DOI: 10.17626/dBEM.ICoM.P00.2015.p033

APPLYING ALTERNATIVE METRICS IN THE QUANTIFICATION OF NEWS

Márk MOLNÁR, Zsuzsanna NAÁR-TÓTH
Szent István University, Gödöllő, Hungary
E-mail: mark.molnar@gtk.szie.hu

Summary: Using a common definition we can define news analysis as the measurement of the various qualitative and quantitative elements of textual news stories. These elements include sentiment, relevance and novelty. By quantifying news stories we can gain a useful way to manipulate and use everyday information in a mathematically concise manner. In this article a framework for news analytics techniques used in finance is provided. Various news analytic methods and software are discussed, and a set of metrics is given that may be applied to assess the performance of analytics. Various directions for this field are discussed. The proposed methods can help the valuation and trading of securities, facilitate investment decision making, meet regulatory requirements, or manage risk.

Keywords: textual news stories, finance, alternative metrics, software, investment decision making

1. Introduction

Quantitative analysis of text (news, tweets, articles, etc) can provide additional information for financial analysis. First of all text contains an additional emotional content (called sentiment) which provide valuable input for further conjectures on a given topic. Another important factor is the opinions and links found in the text to other sources. Third the quantification of some intrinsically qualitative information can be difficult and results in „signal loss". Fourth, textual information contains some additional value over aggregated and composite quantitative information (Loughran, McDonald, 2014).

Evolution of computer hardware, computing power and storage capacity allowed for the birth and fast evolution of data mining. In addition to the vast amount of data generated every day and every hour it is possible to rely on large databases for analysing information. Dictionaries are used for providing a quick assessment of sentiments found in an article by quickly comparing the contents of the respective text with the built-in words and developing a score. One example is the Harvard Inquireer (http://www.wjh.harvard.edu/ inquirer/) which allows for deciding on the optimistic or pessimistic nature of an article. Associative dictionaries are also a novelty, they basically function as a thesaurus and allow for establishing the proper context of a text. To visualise context some webpages offer online graphics to provide an eycatching clue, see e.g. Visuwords or other pages.

Advantage of using such dictionaries is that they provide an unbiased fundament for evaluation, an objective basis which can be referenced and referred in research.

Many techniques exist to reduce the enormous amount of textual input to process thus simplifying analytical work. One interesting and important element is text summarisation. A simple form of summarisation is when we select the sentence(s) with the largest commonality index; that is, a number which represents the similarity beween the text and other elements. One of the basic measures used is the Jaccard formula (Jaccard, 1901), which allow for the composition of the Jaccardi matrix. The (i,j) element of the Jaccardi matrix is given as follows $J_{ij} = \left|\frac{s_i \cap s_j}{s_i \cup s_j}\right| = J_{ji}$. Similarity is calculated by calculating row sums, $S_i = \sum_j J_{ij}$, and a natural ranking according to the significance of a sentence can be starting from the lowest values (e.g. highest information content or novelty).

After having done some analysis and editing tasks and having trimmed the text to our needs the next task is to analyse the text. One important step is to extract sentiments and decide about the message of the text (e.g. optimistic, pessimistic, neutral).

One method for this process is the Bayesian classification, where we use a training set to "teach" the computer to classify documents based on the occurrence of typical terms (so called prior probabilities) and use the definition of the conditional probability to calculate posterior probabilities to classify new documents in the given classes of sentiments.

Another method frequently applied is the support vector model, which spearates the datasets using a distance maximisation method (e.g. distance between data groups is maximised by fitting (a) separating hyperplane(s)) between).

A simple way can be the word count method where we simple count the number of words with positive and negative sentiment and get a net balance of the text.

2. Applying metrics text analysis assessment

When trying to establish the quality of an algorhitm in text minig, it is important to apply certain metrics. Originally metrics mean a measurement of distance in mathematics, in the current context they provide a means to test for the goodness of the text mining algorithm. Here we present only a few examples from the literature (see e.g. Das, 2014 and Das and Chen, 2007).

One important element is the confusion matrix, which describes the goodness of classification using a matrix form. Simply put, assuming a *k* categories, we have a quadratic K x K matrix, where the rows represent actual categories, columns represent assigned categories, and any cell (*i,j*) represents a text which is category *i* and was assigned to category *j*. Obviously only elements in the diagonal of the matrix represent well classified elements, all other elements which are non-zero represent classification error (thus the notation confusion matrix).

