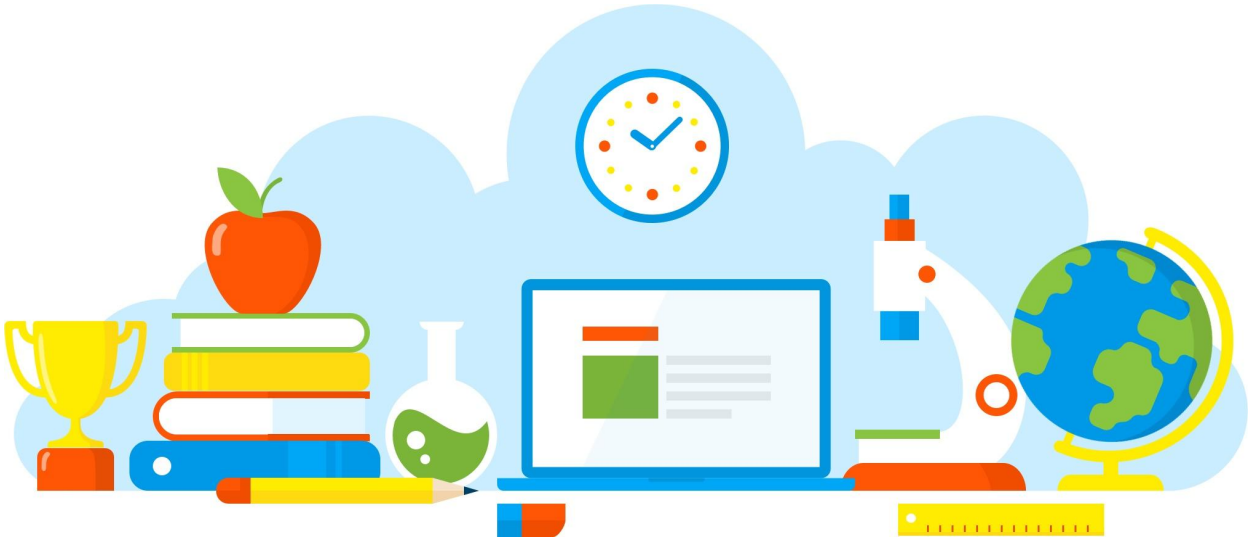


PROJECT REPORT

Investment Analysis with Natural Language Processing (NLP) techniques to exploit Sentiment for Financial Analysis and Validation of a Hypothesis.

SUHAS S

01/06/2021

Introduction

-**Natural Language Processing (NLP)** is set of techniques which helps us gain insights from text

- Applications of NLP in Finance include

- **Context** (e.g. using Topic Modelling to establish context of news articles, firm announcements, business descriptions, etc)
- **Compliance** (e.g. detecting insider trading via emails / chat transcripts)
- **Quantitative Analysis** (e.g. trading strategies using sentiment analysis)

-**Sentiment analysis** (in Finance) involves quantifying and exploiting 'sentiment' / 'emotions' for investment.

-There are broadly 2 ways of estimating 'sentiment':

1)**Lexicon / Dictionary based approach** : We start with a 'prior' on what constitutes words relating to 'sentiment' We then estimate 'sentiment' ($\boldsymbol{\phi}$) as a function of the (cleaned) words ($\boldsymbol{W}*$) in a given document ($\mathbf{d}$) which belong to a sentiment language ($\boldsymbol{\varphi}$).

2)**"Machine learning" approach:** We start with a subsample of the 'corpus' C which displays 'sentiment' ($\boldsymbol{\phi}$).We then apply 'machine learning' (e.g. classification) algorithms to categorise / classify other sample text on its level of 'sentiment' ($\boldsymbol{\phi}$).

Methodology

5 Step Sentiment Analysis Process

1. Creating a testable hypothesis
2. Extracting relevant data (i.e., a 'corpus')
3. Cleaning the text data.
4. Estimating the sentiment measure(s).
5. Testing / validating the hypothesis.

Hypothesis

H0: Returns of firms with stronger net positive tone are statistically equal to the returns of firms with weaker net positive tone.

H1: Returns of firms with stronger net positive tone are statistically greater than the returns of firms with weaker net positive tone.

