News-induced Style Seasonality

C.Gillain[a], A. Ittoo[a], L. Lambert[a]

[a]HEC Liège, ULiège

This version is a very preliminary draft. Please do not cite.

June 1, 2019

Abstract

This paper posits a new methodological approach to test how specialized media could influence the information transmission channels towards investors. We contribute to the literature on the role of media on investor limited attention, on seasonal effects in market anomalies and on the impact of news on market anomalies. Our approach is somewhat different from the current literature as we determine whether we can detect any seasonality in the news coverage of recommendations, analyses or opinions on investment styles provided by specialized press to institutional investors. Our paper not only contributes to the literature on market anomalies and seasonality effects in financial markets but also aligns itself with a new strand of research involving the application of text mining in finance. First, our text corpus gathers articles from specialized press targeting institutional investors. Such a corpus is unique and has never been investigated. Second, we build our own dictionaries from several statistical methods to extract style information from news flow. The method is innovative and our study is the first to investigate the seasonality in the underlying information channel. At this stage, the paper is mainly methodological and centered on small and large styles. Results will be extended to other investment styles in the near future and completed with statistical test of cyclicality and trend analysis.

1 Introduction

Mispricing with regard to the original Capital Asset Pricing Model (Sharpe (1964), Lintner (1965)) due to factors such as the size and value effects has been documented in the US stock market since the early 1980s. Nowadays, the pace of discovery of new abnormal effects has sharply increased with a record of more than 300 anomalies identified in 2016 (see Harvey, Liu and Zhu (2016), Green, Hand and Zhang (2013) and Hou, Xu and Zhang (2018).)

Recent papers have studied the impact of news diffusion on these market anomalies. Engleberg et al. (2018) have shown that market anomaly returns are concentrated on corporate news events. Barber and Odean (2008) have shown that investors are net buyer of attention grabbing stocks: stocks in the news will experience significant abnormal returns. Our thesis is that if news conducting information on market anomalies exhibit seasonality in coverage, anomaly returns should also exhibit this seasonality. Our paper therefore contributes to the literature on market anomalies by revisiting the seasonal effects in market anomalies induced by news diffusion. We rely on textual mining classifiers to infer news coverage on market anomalies among institutional media. The aim is to study the news seasonality and their impact on anomalies returns.

The motivation to look through textual content in finance is that it brings additional information to traditional quantitative data. This incremental information can be related to prospects from management (MD&A in SEC filings), qualitative appraisal in the financial press (e.g. Wall Street Journal, Dow Jones News) or investor sentiment on social network (e.g. Tweeter, StockTwits). An important volume of literature explores the interaction between tone in textual content and market reactions (Antweiler and Franck (2004), Tetlock (2007), Tetlock, Saar-Tsechansky and Macskassy (2008), Li (2010), Garcia (2013), Renault (2017)).

Tone content evaluation is achieved through word (or group of words) classification in two categories : positive and negative. Methods varies from dictionary-based analysis to statistical classifiers. Loughran and McDonald (2011) construct a financial dictionary by examining words occurring in SEC's 10-K filings. They prove that many words from Harvard General inquirer are misclassified in the financial context. Henry and Leone (2016) confirm that domain-specific word lists such as LM dictionary better measure tone in financial disclosures. Moreover, Li (2010) was already advocating that in absence of an adapted dictionary, researchers should rely on machine learning classifier instead of using general word lists. Finally, Henry and Leone (2016) compare word-frequency tone measurement to the Naïve Bayesian classifier developed in Li (2010) : they find that the two methods exhibit similar performance. Textual analysis goes however beyond tone extraction and is a powerful tool that could fill the gap for quantitative "missing" data or unsolved research questions.

Our main contribution is to empirically investigate the news distribution of anomalies. We analyze specialized financial articles using textual mining techniques to detect information relative to anomalies. From subsets of annotated data, we automatically construct anomalies-dedicated lexicons (i.e. dictionaries). The resulting lexicons are domain-specific (i.e. anomaly-specific) and directly adapted to the sample studied.

We then extract news with anomaly content based on lexicon terms detection.. We aggregate this coverage score (daily, weekly and monthly) through our existing news flow, investigating presence of seasonality patterns. Finally, relying on subsets of annotated data allow us to implement machine learning classifier using different techniques : Naïve-Bayesian and Support Vector Machine (Das and Chen 2007). We can therefore select the most accurate method (between lexicon-based or classifiers) to appropriately detect anomalies information.

To conduct our analysis, we select specialized media content targeting institutional investors : Pension Funds and Institutional (PF&I) magazines. We expect this specialized financial press to convey information related to anomalies. We focus on institutional investors because transaction volume from these actors currently overtake retails transactions. We create a new database gathering news from (PF&I) magazines. To the best of our knowledge, this constructed dataset is unique in the literature and no commercial entity provide similar data.

