29</td>
<td>edgar/data/921145/0001193097-08-078491.txt</td>
<td></td>
</tr>
<tr>
<td>N-CSR</td>
<td>ALLENECKEINBLINDE STYLE SERIES INC</td>
<td>1172221</td>
<td>2008-11-06</td>
<td>edgar/data/1172221/0001193097-08-078493.txt</td>
<td></td>
</tr>
</tbody>
</table>
(Semi-annual) shareholder reports – cont.

• How to automatically download files → Python

• Python
  o Programming language
  o Versions 2.7 (no longer supported but still commonly used) and 3.6
  o My recommendation: Anaconda ("Spyder")
    – nice programming environment (editor, variable explorer, etc.)
    – Available as freeware at: https://www.anaconda.com/download/
(Semi-annual) shareholder reports – cont.

- Python can easily handle csv files.
- Input: csv file with 5 examples.

- Download the 5 examples using the urllib package.

- Program structure:
  1. Loop over all lines in the csv file.
  2. Split each line into columns (=variables).
  3. Use the link to download the file.
  4. Save the file as txt on local hard drive.
Identify the letter from the fund’s management

How to identify the letter?
• Complete submission files contain tons of html code → unreadable for humans
• Delete html code using Python’s BeautifulSoup package

• The SEC filing is the full shareholder report incl. tables, figures, disclaimers, etc.
• How to identify the letter?
  o Item or section numbers? → not possible, as no fixed position and/or section
  o Phrases/words that typically indicates the beginning and ending of letters
  o What if there is no start or no end phrase?
    – Start from the top of the report
    – Stop at sections that are typically placed below the letter (e.g. expense example,
      performance figure)
    – Delete text that is not part of a letter (e.g. table of contents, SEC header)
How to measure tone?

• ‘bag of the word approach’, dictionary approach
  o Count the number of words that are included in a specific word list (negative, positive,…)
  o Calculate the fraction of these words by dividing by the total number of words.

Positive vs. negative letters

• Measuring negativity works best
  o Unambiguous meaning (“not as good as expected ” vs. “not as terrible as expected”)
  o Investors’ reaction best explained by negativity (e.g. Tetlock, 2007)
Determine tone (negativity) (2)

Which dictionary?
- General Inquirer Harvard IV-4 psychosocial dictionary
- 77 dictionaries
  - Negative: 2,291 words
  - Positive: 1,902 words
  - Passive: 911 words
  - Pleasure: 168 words
- The dictionaries are available at: http://www.wjh.harvard.edu/~inquirer/homecat.htm
Determine tone (negativity) (3)

1) Two large valence categories (new)

- **Positiv**: 1,015 words of positive outlook. (It does not contain words for *yes*, which has been made a separate category of 20 entries.)
- **Negativ**: 2,291 words of negative outlook (not including the separate category *no* in the sense of refusal).

We plan to develop further subcategories of these categories.

**Harvard IV-4 categories:**

2) "Osgood" three semantic dimensions.

These categories reflect Charles Osgood’s semantic differential findings regarding basic language universals.

An earlier version had three different “intensity” levels for each category, but these were combined. A word may be more than one dimension, if appropriate. For example, “celebration” in the Harvard dictionary is **Positiv/Pstv/Affil/ActiveRitual**

- **Pstv**: 1,045 positive words, an earlier version of *Positiv*.
  - A subset of 557 words are also tagged *Affil* for words indicating affiliation or supportiveness.
- **Ngtv**: 1,150 negative words, an earlier version of *Negativ*.
  - A subset of 833 words are also tagged *Hostle* for words indicating an attitude or concern with hostility or aggressiveness.
- **Strong**: 1,002 words implying strength.
  - A subset of 689 words are tagged *Power*, indicating a concern with power, control or authority.
- **Weak**: 755 words implying weakness.
  - A subset of 284 words are also tagged *Submit*, connoting submission to authority or power, dependence on others, vulnerability to others, or withdrawal.
- **Active**: 2,045 words implying an active orientation.
- **Passive**: 911 words indicating a passive orientation.
Loughran and McDonald (2011)

