BRAIN

Brain Language Metrics on Company Filings

Product Summary

The Brain Language Metrics on Company Filings (BLMCF) dataset has the objective of monitoring several language metrics on 10-Ks and 10-Qs company reports for approximately 6000+ US stocks.

Recent literature works claim inefficiencies in the market response to company filings information due to the increased complexity and length of such reports (see for example *"Lazy Prices" Cohen et al. 2018 or " The Positive Similarity of Company Filings and the Cross-Section of Stock Returns", M. Padysak 2020).*

Our dataset is made of two parts; the first one includes the language metrics of the most recent 10-K or 10-Q report for each firm, namely:

1. Financial sentiment
2. Percentage of words belonging to financial domain classified by language types:
    - *"Constraining"* language
    - *"Interesting"* language
    - *"Litigious"* language
    - *"Uncertainty"* language
3. Readability score
4. Lexical metrics such as lexical density and richness
5. Text statistics such as the report length and the average sentence length

The second part includes the differences between the two most recent 10-Ks or 10-Qs reports of the same period for each company, namely:

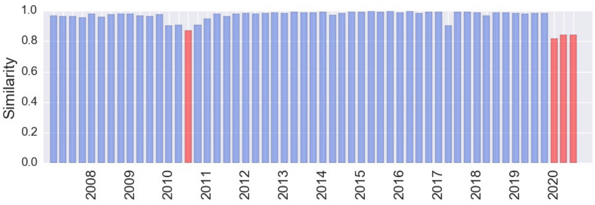
1. Difference of the various language metrics (e.g. delta sentiment, delta readability score delta, delta percentage of a specific language type etc.)
2. Similarity metrics between documents, also with respect to a specific language type (for example similarity with respect to *"litigious"* language or *"uncertainty"* language)

Our dataset includes the metrics and related differences both for the whole report and for *specific sections* (Risk Factors and Management Discussion and Analysis)

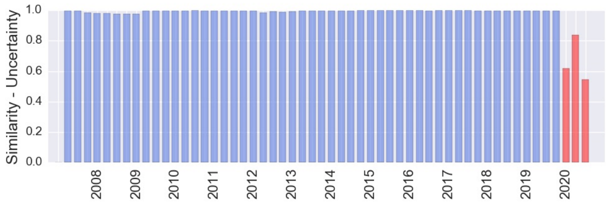
Examples of Metrics

One of the language metrics included in the dataset is the similarity between 10-K and 10-Q reports as shown in the following plots for AAPL stock.

Similarity with focus on generic financial domain language(whole report) - AAPL

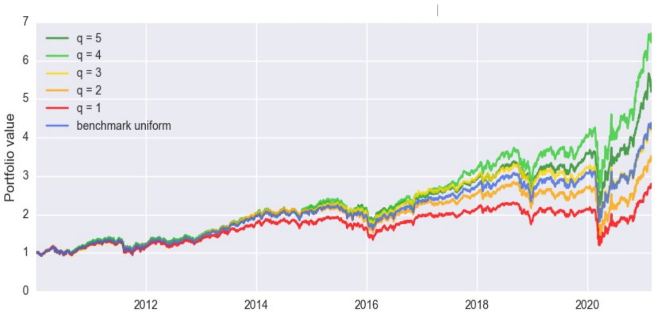


Similarity with focus on "uncertainty" financial domain language (Risk Factors section) - AAPL



Use Case

The backtesting of Russell 3000 universe ranked by the similarity metrics of the Risk Factor (RF) section shows a promising separation in quintiles in the interval 2010-2021 (quarterly rebalancing and uniform weights); Stocks that show more changes in the RF section seem on average to underperform the universe (top quintile, dark green) and, vice versa, stocks that show less changes in the RF section seem on average to outperform the universe (bottom quintile, red color).



Dataset Frequency

The dataset is updated with a daily frequency since new 10-Ks and 10-Qs reports are released every day for some of the universe companies. Clearly the largest update will be around February, April, August and November when the largest number of reports is released. The historical dataset is available from year 2010.

Contacts

BRAIN is a Research Company that develops proprietary signals based on alternative data and algorithms for investment strategies on financial markets.

- EMAIL: contact@braincompany.co
- WEB: http://www.braincompany.co

Last update: 03/05/2021

DISCLAIMER: the content of this presentation is not to be intended as investment advice. The material is provided for informational purposes only and does not constitute an offer to sell, a solicitation to buy, or a recommendation or endorsement for any security or strategy, nor does it constitute an offer to provide investment advisory or other services by Brain. Brain makes no guarantees regarding the accuracy and completeness of the information expressed in this document.