Our first results concern the controversial “January effect”. While some evidence shows that this anomaly is persistent and is particularly high for some market anomalies such as the size effects (see Reinganum (1981), Roll (1981), Keim (1983), and Asness et al. (2015)). Other claim that it has completely disappeared after the Tax Reform Act of 1986. We revisit this January effect by investigating whether there exist seasonal patterns within the information diffusion on small and large capitalization stocks.

Our intended contribution to the literature is twofold : to identify the most appropriate method to automatically acquire small and large capitalization lexicons and to explore the potential existence of seasonal pattern in news related to this investment styles.

2 Literature Review

Analyzing the seasonal effects in market anomalies through the seasonal coverage of recommendations, investment analyses and opinions found in specialized press relies on the assumption that expert recommendations do influence investors. Kumar (2009) shows that bullish and bearish signals to investment styles of individual investors is significantly related to the information content of expert newsletters in the previous month. Other papers on the impact of expert advice can be found in Goetzmann and Massa (2003) and Fisher and Statman (2000)

The current version of the paper focuses on the size effect and small versus large investments. Seasonal patterns have been shown to be particularly concentrated among small capitalization (see Reinganum (1981), Roll (1981), and Keim (1983), Asness et al. (2015)). This "*turn-of-the-year effect*" implies that the size effect is weak over time except during the month of January. In particular, Rozeff and Kinney (1976) defines the January effect as the abnormal return phenomenon exhibited by stocks in January. Two main hypotheses are put forward to explain this effect: i) tax-loss-selling among retail investors (Ritter (1988)) : individuals sell losers in December to realize tax losses and wait until January to reinvest; ii) window-dressing by institutional investors : managers tend to buy winners and sell losers during December to present attractive year-end portfolios (Lakonishok et al. (1991)). This anomaly still persists, challenging the efficient market hypothesis (Haugen and Jorion (1996), Haug and Hirschey (2006)).

More recently, Sikes (2014) documents that institutional investors demonstrate both tax-losses and window-dressing incentives (varying with their profile and clientele). For example, high net-worth advisers are highly tax-sensitive while pensions funds have strong window-dressing incentives. She concludes that institutional investors contribute to this turn-of-the-year effect whatever their incentives (tax-sensitive institutional investors seem to have more powerful impact). This paper confirms again the persistence of the January effect from 1987 to 2010.

The principle of attention constraints developed by Kahneman and Tversky [1979]) gives a framework to revisit the impact of news on anomaly seasonality. Recent empirical studies on the impact of news have grounded their results into this framework. Fedyck (2018) shows that limited attention is responsible of gradual information diffusion rather than heterogenous beliefs. Barber and Odean (2008) and Odean (2009) show the impact of limited attention to style investing. On the contrary, Merton (1987) establishes the "investor recognition hypothesis" or the fact that individual investors will hold an under-diversified portfolio and follow a few number of stocks they know. Even in this situation, our framework could stay valid as we focus at the impact of aggregate style information spread in the news rather than information on individual firms.

3 Data

Several media groups deliver financial information through their portfolio of magazines. Their general mission statement includes production of trusted information targeting financial decision makers (such as investment managers, advisers, pension trustees and financial intermediaries). We identified nine different Pension Funds and Institutional (PF&I) magazines from five well-known financial media groups. Table 1 presents a description of those media and magazines covered in this study. Although the audience differ through our PF&I magazine sample, the majority of readers are institutional investor, asset manager, advisers and consultants.

News released in PF&I magazines cover a broad range of topics such as macro-economics, market analysis, expert insight, portfolio strategic allocation. In any case, equity related information is essentially discussed at category-based level, i.e. emerging market and developed market equities, equity investment style or industry sectors. This news structure represents a great advantage : we can directly extract style information at an aggregate level. This approach is significantly different from the information we could get from conventional newspapers such as Reuters, Bloomberg, Wall street Journal, and Dow Jones News. Those media are featuring news on individual firms rather than opinions, recommendations or portfolio allocation to equity style investment. One could have extracted information on a company level and aggregate the individual companies' score with regard to styles portfolio. Two caveats are in order: first, this would assume investors are able to process all this information, second style rankings among companies are time-varying.

We use *Scrapy* and *Beautiful Soup* library from python to collect news content from PF&I magazines' websites. Each news collected is converted in a plain text file with related information : magazine, date, title, author, section and textual content (See Appendix 1) . We gather more than 100,000 news from the web from January 1996 to June 2018. This non exhaustive sample seems sufficient to explore our research question. Table 2 presents descriptive statistics of our database.