Dataset

Testing this particular hypothesis requires 2 sets of data:

1)**Returns data** (obtained from stock prices) -

Source:Google Finance (via Google Sheets) and Pandas Datareader

2) **Tone data** (obtained by estimating tone using words from a corpus)-

Source: 10-K Annual Reports* (SEC filings) Particularly, the Management Discussion & Analysis (MD&A) section – Item 7

Cleaning Text Data

'Cleaning' text data involves transforming raw text into a format suitable for textual analysis.

"Word Tokenization" involves transforming sentences into lists of words.

3 Step Process after Word Tokenization:

- Remove numbers, symbols, and non alphabetic characters.
- Harmonise letter case (e.g. all lower case) Ø
- Remove the most common words ("stopwords")

Depending on the hypothesis, text cleaning can also include:

- Stemming or Lemmatizing
- Removal of the most common words within the corpus

Sentiment Estimation

We use Lexicon Approach to estimate Sentiment

3 Step Process:

- Start with a 'prior' on a sentiment language
- Transform raw text into a form suitable for textual analysis (i.e., clean the text)
- Estimate sentiment as a function of the words belonging to the sentiment language.

Consider a sample of J firms, for which we have a Corpus C comprising of d documents over t time such that:

$$\mathcal{C} = \{d_{1t1}, d_{1t2}, \dots, d_{(J-1)T}, \dots, d_{J(T-1)}, d_{JT}\};$$
$$\forall\, j \in \{j_1, j_2, \dots, J\}$$
$$\mathcal{D} = \sum d \equiv \mathcal{C}$$

Each document d has $\mathcal{W}_d$ words.

$$\mathcal{C} = \begin{pmatrix} d_{1t1} \\ d_{1t2} \\ d_{2t1} \\ d_{2t2} \\ d_{Jt1} \\ d_{Jt2} \end{pmatrix} \equiv \begin{pmatrix} [w_1, w_2, \dots, \mathcal{W}]_{d_{1t1}} \\ [w_1, w_2, \dots, \mathcal{W}]_{d_{1t2}} \\ [w_1, w_2, \dots, \mathcal{W}]_{d_{2t1}} \\ [w_1, w_2, \dots, \mathcal{W}]_{d_{2t2}} \\ [w_1, w_2, \dots, \mathcal{W}]_{d_{Jt1}} \\ [w_1, w_2, \dots, \mathcal{W}]_{d_{JT}} \end{pmatrix}; \neg\square(\mathcal{W}_{d_{jt}} = \mathcal{W}_{d_{kt}})$$

Sentiment ϕ can be estimated for each document d as a function of the 'cleaned words' $W_d{}^*$ which belong to a 'sentiment language' $\boldsymbol{\psi}$.

If the document d is at the firm level (over different time periods), this estimate can be proxied as the level of sentiment ϕ of firm J at time t.

$$\phi_{s,jt} = f\left(\mathcal{W}^*_{d\,jt}, \psi_s\right)$$

Where:

$\phi_s = s$ Sentiment Estimate for firm j at time t;
$s \in$ {'positive', 'negative', …}

$\mathcal{W}^*_{d_{jt}}$ = Cleaned words within document d for firm j at time t

ψ_s = Sentiment language for a specific sentiment s

Sentiment could be estimated as:

- The frequency counts of 'sentiment language' used in a document.
- The number of unique times a 'sentiment language' was used in a document.
- The proportion of sentiment language used.

Here we use Proportion Approach:

'Sentiment' ϕ as the proportion of the frequency counts of "sentiment language"***ψ*** used, relative to the total number of (cleaned) words ∑W∗ in a given document **d**.

$$\phi_{s,jt} = \frac{\sum 1_{w^*_{d_{jt}} \in \psi_s}}{\sum w^*_{d_{jt}}}$$

Where:

$\phi_{s,jt} = s$ Sentiment Estimate for a firm j at time t

$w^*_{d_{jt}}$ = Cleaned words in a document d for firm j at time t

$\psi_s = s$ Sentiment Language / Lexicon / Dictionary

Estimating Net Positive Tone

$$\phi_{NPT,jt} = \frac{\phi_{pos,jt} - \phi_{neg,jt}}{\phi_{pos,jt} + \phi_{neg,jt}}$$

Where:

$\phi_{NPT,jt}$ = Net Positive Tone Estimate for a firm j at time t

ϕ_{pos} = Positive Sentiment Estimate for a firm j at time t

$$\phi_{pos} = \frac{\sum 1_{w^*_{d_{jt}} \in \psi_{pos}}}{\sum w^*_{d_{jt}}}$$

-Lexicons for positive and negative sentiment can be obtained from Python's NLTK framework via opinion_lexicon

-Here we estimate sentiment ϕ for all firms by iterating over each file. We do not use Document Term Matrix because the data is small and can be iterated quickly.

