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 interesting 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).*

The 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:

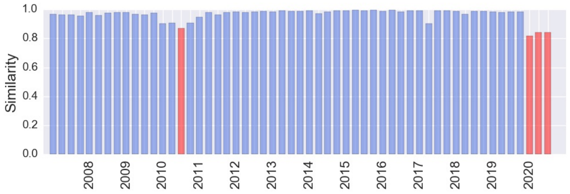
1. Differences of the various language metrics (e.g. delta sentiment, delta readability score delta, delta percentage of a specific language type etc.)
2. Similarity metrics between documents, also with respect to a specific language type (for example similarity with respect to *"litigious"* language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

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

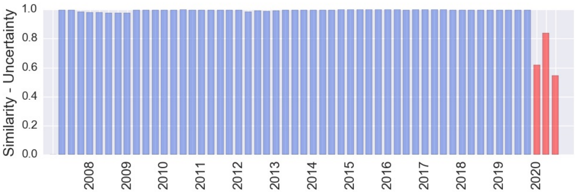
Some Examples

One metric included in the dataset is the similarity between 10-K and 10-Q reports of the same period. In the following plots we are showing some examples; the similarity of the whole AAPL report with respect to generic financial domain language (first plot) and then with focus on "uncertainty" language in financial domain in the Risk Factors section (second plot).

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

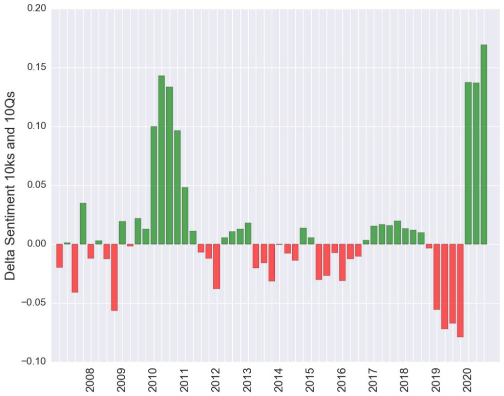


Similarity with focus on "uncertainty" financial domain language (Risk Factors section)



Another metric included in the dataset is the difference of the sentiment scores of 10-K and 10-Q reports with scores of the same period of the previous year, see below

Difference of Sentiment Score in AAPL filings



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 most relevant updates 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: https://www.braincompany.co

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.