

# Abstract

## Using Financial Event Phrases and Keywords to Classify Form 8-K Disclosures by Likely Share Price Response

Darina M. Slattery

It is generally agreed that there are three different types of financial information: information in past stock prices, information that is available to all the public, and information that is both available to the public and available privately to insiders (Fama 1970; Haugen 1990; Hellstrom and Holmstrom 1998; Elton et al 2003). There is considerable debate about the possible impact that different kinds of information can have on the value of financial instruments. On the one hand, the efficient markets hypothesis (EMH) states that the price of a financial instrument properly reflects all available information immediately (Fama 1970). If security prices respond to all available information quickly, then the market is deemed efficient and no excess profits or returns can be made. On the other hand, fundamental and technical analysts argue that the market is inefficient because information disseminates slowly through the market and prices under- or over-react to the information (Haugen 1990).

A number of different data sources, features, goals, and methods have been used to automatically analyse content in financial documents. However, there has been very little research undertaken in the area of *automatic event phrase recognition and classification* of online disclosures. Our research study focuses on content contained in Form 8-K disclosures filed on EDGAR, a system maintained by the Securities and Exchange Commission (SEC). In our research study, we developed a prototype automatic financial event phrase (FEP) recogniser and we automatically classified a small sample of 8-Ks by likely share price response, using the automatically recognised FEPs and hand-chosen keywords as features. In four comparative classification experiments, we used the C4.5 suite of programs and the SVM-Light support vector machine program. Our datasets comprised 8-Ks filed by 50 randomly-chosen S&P 500 companies from 1997 to 2000 and 2005 to 2008.

Our research experiments yielded some interesting findings. In an experiment on the 2005 to 2008 dataset comprising 280 8-Ks, C4.5 was able to correctly classify 63.2% of the ‘ups’<sup>1</sup> (as against 58.2% at chance), when using FEPs and keywords. We also found that C4.5 appears to be better at identifying patterns in the training cases than SVM-Light, regardless of whether they were ‘ups’ or ‘downs’. When we compared the results from our FEP experiments with the results from two baseline approaches—n-gram classification and Naïve Bayes bag-of-words classification—we found that C4.5 using FEPs and keywords yielded marginally higher overall classification accuracy than C4.5 using n-grams or Naïve Bayes bag-of-words. A detailed description of the classification experiments is provided in the thesis, along with a discussion of the strengths and limitations of the research study. Recommendations for future work include further refinement of the FEPs and keywords, classification of larger datasets, and incorporation of additional classification variables beyond financial event phrases and hand-chosen keywords.

---

<sup>1</sup> By ‘ups’/ ‘downs’ we mean 8-Ks that had an increase/ decrease in share price around the filing date.