Table 1
Description of media Group and magazines included in PF&I database

Media Group	Magazines
Euromoney Institutional investor : Euromoney Institutional Investor PLC ("Euromoney") is a global, multi-brand information business which provides critical data, price reporting, insight, analysis and must-attend events to financial services, commodities, telecoms and legal markets. Euromoney is listed on the London Stock Exchange and is a member of the FTSE 250 share index.	**Euromoney :** Euromoney, founded in 1969 to chart the liberalisation of cross-border capital flows, is the leading publisher on the world's banking and financial markets. Our coverage provides unrivalled insight into the finance houses at the heart of global finance through our privileged access to their senior leaders. **Institutional Investor :** For 50 years, Institutional Investor has built its reputation on providing must-have information for the world's most influential decision-makers in traditional and alternative asset management.
FTAdviser : FTAdviser.com is dedicated to the financial intermediary market covering investments, mortgages, pensions, insurance, regulation and other key issues. The strength of FTAdviser.com comes from dedicated up-to-the-minute news articles and in-depth commentary written by the FTAdviser.com team, combined with the expertise of Financial Adviser and Money Management magazines, whose content feeds directly into the site.	**Financial Adviser :** The premier weekly newspaper for the UK´s financial intermediary community, Financial Adviser was launched in 1988 after the Financial Services Act 1986 defined for the first time the role of the independent financial adviser. Financial Adviser offers comprehensive and in–depth coverage of the retail finance landscape.
Global Fund Media : Founded in 2002, GFM Ltd is the most targeted digital news publisher serving institutional investors/wealth managers and their investment managers/advisers across all asset classes with seven daily global newswires and real-time news-driven web sites.	**AlphaQ :** Compendium of investment ideas, skills and talent across all asset classes **Institutional Asset Manager :** Institutional investors/pension funds and their managed funds/investment managers **Wealth adviser:** Private client/wealth managers, family offices, trustees and their investment advisers.
IPE International Publishers Ltd	**Investment & Pension Europe (IPE) :** IPE is the leading European publication for institutional investors and those running pension funds. It is published by IPE International Publishers Ltd, an independently-owned company founded in July 1996.
PLANSPONSOR/PLANADVISER, with its reputation for editorial integrity, objectivity, and leadership, is the trusted information and solutions resource for America's retirement benefits decision makers. With its powerful array of customer-driven marketing programs, PLANSPONSOR/PLANADVISER offers industry providers an unparalleled ability to reach this influential audience. With all of the changes within the retirement industry, plan sponsors and advisers rely on PLANSPONSOR/PLANADVISER magazine to help them stay informed of crucial issues and important new innovative solutions.	**Plan Sponsor :** Since 1993, PLANSPONSOR has been the nation's leading authority on retirement and benefits programs and has been dedicated to helping employers navigate the complex world of retirement plan design and strategy. **Plan Adviser :** Over the past 10 years, retirement plan advisers have reshaped the face of retirement benefits programs and PLANADVISER has been there every step of the way—providing deep insight into the most pressing retirement plan challenges and strategies facing this specialized group. Our mission, through diverse media channels, is to identify and explore the most critical selling and servicing strategies and tactics facing retirement plan advisers and their clients.

Table 2
Descriptive statistics - PF&I database
Panel A presents the distribution of news collected by year
Panel B presents the distribution of news collected by magazine

Panel A : Frequency of news collected by year

Year	*Frequency*	*Percent*
1996	60	0.06%
1997	152	0.15%
1998	158	0.16%
1999	149	0.15%
2000	165	0.16%
2001	190	0.19%
2002	1912	1.88%
2003	3545	3.49%
2004	3752	3.69%
2005	3707	3.65%
2006	4006	3.94%
2007	4651	4.58%
2008	4795	4.72%
2009	5437	5.35%
2010	7069	6.96%
2011	9321	9.17%
2012	5466	5.38%
2013	5012	4.93%
2014	7389	7.27%
2015	8370	8.24%
2016	10054	9.89%
2017	11588	11.40%
2018	4681	4.61%
Total	*101629*	*100.00%*

Panel B : Frequency of news collected by magazine

Magazine	*Frequency*	*Percent*
AlphaQ	386	0.38%
Euromoney	15129	14.89%
FT Adviser	8000	7.87%
Institutional Asset	3394	3.34%
Institutional Investor	11981	11.79%
IPE	4660	4.59%
Plan Adviser	8024	7.90%
Plan Sponsor	34417	33.87%
Wealth Adviser	15638	15.39%
Total	*101629*	*100.00%*

4 Methodology

4.1 Creating an annotated subsample

From PF&I magazines' database, we gather a subsample of 141 news. The subsample selection was made by screening news titles containing "small" and "large" words. For example, we select the news titled : "US ***small*** caps stay safe". We complete the subsample with random selection to add some news without any style content.