Merging Tone and Returns Data

-We now have the Returns Data and Tone Data. Testing the hypothesis will require working with both datasets in tandem.

-Merging multiple datasets requires “common identifiers”.In our case, we’re working with firm and date level identifiers for each dataset. But they are not common

-The datasets have different time frequencies in that we have daily Returns Data and annual Tone Data. There are 250 trading days each year, and typically only 1 filing (10-K) each year.

-Merging the datasets on the existing date level (price date, filing date) will result in approximately 249 missing observations per firm per year.

-We need to convert the annual Tone Data into daily data. We’ll do this by ‘filling’ the missing Tone observations.

-Annual data can be converted to daily by: Assuming it remains unchanged across all days in a given year.

-The idea that tone doesn’t change across all days in a given year relies on the premise that the management “set the tone” for the year.

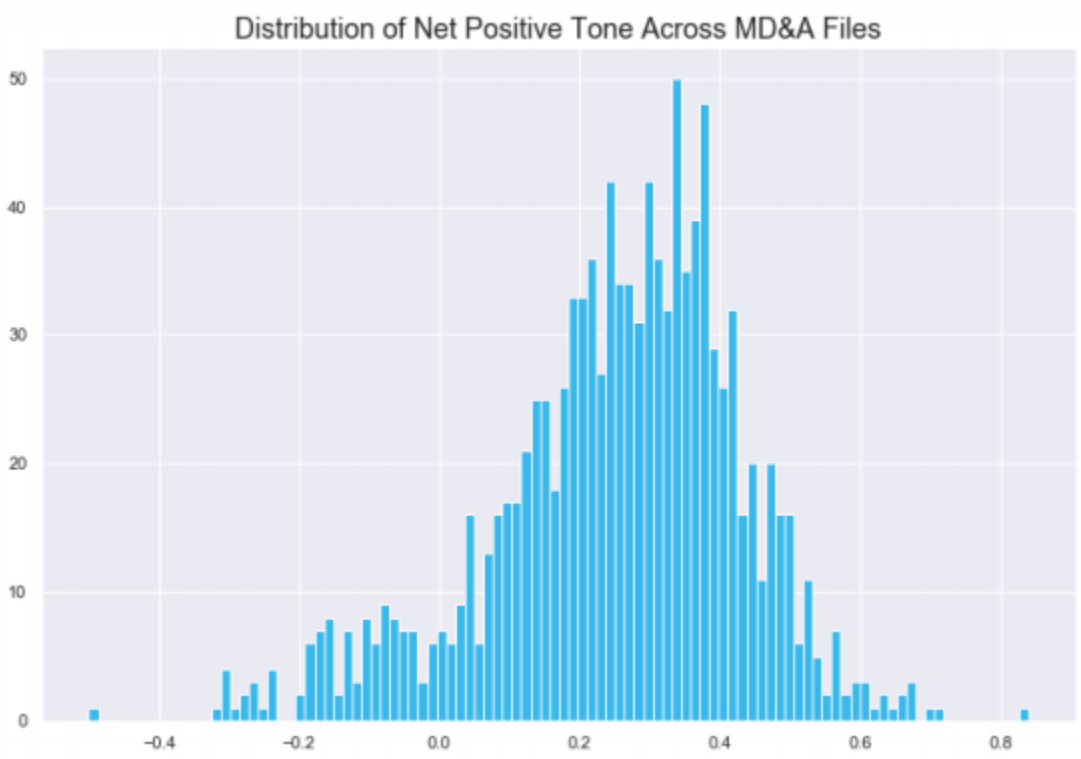
-We can merge data by using the .merge method that’s built into Pandas. The .merge arguments ‘on’, ‘left_on’, ‘right_on’ allow us to specify common identifiers across datasets. The argument ‘how’ specifies inner, outer, left, or right merges.