The test is based on a χ^2 critical value, the null hypothesis that in the case of random guessing (a completely useless algorithm) the rows and columns would be independent. Denote with O(*i,j*) the actual elements of the confusion matrix and E(*i,j*) the expected element under the assumption of no classification (uniformly distributed random values, e.g. the number of observations in the *i*th row and *j*th column divided by the total number of observations).

$$\chi^2_{(K-1)^2} = \sum_i \sum_j \frac{\left(O(i,j) - E(i,j)\right)^2}{E(i,j)}$$

Depending on this test statistics we can decide about accepting the algorithm.

Based on the elements of the confusion matrix accuracy can also be tested with the following metrics using the previous notations

$$A = \frac{\sum_i O(i,i)}{\sum_j M(j)} = \frac{\sum_i O(i,i)}{\sum_i M(i)}:$$

This is simply the sum of the diagonal elements divided by the sum of all elements of the matrix.

Incorrect classification can be sometimes more harmful than no classification at all. Incorrect classification can be simply counted as the percentage of elements which are not correctly assigned (this can be weighted). A logical assumption is that the categories are arranged in a manner where neighbouring categories have proxmitiy in their sentimental content, too. Under such arrangement it is expected that a classification which puts a given category to a category with distinctively different meaning causes much more harm than a misclassification to a category in the "vicinity". In our proposition below we try to give a way to resolve this issue by introducing a vicinity factor in misclassification.

3. Proposed new metrics in text analysis

One important element in text analysis is classification of text. Besides that, in our proposed method it is possible to identify the main market tendencies according to the followings.

Assume that the information from the market is organised into *n* documents (sources) and that at most *k* distinctive terms are

Suggestion for systemic error testing

If a classification algorithm is completely precise, we would only receive elements in the main diagonal, that is, the rank of the matrix would be full (equaling the number of rows). If on the other hand we have a systemic error in the algorithm, this would mean a tendency of false classification. In this case a category could be replaced by one or more other categories and the classification would not suffer any loss. For this we suggest a rank probe, that is to calculate the rank of the confusion matrix. If the rank is lower than the order of the matrix (*k*) that means that one category can be reproduced as a linear combination of other (one or more) categories. In that case the algorithm is generating systemic, inherent errors. If the rank of the confusion matrix is full, then the algorithm contains only random errors.

Communality matrix and determination of principal vectors in news

Concerning miscategorisation as a grave error it is logical to identify a measure to deal with this problem. Assume that categories are assigned in a logical order (e.g. decreasing sentiment, etc.) and that the the algorithm is not degenerative, that is the *K* confusion matrix is full rank. In that case it is possible to apply linear transformation and gain the Jordan canonical form (Molnár - Szidarovszky, 2002). In that case there exist at least one real eigenvalue of the matrix, but more importantly the basis of the Jordan-form matrix is composed of the eigenvectors of the original matrix (transformation, or in our case classifying algorithm).

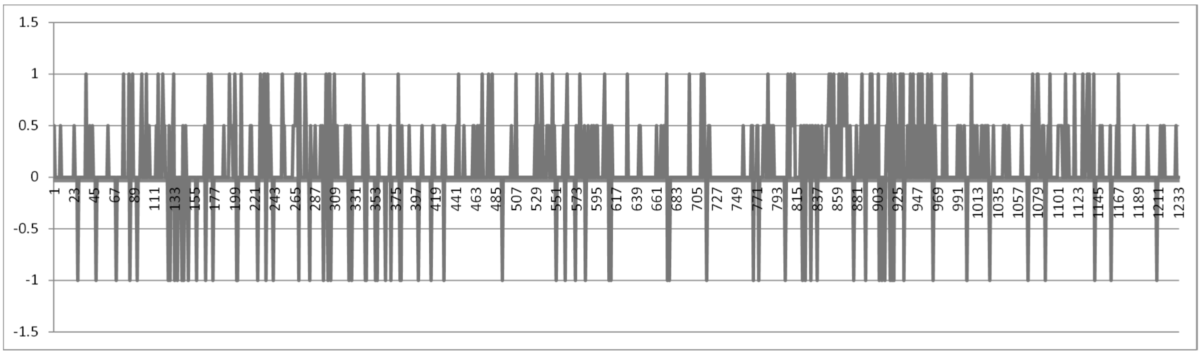
A measure for the degree of miscategorisation can be a simple euclidean distance of the standardised eigenvectors. If the distance is less than a given threshold, then the categorisation can be accepted. If the distance is very large than the algorithm can be considered risky from the aspect of miscategorisation.