We read all 141 news included in the subsample : 76 news contains small capitalization information and 89 news contains large capitalization information.
We classify news under two different scheme : one related to small style content and the other related to large style content. If the news contains small (resp. large) style information, it belongs to "small" (resp. "large") class. Otherwise, the news belong to "neutral" class. This lead to four classification possibilities :

- Small/neutral : news with only small content
- Neutral/large : news with only large content
- Small/large : news with both large and small content
- neutral/neutral : news with no style content

We also proceed to an annotation at a sentence level. We annotated manually 2761 sentences : 267 "small/neutral", 106 "neutral/large", 146 "small/large" and 2242 "neutral/neutral". We illustrate our annotation process with examples for each class :

- Small/neutral sentence : "However, we believe the US small-cap landscape will be less impacted fundamentally."
- Neutral/large sentence : Over the 15-year period ended June 30, 2017, only 48% of large-cap funds survived."
- Small/large sentence : "On the other hand, however, during crises and periods of heightened risk aversion, as investors flee, small caps underperform large caps."
- Neutral/neutral sentence : "This results in a greater increase in profits and share prices."

This annotated data will be used to train different algorithms and construct small and large styles lexicons. But we have first to pre-process textual data in a convenient format.

4.2 Data pre-processing

In order to prepare news data for lexicon creation and algorithm training, we perform various pre-processing tasks using *nltk* library in Python :

1. Lowering all characters
2. Deleting all special characters and numbers
3. Applying word tokenization (i.e. converting news in a list of words)
4. Removing stop words and one character words to focus only on informational content
5. Stemming words to their root form

We apply this pre-processing to the annotated data and to the overall database.

4.3 Lexicon acquisition, algorithm training and news classification

We use three different methods to classify news : a lexicon-based approach, Naïve Bayes (NB) and support vector machine (SVM) methods. We use our subsample of annotated data in two ways : train classifiers and construct lexicons dedicated to small and large style. We subsequently compare the performance of each methods to classify news from our database.

In order to construct lexicons dedicated to style investing, we use two different sets of corpus : i) one with annotated news ii) another with annotated sentences. We explore those two solution arguing that annotation is generally available at a document level but we have here the advantage to get more granular annotation at the sentence level. We expect the set of sentences to be qualitatively more informative than the set of news since the latter contains all informational content from the news.

We automatically extract term referring to style investing using a frequency count method, i.e. sorting n-grams (group of n words) by decreasing frequency. We gather the 20 most frequent unigrams, bigrams and trigrams to discuss the construction of our style lexicons. Finally, we classify news with those lexicons : if a term from a style lexicon appears in a news, the news is then classified as a style news. Otherwise, news is classified in the neutral class.

We use three different sets of data to train different classifiers for each statistical methods (NB and SVM) : i) annotated sentences ii) annotated texts iii) annotated texts excluding common word between style and neutral classes. We argument that if we exclude common words between style and neutral classes, this could potentially lead to better performance classification. In total, we get three different classifiers for each method. Each classifier determine if a news belong to the style class or the neutral class.

Finally, we compare lexicon and statistical classifiers to select the most performant one. We manually annotate 200 news which were not included in our subsample annotated data. For each news, we compare our manual annotations to the classification results.

4.4 Investigating seasonality patterns

We count small and large style news identified by the most performant classifier. We aggregate this information by week, month and year. We didn't perform daily aggregation since we observe infrequent style information on a daily basis. We compute style coverage score which represents the proportion of style news among all news flow :

$$Style\ coverage_t = \frac{\sum style\ news_t}{\sum news_t},\ \ t = week, month, year$$

We currently get two coverage score : small and large style coverage. The difference between coverage scores represent the relative attention between small and large capitalizations in the news. We estimate the spread of coverage :

$$Coverage\ spread_t = Small\ coverage_t - \ Large\ coverage_t = \frac{\sum small\ news_t - \sum large\ news_t}{\sum news_t}$$

This coverage spread is positive (resp. negative) when attention in the news is directed to small style (resp. large style). The larger the spread, the larger the attention. We expect small cap (resp. large cap) to exhibit higher returns when coverage spread is positive (resp. negative). We therefore make the following hypothesis : *small cap returns are higher when attention is abnormally high*.

