Incorporating Data Mining and Advanced Big Data Analytics into Prediction Models

What is Big Data and what is our initial approach?

A wealth of unstructured data lies at our fingertips, with only limited developed methodology at hand to investigate and exploit this data. Our mission is to draw insightful and well informed conclusions about a given area of interest - all the while efficiently utilising the abundance of unstructured data in conjunction with typically used market data - which are superior to those drawn without the inclusion of unstructured information. In brief, we aim to improve well established methods of modelling and prediction through the implementation cutting edge data analytics.
The following steps summarise the workflow:

1) We use established ***data mining*** techniques to acquire a large amount of unstructured data from sources such as Twitter, Google and business new RSS feeds. This data collection process shall focus on one specific financial subject at a time, for example S&P500 Index, BMW shares, interest rates, consumer debt etc.
2) We clean, standardise and organise the data. Examples are: removing useless words (and, it, of, etc.), selecting only one language, storing data in an easily manipulated structure e.g. a Corpus within R.
3) We develop a system of intelligent algorithms that evaluate the data for its usefulness and the writer's sentiment, defining features demonstrated by the data and subsequently devising a system to evaluate and rate the significance of individual data. These measurements must then be amalgamated to produce a single/set of factor(s) describing the magnitude of influence to further use in a model.
4) We create a larger model incorporating our quantified output from the originally unstructured data with relevant notable (observable and measurable) market data e.g. index/stock movements, interest rates and implied volatility. This will resemble some form of regression model, likely being autoregressive. The following illustrates the basic form:

$$f(Z) = \beta_0 + \beta_t \cdot X_t + \beta_{t-1} \cdot X_{t-1} + \gamma_t \cdot Y_t + \varepsilon$$

In this generalised regression expression we are predicting Z. The terms X_t correspond to measurable market indicators, with their β-values their equivalent estimators. We also see there is an autoregressive factor, X_{t-1} and its estimator. The Y_t terms are those derived from our analysis of unstructured data with analogous estimator γ_t.

Possible Shortfalls?

1) It has proven to be a very difficult task to subjectively classify statements without human assistance. The validity of results will therefore require a high level of scrutiny and sanity checking before they can be implemented with confidence.
2) Determining the direction of causality may prove difficult. Is sentiment bad because of what has happened or did something bad happen as a result of the public sentiment? Can this be clarified with Granger causality tests alone?
3) Use of sentiment analysis in conjunction with auto-regressive factors may complicate the interpretation significance of many model factors.

Delving Deeper – Gradient Boosting

Applying this methodology allows a sequential classification algorithm, whereby each step improves on all preceding steps, meaning irrelevant predictors are methodically phased out.

Advantages:

1) A broad selection of classification models can be incorporated due to a very general loss function.
2) Boosting can be applied to wide data, where p>>n.
3) Reduced possibility of multicollinearity effects

Disadvantages:

1) The 'over fitting' of noisy data is a danger, something unstructured data may be vulnerable to. We may end up too focussed on training data.

Literature:

- A good introduction to text mining using R
  → *http://www.jstatsoft.org/v25/i05*
- How labels and values can be assigned
  →*Sentiment Analysis and Subjectivity, Bing Liu*
- All-round coverage of statistical methods in R:
  An Introduction to Statistical Learning
  → http://www-bcf.usc.edu/~gareth/ISL/getbook.html
- Boosting →*Elements of Statistical Learning* (§10)
  http://tinyurl.com/mdch4am