4. Empirical results of news analytics

Some elements of the theoretical results were applied to a specific case along the following lines. The Hungarian Oil Companies (MOL) and the Croatian Industrianafte (INA) formed a strategical alliance in 2003 and MOL became the owner of almost 50% of the INA shares. In our short analysis we analysed approximately 850 articles from Hungarian websites (primarily, portfolio.hu). These articles were grouped into three categories based on keyword

assessment: bearish (pessimistic), bullish (optimistic) and neutral, and were scored accordingly. In many cases the articles were of political nature and thus had additional layers of information. In cases where multiple messages (perhaps of mixed positive and negative nature) were found the overall aggregate value was considered for that day.
This was matched with the daily movement of MOL share prices on the Budapest Stock Exchange (BUX).
These results were combined in simple difference values as follows. If the information gained from news analytics (three discrete values were possible) were matching the daily movement of prices then we assigned a +0.5 value to the forecast.
If the information derived from the analysed news were different from the share price movement we generated a +1 or -1 value depending on the direction of share price change compared to the forecast.
The information is summarised in the following chart, Figure 1.

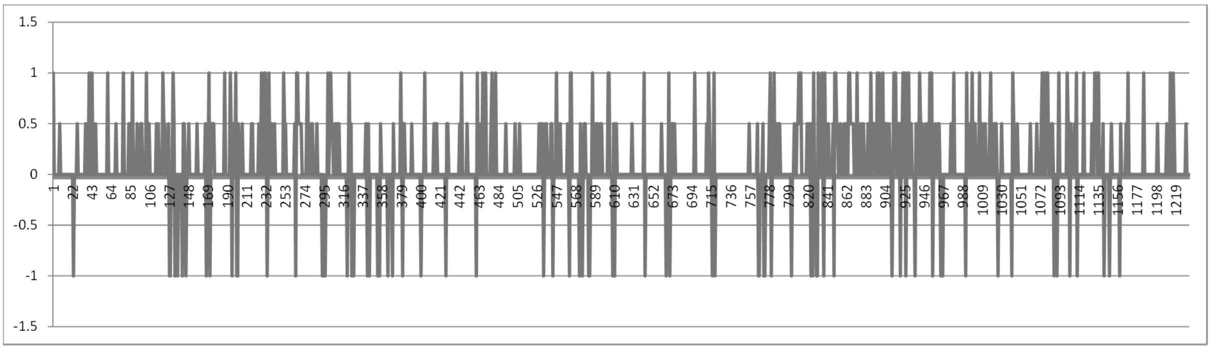
Figure 1. Differences between information of analysed news and share price movements, no lag



(1=share price increase w. negative forecast, 0.5 identical movement, correct forecast, -1=share price decrease with positive forecast)
Source: own calculations

It is well visible that the news analytics performed only partially well in forecasting price movements. As this was the contemporaneous (daily change) it is worth to check for the lag phase behaviour of the forecast. This is shown in Figure 2.

Figure 2. Differences between information of analysed news and share price movements, 1 day lag



(1=share price increase w. negative forecast, 0.5 identical movement, correct forecast, -1=share price decrease with positive forecast)
Source: own calculations

From this chart it can also be observed that a given day prediction from news analysis typically resulted in the next day share price movement following the sentiment of the news text.

5. Remarks

Although the above results are of limited scope they show that news analytics require increased attention both from the theoretical view and from the view of technical analysis. There is evidence that the market is not fully informed, at least that that full information principle only holds in a weaker form as news analysis is able to provide additional predictive abilities.

References

1. S. R. Das. Text and Context (2014): Language Analytics in Finance. Foundations and Trends R in Finance, vol. 8, no. 3, pp. 145–260.
2. S. Das and M. Chen (2007): Yahoo for amazon! sentiment extraction from small talk on the web. Management Science, 53:1375–1388.
3. Jaccard P (1901). Étude comparative de la distribuition florale dans une portion des Alpes et des Jura. Bull. Soc. Vaudoise Sci. Nat. 37: 547-579.
4. T. Loughran and W. McDonald (2014): Measuring readability in financial disclosures. Journal of Finance, 69:1643–1671.
5. Molnár S., Szidarovszky F. (2002): Introduction to Matrix Theory with Applications to Business and Economics, World Scientific, London.