Abnormal attention is defined by high absolute coverage spread controlling for market sentiment and macroeconomic variables. We use investor sentiment index constructed by Baker and Wurgler (2006) and 5 control variables for macro-economics conditions presented in Stambaugh et al. (2012) : the default premium, the term premium, the real interest rate, the inflation rate and the consumption wealth ratio. We regress the coverage spread on those 6 variables :

$$Coverage\ spread_t = a + bS_{t-1} + \sum_{i=1}^{5} c_i M_i + \ \varepsilon_t$$

Residuals terms from this regression can be interpreted as abnormal attention. We currently investigate the right methodology to see the impact of abnormal attention on small and large stocks returns.

5 Results

5.1 Lexicon construction

Table 4 reports lexicon acquisition results by style and corpus. For sake of clarity, we report only the 20 most frequent n-grams. Unigrams are too simple to solve our feature extraction problem. We can however consider “smallcap” and “midcap” unigrams as good candidates. Trigrams are not quite different than bigrams. For example, trigrams like (‘small’, ‘cap’, ‘funds’) or (‘small’, ‘mid’, ‘cap’) respectively holds (‘small’, ‘cap’) and (‘mid’, ‘cap’) bigrams. Bigrams are therefore more suitable than unigrams or trigrams to form styles lexicons. Our subsequent discussion will only consider bigrams. Finally, we get most suitable style bigrams using the corpus of sentences rather than to the corpus of news. We can illustrate this point from Table 4. We find 8 bigrams directly referring to small style with the corpus of sentences : ('small', 'cap'), ('mid', 'cap'), ('small', 'caps'), ('smaller', 'companies'), ('international', 'small'), ('us', 'small'), ('small', 'mid'), ('small', 'companies'). In comparison, we only find 4 bigrams with a direct small style reference with the corpus of news : ('small', 'cap'), ('mid', 'cap'), ('small', 'caps'), ('smaller', 'companies').

From this qualitative appraisal, we construct style lexicons as followed :

- Small lexicon : ‘smallcap’, ‘midcap’, (‘small’, ‘cap’), (‘mid’,’cap’)
- Large lexicon : (‘larg’, ‘cap’)

We used stemmed term to correctly detect information in pre-processed news. We assumed that those restricted version of style lexicons were effective to detect style information.

Table 4

Panel A : Small lexicon acquisition by raw frequencies

	Set of news	Set of sentences
Unigrams	('funds'), ('cap'), ('fund'), ('small'), ('year'), ('markets'), ('per'), ('cent'), ('equity'), ('large'), ('market'), ('companies'), ('index'), ('years'), ('growth'), ('investors'), ('mid'), ('emerging'), ('investment'), ('us')	('small'), ('cap'), ('funds'), ('fund'), ('companies'), ('mid'), ('us'), ('equity'), ('caps'), ('index'), ('growth'), ('smaller'), ('uk'), ('stocks'), ('year'), ('international'), ('market'), ('per'), ('manager'), ('cent')
Bigrams	('small', 'cap'), ('per', 'cent'), ('cap', 'funds'), ('large', 'cap'), ('emerging', 'markets'), ('mid', 'cap'), ('actively', 'managed'), ('small', 'caps'), ('long', 'term'), ('fixed', 'income'), ('equity', 'funds'), ('cap', 'stocks'), ('five', 'years'), ('hedge', 'fund'), ('three', 'years'), ('last', 'year'), ('cap', 'growth'), ('smaller', 'companies'), ('active', 'funds'), ('standard', 'poor')	('small', 'cap'), ('mid', 'cap'), ('small', 'caps'), ('cap', 'funds'), ('smaller', 'companies'), ('international', 'small'), ('per', 'cent'), ('cap', 'stocks'), ('us', 'small'), ('small', 'mid'), ('cap', 'growth'), ('equity', 'fund'), ('cap', 'equity'), ('emerging', 'markets'), ('small', 'companies'), ('cap', 'companies'), ('cap', 'index'), ('growth', 'fund'), ('long', 'term'), ('actively', 'managed')
Trigrams	('small', 'cap', 'funds'), ('mid', 'cap', 'funds'), ('large', 'cap', 'funds'), ('cap', 'funds', 'smallcap'), ('cap', 'funds', 'midcap'), ('small', 'cap', 'stocks'), ('per', 'cent', 'per'), ('small', 'mid', 'cap'), ('small', 'cap', 'growth'), ('us', 'small', 'cap'), ('funds', 'midcap', 'outperformed'), ('midcap', 'outperformed', 'mid'), ('outperformed', 'mid', 'cap'), ('outperformed', 'large', 'cap'), ('large', 'cap', 'stocks'), ('past', 'five', 'years'), ('actively', 'managed', 'funds'), ('small', 'cap', 'equity'), ('funds', 'smallcap', 'outperformed'), ('outperformed', 'actively', 'managed')	('small', 'cap', 'funds'), ('small', 'cap', 'stocks'), ('mid', 'cap', 'funds'), ('us', 'small', 'cap'), ('small', 'cap', 'growth'), ('international', 'small', 'cap'), ('small', 'cap', 'equity'), ('small', 'mid', 'cap'), ('cap', 'equity', 'fund'), ('international', 'small', 'caps'), ('small', 'cap', 'companies'), ('small', 'cap', 'index'), ('uk', 'smaller', 'companies'), ('eafe', 'small', 'cap'), ('small', 'mid', 'caps'), ('cap', 'growth', 'fund'), ('active', 'funds', 'scorecard'), ('funds', 'scorecard', 'spiva'), ('indices', 'versus', 'active'), ('medium', 'sized', 'companies')