-Missing observations can be filled (forward only) by using ffill(), grouping by a firm level identifier if necessary.

Estimating Sentiment Portfolio Returns

-Testing the hypothesis requires creating portfolios of firms with stronger and weaker net positive tone. Then statistically testing the difference between the returns on both portfolios.

-We can identify firms with stronger and weaker net positive tone based on their relative value of net positive tone.



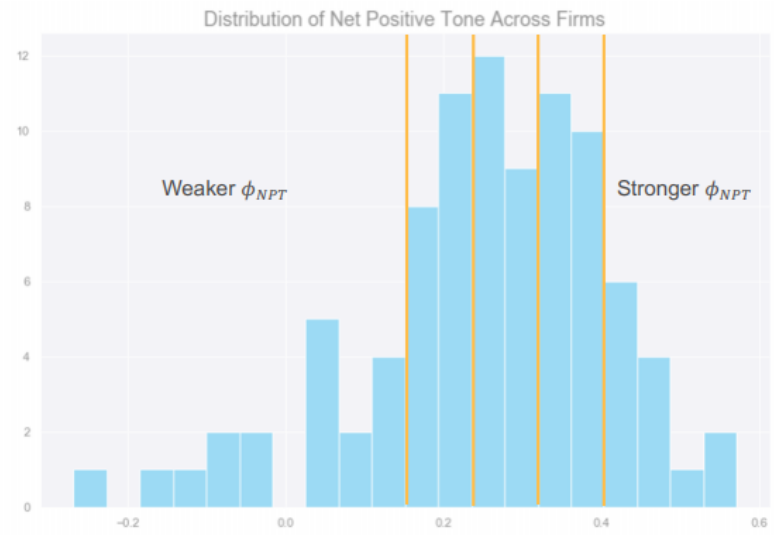
-At each date d, we sort all firms into “buckets” based on their ϕ values.

-It’s important to ensure there are at least approximately the same number of firms in each bucket. Otherwise, the results may be influenced by the effects of diversification.

-After grouping firms into stronger and weaker net positive tone “buckets”, we can estimate the average return for each bucket each day.

-We can sort firms into sentiment buckets using Pandas ‘qcut’.

-The returns of equal weighted sentiment portfolios can be estimated by calculating the simple mean for each “sentiment bucket”.



Testing & Validating the Hypothesis

-Returns of firms in Quintile 5 (strong ϕ) and Quintile 1 (weak ϕ).

-We can test the core hypothesis by statistically comparing the mean return of both portfolios. The mean difference can be tested for statistical significance using a 't-test'.

-A 't-test' (aka "Student's t-test") is a statistical test to compare the means of two groups. The null hypothesis of the t-test is that the means of 2 groups are identical.

$$tStat = \frac{\bar{x}_{d_{j,k}}}{\hat{\sigma}_{d_{j,k}}/\sqrt{n}}$$

Where:

$\bar{x}_{d_{j,k}}, \hat{\sigma}_{d_{j,k}}$ = *Mean and standard deviation of the difference between the values of group j and group k*

n = *Number of observations*

-The greater the t-statistic, the more likely it is that our (alternative) hypothesis holds. We compare the t-statistic to a 'critical value' and reject the null if the t-statistic is greater than or equal to the critical value.

-The 'critical value' depends on what level of significance one wants.

-A good heuristic / rule of thumb when working with financial data is to use a 'critical

value' of 2. I.e., reject the null if the t-statistic is greater than or equal to 2.

-Generally, a t-statistic that is greater than or equal to 2 implies that the result we see holds at least 95% of the time.

RESULTS

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
t_stat	-0.786255	0.687947	2.429968	0.687843	-0.459497	1.041805	0.424689	-0.328352	-0.282712	-1.228795	0.351794	0.492429	-1.266477

. - Our Core Hypothesis does not hold in most of the years.

Thus we can conclude that:**Generally Returns of firms with stronger net positive tone are statistically equal to the returns of firms with weaker net positive tone.**

However we can observe the t_stat value >2 in the year 2007. 2007 was the year of Global Financial Crisis.

Thus we can conclude that there is evidence to suggest that :**Returns of firms with stronger net positive tone are statistically greater than the returns of firms with weaker net positive tone during times of Crisis**

Thank You.