Sentiment classification for specialized applications: Finance

Introduction: While general sentiment analysis has made substantial progress in both its accuracy and relevance, progress in specialized data sets has been less forthcoming. In particular, classification of news articles is an important component of financial markets. And yet, technologies to classify financial news articles have been rather unsuccessful, despite significant efforts to do the same [1,2,3,4]. The reasons for this are many. In this article, we will discuss some of the state of the art technologies used for news sentiment parsing, and discuss some of the shortcomings of the current methods.

Data sources: There are a few prominent sources of data for the financial markets. First, we have the generic news feeds such as cnn, bbc, etc. While there is a great deal of readership associated with these news articles, they are often not finance focused and are not incentivized to provide price-moving information. For such news, most market participants turn to specialized news providers such as Bloomberg, Wall Street Journal (WSJ), Financial Times (FT), and other outlets. Thirdly, there are the news wire services like Reuters that provide a substantial volume of financial news. As expected, most of these specialized news source are expensive to obtain. Fourth, there are the social media platforms like twitter and stocktwits and public blogs that are used to disseminate information both by companies, expert, and novice commentators. Finally, there is a wealth of regulatory and publicly available data sources that are filed as a part of government requirements. One such example is the Security and Exchange Commission's (SEC) EDGAR data base. The EDGAR DB is maintained by the SEC, a US government agency and contains a wealth of data from various companies that file into the database on a regular basis.

Challenges:

1. Heterogeneity: The primary difficulty with parsing financial sentiment data is the wide variety of news sources that are available. As one can surmise easily, the techniques used for reading a long-form Bloomberg article are very different from the techniques used for reading twitter feeds. This results in the need to develop specific ML models (and even language models) that are tuned to the source of the data. Differences also exist in terms of linguistic contents between various types of news sources – E.g. Bloomberg vs Reuters news feeds contain very different language in their news feeds.
2. Market moving news vs commentaries: The second difficulty is the problem of distinguishing between news that is market moving vs commentaries that are done after the market moving news has already moved the markets. For example, it is important to find a breaking news that is associated with a mergers and acquisitions transaction. On the other hand, a long-form commentary on this event that discussed the various participants, and their performances are likely to be less useful.
3. Quantitative vs soft news: The next difficulty is in parsing quantitative news well for financial news. Financial news can be both quantitative (e.g. earnings estimates, revenue estimates, from earnings releases). And yet, there are also numerous articles that discuss soft news. For example, Benzinga could be discussing a company and why it should be purchased. Segmenting these into different categories and making sure that we read the news in context is a very challenging problem to solve.
4. Peer comparisons, and numerical assignments: This is a crucial stumbling block for finance. When a lot of earnings news are released, it is relatively easier for a human to parse all the numbers, and compare it with their peer companies to understand the performances better. And yet, it is difficult, if not impossible for machines to do the same. This is because, these tables with numeric information

about a companies performance are often non-standardized. As a result, one if forced to manually read, understand the context, and then compare with peers. Although several datasets to compute peer stocks exist, yet, understanding the revenue drives, and revenue change drivers from particular segments are important.

For example, recently, Intel delivered their earnings, and it turned out that one of their divisions lagged their revenue growth (not even their revenues!) and it was picked up by humans, and the intel stock suffered a 10% reduction in price overnight. This was not picked up by the usual data feeds, as there was no prior about this particular segment in most news articles.

State of the art technology:

While traditionally, feature engineering news articles has been the best way of measuring news sentiment [2], recent advances in Keras and transformers promise to provide massive improvements in financial text data.

At this point, it should also be mentioned that the Named Entity Recognition (NER) problem in finance has been solved reasonably well, at least in so far as identifying the primary company's name. And yet, finding relationships between the various companies is an unsolved problem.

Keras has been used successfully to identify guidance and quantitative information from various news feeds including Reuters feeds. The idea here is to create a hand-made training data set of information that has been deemed useful by the markets and that have resulted in market moves in the past. This data is them lemmatized, stemmed, and used to train Keras models. These models are then used in live news feeds, and updated over time.

Finally, transformer technologies have also been ported for finance for sentiment classification. One example of this are the FinBERT series of models based on Google's BERT (BiDirectional Encoder Representations from Transformers). However, despite the hype, the outperformance of such models have been mediocre at best. In fact, in a peer review recently, FinBERT was not deemed to have substantial improvements over the BERT models to warrant inclusion in a top conference.

In summary, we have discussed the numerous challenges associated with sentiment scoring in financial data. Admittedly, this is a broad 30k feet overview rather than a detail. There is still a lot of work to be done in the area to beat human parsing.

References

[1] Good news, Bad news, No news : The media and the cross-section of stock returns [2020]

[2] Measuring news sentiment[2020]

[3] Social media sentiment, limited attention and stock returns [2019]

[4] News media analytics in finance [2019]