Panel B : Large lexicon acquisition by raw frequencies

	Set of news	Set of sentences
Unigrams	('cap'), ('funds'), ('fund'), ('small'), ('year'), ('large'), ('markets'), ('per'), ('cent'), ('equity'), ('companies'), ('index'), ('market'), ('growth'), ('years'), ('investors'), ('investment'), ('says',), ('mid'), ('us')	('large'), ('cap'), ('fund'), ('funds'), ('index'), ('growth'), ('stocks'), ('value'), ('year'), ('russell'), ('said'), ('high'), ('companies'), ('managed'), ('market'), ('europe'), ('stock'), ('developed'), ('per'), ('cent')
Bigrams	('small', 'cap'), ('per', 'cent'), ('large', 'cap'), ('cap', 'funds'), ('emerging', 'markets'), ('mid', 'cap'), ('actively', 'managed'), ('small', 'caps'), ('long', 'term'), ('cap', 'stocks'), ('equity', 'funds'), ('fixed', 'income'), ('five', 'years'), ('cap', 'growth'), ('three', 'years'), ('cohen', 'steers'), ('usd', 'billion'), ('hedge', 'fund'), ('real', 'estate'), ('last', 'year')	('large', 'cap'), ('cap', 'growth'), ('developed', 'europe'), ('per', 'cent'), ('russell', 'developed'), ('cap', 'fund'), ('europe', 'large'), ('high', 'efficiency'), ('actively', 'managed'), ('cap', 'stocks'), ('growth', 'fund'), ('cap', 'funds'), ('cap', 'value'), ('cap', 'core'), ('cap', 'high'), ('large', 'caps'), ('european', 'large'), ('gcc', 'large'), ('managed', 'large'), ('median', 'large')
Trigrams	('small', 'cap', 'funds'), ('mid', 'cap', 'funds'), ('large', 'cap', 'funds'), ('cap', 'funds', 'smallcap'), ('cap', 'funds', 'midcap'), ('small', 'cap', 'stocks'), ('per', 'cent', 'per'), ('large', 'cap', 'stocks'), ('us', 'small', 'cap'), ('funds', 'midcap', 'outperformed'), ('midcap', 'outperformed', 'mid'), ('outperformed', 'mid', 'cap'), ('small', 'cap', 'growth'), ('large', 'cap', 'growth'), ('outperformed', 'large', 'cap'), ('small', 'mid', 'cap'), ('past', 'five', 'years'), ('actively', 'managed', 'funds'), ('cap', 'growth', 'fund'), ('funds', 'smallcap', 'outperformed')	('large', 'cap', 'growth'), ('russell', 'developed', 'europe'), ('developed', 'europe', 'large'), ('europe', 'large', 'cap'), ('large', 'cap', 'fund'), ('cap', 'growth', 'fund'), ('large', 'cap', 'stocks'), ('large', 'cap', 'funds'), ('large', 'cap', 'value'), ('cap', 'high', 'efficiency'), ('large', 'cap', 'core'), ('large', 'cap', 'high'), ('gcc', 'large', 'cap'), ('median', 'large', 'cap'), ('us', 'large', 'cap'), ('cap', 'core', 'fund'), ('cap', 'value', 'fund'), ('index', 'russell', 'developed'), ('managed', 'large', 'cap'), ('actively', 'managed', 'large')

5.2 Classifiers performance

We report performance scores of classifiers in Table 5. The following measures are estimated : accuracy, precision, recall and F1 score. In this research, we want to detect all news related to style. We therefore try to minimize Type II error, i.e. minimize the number of false negative classification. In other words, we want to avoid as much as possible to wrongly classify style news in the neutral class. This means maximizing recall score which represents the fraction of style news successfully retrieved by the classifier.

Lexicon classifier exhibits best performance scores relatively to other statistical classifiers. Recall scores for lexicon classifier are 0.95 for small class and 0.93 for large class. These results are yet achieved with very few lexicons terms. Nevertheless, small cap, mid cap and large cap terms are so specific to small and large capitalization stocks that they appear in all related news.