- Almost 75% of the words in the Harvard IV psychosocial dictionary are misclassified in business contexts.
  - Neutral meaning
    - tax, costs, capital, expense, liabilities, service → noisy
  - Systematic bias
    - Capital → banking and insurance
    - Crude → oil industry
    - Mine → precious metals and coal
    - In the 1999 10-K of Coeur d’Alene Mines Corporation the word mine accounts for 25% of all negative words.
Loughran and McDonald propose 6 lists:

- **Negative:** 2,337 words
  - 1,121 overlap with Harvard negative
  - Restated, litigation, termination, unpaid, investigation, serious, deterioration, etc.
- **Positive:** 353 words
  - Achieve, efficient, improve, profitable, etc.
- **Uncertainty:** 285 words
  - General notion on imprecision, not only risk
  - Approximate, depend, fluctuate, indefinite, uncertain, etc.
- **Litigious:** 731 words
  - Claimant, deposition, testimony, etc.
- **Modal strong:** 19 words
  - Always, highest, must, etc.
- **Modal weak:** 27 words
  - Could, depending, might, etc.
Determine tone (negativity) (6)

• How were the word lists created?
  o Create a list of all words contained in the 10-Ks.
  o All words that occur in at least 5% of the filings are manually examined.

• Word lists can be downloaded from Bill McDonald’s webpage
  http://www3.nd.edu/~mcdonald/Word_Lists.html

• They use inflected versions of the word lists
  o Accident, accidental, accidentally, and accidents
  o The expand the original Harvard negative list from 2,005 (word stem) to 4,187 words (incl. inflections)
  o Problem with stemming: odd and odds

• They control for negations when using the positive word list
  o Simple negation is taken to be observations of one of six words (no, not, none, neither, never, nobody)
    occurring within three words preceding a positive word.
Determine tone (negativity) (7)

Outlook – other dimensions of writing style

• Forward-looking information
• Blaming
• Hard facts vs. vague talk: numbers

• Readability and complexity
  o Word-per-sentence
  o Document length
  o Financial terms
  o Fog index
Main Finding – Impact of Letter Tone on Fund Flows

Cumulative net flows for up to 15 days after the filing date of the shareholder letter
Letter Tone and Fund Flows – Monthly Analysis

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMD^-</td>
<td>-0.166***</td>
<td>-0.135**</td>
</tr>
<tr>
<td></td>
<td>(-3.02)</td>
<td>(-2.32)</td>
</tr>
<tr>
<td>LMD^-_adj.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Return Rank</td>
<td>0.008</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>(1.33)</td>
<td>(0.96)</td>
</tr>
<tr>
<td>Return Rank^2</td>
<td>0.012**</td>
<td>0.014**</td>
</tr>
<tr>
<td></td>
<td>(2.06)</td>
<td>(2.26)</td>
</tr>
<tr>
<td>Expense Ratio</td>
<td>0.028</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Fund Risk</td>
<td>-0.128***</td>
<td>-0.106**</td>
</tr>
<tr>
<td></td>
<td>(-3.01)</td>
<td>(-2.28)</td>
</tr>
<tr>
<td>LN(Words)</td>
<td>-0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(-0.06)</td>
<td>(-0.25)</td>
</tr>
<tr>
<td>LN(WPS)</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>(0.76)</td>
<td>(1.25)</td>
</tr>
<tr>
<td>Time Difference</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>Flow Segment</td>
<td>0.616***</td>
<td>0.762***</td>
</tr>
<tr>
<td></td>
<td>(5.11)</td>
<td>(6.14)</td>
</tr>
</tbody>
</table>

Dependent variable: net fund flow in the Filing Month

One std. dev. increase in LMD is associated with $2.95 million (0.21%) lower flows → economically meaningful
Shareholder Letter Tone and Fund Performance

**Dependent variable:** performance in the six months after the Filing Month

→ Prediction of future fund performance

**Result:** No significant relation between tone and future performance
• Mutual fund investors strongly react to the writing style of shareholder letters.
• A more negative tone leads to lower fund flows.
• Investment companies should pay attention to the wording of shareholder letters and avoid a too pessimistic language.
• Tone is not predictive for future fund performance.
Thank you very much for your attention!