Diversification Under Market Inefficiency, an NLP Approach

Robert Brink   Marcus Gawronsky   Christpher Kleyweg
Ryan Kruger

19 September 2018

Abstract

A recent topic of interest in the realm of financial research has been the use of Artificial Intelligence (AI) in financial prediction. This paper explores the use of various techniques in the realm of Natural Language Processing (NLP), using a popular computational technique in deriving meaning from text data, to analyze the relationship between company association and portfolio diversification for companies on the Johannesburg Stock Exchange (JSE). Using a novel take on the Word2Vec word embedding technique, we show word-vector association between companies to track portfolio volatility over time.

1 Introduction

The Capital Asset Pricing Model (CAPM) remains popular amongst financial professionals given its simplicity and ease of use. Central to the CAPM is the assumption of efficient markets, whereby investors are able to fully diversify away unsystematic risk leaving systematic risk represented by Beta as the sole risk determinant (Strugnell, Gilbert & Kruger, 2011). However, the concept of perfect diversification prescribed by the CAPM remains challenging In South Africa or similar markets which exhibit low forms of efficiency. This leads to an underestimation of risk by South African investors who may be misled in their investment decisions through the use of CAPM or its extensions. This paper, therefore, presents an alternative risk metric known as "Association Risk" which offers investors a way in which to determine the level of diversification of a given portfolio. This metric is devoid of CAPM assumptions, using Natural Language Processing (NLP) techniques on qualitative news articles and analyst reports. This eradicates the problem of quantitative pricing data which may be unreliable or corrupted. Portfolios with lower levels of association are shown to be more diversified, exhibiting lower levels of volatility and therefore lower levels of risk.

This paper is organized as follows. Section 2 reviews the literature on the Capital Asset Pricing Model (CAPM) and introduces the use of Natural Language Processing (NLP) in the financial domain. Section 3 details the methodology, data, environment and data preprocessing. Section 4 discusses the model specification, benchmarking, experimental design and method. Section 5 contains the results of this study and is split into two main sections namely Portfolio ANCOVA and Portfolio ANCOVA with time blocking. Finally, in Section 6, we summarise and conclude our findings.

Strugnell, D., Gilbert, E. & Kruger, R. 2011. Beta, size and value effects on the JSE, 1994–2007. *Investment Analysts Journal.* 40(74):1–17. DOI: 10.1080/10293523.2011.11082537.