Naïve Bayes classifiers using annotated texts detect 79% to 84% of small style news and 68% to 88% of large style news. Naïve Bayes classifier using sentences get poor recall scores. Support vector machine classifiers detect 64% to 86% of small style news and 73% to 95% of large style news. We observe best performance with subsample of annotated sentences. We observe that recall scores are decreasing with the amount of words contained in the annotated subsample for NB classifiers. This is the opposite for SVM classifiers. We still need to investigate the potential explanation for those observations.

We also want to remind the reader that we use less than 150 news and 3000 sentences to train NB and SVM algorithms. Performance scores could potentially be improved with an extended annotated subsample. Moreover, we may not find strong performance for lexicon classifier in the case of other anomalies. From our current lecture, we advocate that there exists no specific terms such as small cap and large cap in the case of size anomaly.

We conclude this section by selecting the lexicon classifier to investigate seasonality in size anomaly.

Table 5 - Classifiers performance

Panel A : Small style classification

Performance scores for small class				
	Accuracy	**Precision**	**Recall**	**F1_score**
Lexicon	0.97	0.98	0.95	0.97
NB_text	0.69	0.65	0.84	0.74
NB_text (no common words)	0.71	0.68	0.79	0.74
NB_sentences	0.61	1.00	0.23	0.61
SVM_text	0.71	0.76	0.64	0.70
SVM_text (no common words)	0.67	0.67	0.71	0.69
SVM_sent	0.79	0.75	0.86	0.81

Panel B : Large style classification

Performance scores for large class				
	Accuracy	**Precision**	**Recall**	**F1_score**
Lexicon	0.96	0.99	0.93	0.96
NB_text	0.68	0.64	0.88	0.76
NB_text (no common words)	0.58	0.57	0.68	0.63
NB_sentences	0.51	1.00	0.03	0.51
SVM_text	0.64	0.61	0.79	0.70
SVM_text (no common words)	0.58	0.56	0.73	0.65
SVM_sent	0.72	0.66	0.95	0.80

5.3 Seasonality in anomalies

Results discussed in this section use news classification from lexicon classifier. We present yearly news frequency and style coverage in Table 6. We limit the period from January 2003 to December 2017 because it is the most representative part of our database in term of news collected and magazine covered. News frequency increases over time from 3500 news in 2003 to more than 11000 news in 2017. Small and large coverage are below 3% on yearly basis. This apparent low coverage isn't surprising : PF&I cover broad range of topics and asset classes.

We are more interested in the variation of coverage through the year and specially by the coverage spread. We found that the mean of weekly coverage spread is equal to +0.5% and is significantly different from zero. A positive spread means that news are more talking about small capitalizations as a strategical sub-asset class. The opposite pattern is found in conventional media : large companies have more coverage.

Table 6 - Yearly News frequency and style coverage

Year	News frequency	Small coverage	Large coverage
2003	3535	1.87%	1.47%
2004	3752	2.08%	1.63%
2005	3698	2.03%	1.60%
2006	4021	1.67%	1.02%
2007	4706	1.66%	1.40%
2008	4890	1.68%	1.41%
2009	5576	1.69%	1.61%
2010	7080	2.40%	1.57%
2011	9323	2.04%	1.31%
2012	5493	2.18%	1.57%
2013	5032	2.42%	2.01%
2014	7406	2.86%	1.92%
2015	8382	2.43%	1.84%
2016	10058	2.21%	1.48%
2017	11582	2.14%	1.50%

We currently work on abnormal coverage and it relation with SMB factor. Our results will be updated.

6 Conclusion

The paper is currently at a very early stage as we focused on the construction of the algorithm and the text manual collection. Results are promising and will be further developed. It currently focus on the size anomaly but will be extended to other anomalies.

References

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance*, *59*(3), 1259–1294.

Asness, C. S., Frazzini, A., Israel, R., Moskowitz, T. J., Pedersen, L. H., 2015. Size matters, if you control your junk. Fama-Miller Working Paper.

Barber, B. and T. Odean (2009), All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors, The Review of Financial Studies, 21(2), 785–818.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. Journal of Finance 61, 1645–1680.

Das, S.R. , Chen, M.Y. , 2007. Yahoo! for amazon: sentiment extraction from small talk on the *web. Manag. Sci. 53* (9), 1375–1388.

Engelberg, J., R. D. Mclean, and J. Pontiff. 2018. Anomalies and news. Journal of Finance 73(5), 1971-2001.

Fedyk, A. (2018), News-Driven Trading: Who Reads the News and When?, Working Paper - UC Berkeley, Haas School of Business.

