

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.

In recent papers there has been a growing attention towards the language analysis of company reports and the study of possible relations with firms' future performance. See for example *"Lazy Prices" Cohen et al. 2018* or *"The Use of EDGAR Filings by Investors", Loughran and McDonald, 2017*.

Some literature works claim inefficiencies in the market response to company filings information due to the increased complexity and length of such reports; over the last 20 years, the length of the average 10-K has in fact increased dramatically.

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

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:

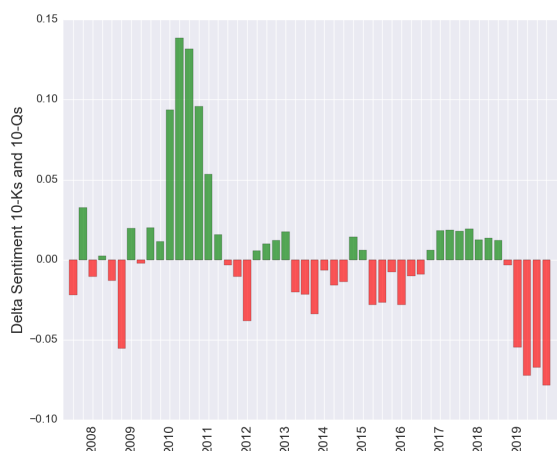
1. Sentiment difference
2. Difference of percentage of words for specific types of language (e.g. "litigious" language)
3. Similarity metrics between documents, also with respect to a specific language type (for example similarity with respect to "litigious" language or "uncertainty" language)

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 2007.

Some Examples

The following graph shows the sentiment differences of 10-K and 10-Q reports of the same period for AAPL stock since 2006. We note that the sentiment differences in 2019 show the lowest values in the last 12 years.



Another language metrics 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 the similarity of AAPL reports with respect to generic financial domain language (first plot) and then with focus on "litigious" language in financial domain (second plot). The interval 2010 – 2011 shows the lowest similarity, in particular in terms of "litigious" language.



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>