Fisher, K. L., and M. Statman. "Investor Sentiment and Stock Returns." Financial Analysts Journal, 56 (2000), 16-2

Garcia, D. (2013). Sentiment during Recessions. *Journal of Finance*, *68*(3), 1267–1300.

Goetzmann, W. N., and M. Massa. "Index Funds and Stock Market Growth." Journal of Business, 76 (2003), 1-2

Green, J., J. R. M. Hand, and X. F. Zhang. 2017. The characteristics that provide independent information about average U. S. Monthly stock returns. Review of Financial Studies 30:4389-436.

Haugen, R. A. and Jorion, P. (1996). The January effect: still there after all these years, *Financial Analysts Journal, 52*, 27–31.

Haug, Mark, and Mark Hirschey (2006). "The January Effect". *Financial Analysts Journal, 62* (5), 78-88.

Henry, E., Leone, A.J. (2016). Measuring qualitative information in capital markets research: comparison of alternative methodologies to measure disclosure tone. *Account. Rev. 91* (1), 153–178.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. … and the cross-section of expected returns. *Review of Financial Studies* 29: 5–68.

Hou, K., C. Xue, and L. Zhang. 2017. Replicating anomalies. Review of Financial Studies (forthcoming)

Keim, D. B., 1983. Size-related anomalies and stock return seasonality: further empirical evidence. Journal of Financial Economics 12, 13–32.

Kumar, A. (2009), Dynamic style preferences of individual investors and stock returns, Journal of Financial and Quantitative Analysis 44(3), 607-640.

Lakonishok, J., Shleifer, A., Thaler, R., Vishny, R. (1991). Window dressing by pension fund managers. *American Economic Review 81* (2),227–231.

Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research, 48* (5): 1049–1102.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics 47*, 13-37.

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65.

Merton, R.C., 1987. A simple model of capital market equilibrium with incomplete information. Journal of Finance 42, 483–511

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, *85*, 62–73.

Reinganum, M. R., 1981. Misspecification of asset pricing: empirical anomalies based on earnings' yields and market values. Journal of Financial Economics 9, 19–46.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking and Finance*, *84*, 25–40.

Ritter, J. (1988). The buying and selling behavior of individual investors at the turn of the year. *Journal of Finance 43*,701–717.

Roll, R., 1981. A possible explanation of the small firm effect, Journal of Finance 36, 879–888.

Rozeff, M., Kinney, W. (1976). Capital market seasonality: the case of stock returns. *Journal of Financial Economics 3*, 379–402.

Sharpe, W. F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance 19*, 425–442.

Sikes, S. A. (2014). The turn-of-the-year effect and tax-loss-selling by institutional investors. *Journal of Accounting and Economics*, *57*(1), 22–42.

Stambaugh, Robert F., Jianfeng Yu, and Yu Yuan, 2012, The short of it: Investor sentiment andanomalies,Journal of Financial Economics104, 288–302.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment : The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance 63*, 1437–1467.

Appendix 1

Example of news converted in plain text :

magazine : FTAdviser

section : Investments

date : 20160720

title : US small caps stay safe

author :Simoney Kyriakou

url : https://www.ftadviser.com/2016/07/27/investments/north-america/us-small-caps-stay-safe-uUTxM2pMpjjOSSiBLwIXnO/article.html

TEXT

Brexit woes may hit global large-caps, but for investors who want a long-term holding, US small caps will be relatively sheltered from the effect of the vote to leave, a fund manager has claimed. According to Chris Berrier, co-portfolio manager with David Schuster of the $127m (£97m) Brown Advisory US Small-Cap Blend Fund, the Brexit vote has created a "great deal of uncertainty in markets". Mr Berrier said: "We are expecting a heightened period of volatility to last for some time. However, we believe the US small-cap landscape will be less impacted fundamentally. "Today, US small-cap valuations – when compared to their large-cap peers - look relatively attractive, and we would expect a premium for secular growth when it is scarce." The comments came as the fund marked its three-year anniversary on 8 July delivering strong outperformance since its inception of 8.5 per cent (net of fees on an annualised basis) against its benchmark, the Russell 2000 Index, at 6.7 per cent. The Dublin-based Undertakings for Collective Investments in Transferable Securities (Ucits) fund aims to provide investors with access to the best US small-cap companies across the growth and value spectrum. In Financial Adviser's sister website FTAdviser's latest CPD-qualifying guide to US small and mid-caps, Jenny Jones, manager of the £1.37bn Schroder US Mid-Cap fund, said it was important to remember long-term investing means more than just three to five years, and for patient investors, putting money away now into small caps can reap rewards over the real long-term.