

Sự truyền tải thông điệp của dữ liệu phi cấu trúc trong dự báo kiệt quệ tài chính của các doanh nghiệp Việt Nam

TS. TRẦN THỊ HẢI LÝ

NGUYỄN THỊ HỒNG TRÂN & NGUYỄN NGỌC MY

Trường Đại học Kinh tế TP.HCM

Nghiên cứu này sử dụng nguồn dữ liệu văn bản trên Báo cáo thường niên và nguồn tin tức trực tuyến của các doanh nghiệp VN, một cách tiếp cận mới dựa trên nguồn dữ liệu phi cấu trúc, vào việc dự báo kiệt quệ tài chính. Thông tin phi cấu trúc được sử dụng riêng biệt hoặc kết hợp với các yếu tố dự báo khác dựa trên dữ liệu cấu trúc truyền thống, hướng tới mục tiêu cải thiện khả năng dự báo của các mô hình. Kết quả nghiên cứu cho thấy xét một cách độc lập thông tin phi cấu trúc có khả năng truyền tải nhiều thông điệp có ý nghĩa trong dự báo tình trạng kiệt quệ tài chính của các công ty niêm yết tại VN, đồng thời việc kết hợp giữa dữ liệu này và dữ liệu cấu trúc đại diện bởi các nhân tố truyền thống (tỷ số tài chính, đặc điểm quản trị công ty, chỉ số vĩ mô) giúp cải thiện độ chính xác của dự báo.

Từ khóa: Dữ liệu phi cấu trúc, dữ liệu cấu trúc, kiệt quệ tài chính.

1. Giới thiệu

Lĩnh vực nghiên cứu kinh tế tài chính thông thường khai thác trên nguồn thông tin dạng cấu trúc như số liệu kế toán, dòng tiền, giá cổ phiếu ... Trong thời đại bùng nổ công nghệ thông tin, nhiều lĩnh vực trên thế giới đã và đang tận dụng giá trị thông điệp của nguồn dữ liệu phi cấu trúc khổng lồ mà lĩnh vực kinh tế tài chính cũng không là ngoại lệ. Theo Turner (2013, tr.1), “Thông tin phi cấu trúc, một cách thừa nhận, truyền đạt nhiều tri thức ý nghĩa cho con người...không còn gói gọn trong lĩnh vực công nghệ thông tin mà còn lan rộng trong lĩnh vực tài chính ngân hàng. Thị trường tài chính ngân hàng cần tận dụng thông tin phi cấu trúc như một cầu nối truyền đạt tiến trình,

phân tích và tổ chức...” Các mảng tin tức trực tuyến, Báo cáo thường niên hoặc các văn bản phân tích cổ phiếu của các chuyên gia là các nguồn dữ liệu phi cấu trúc dạng văn bản phổ biến. Vấn đề phân tích mức độ truyền tải thông điệp có ý nghĩa của dữ liệu phi cấu trúc trong nghiên cứu và dự báo tài chính đang dần trở thành một xu hướng nổi bật trong thời điểm hiện nay và Li (2011, tr. 19) đã đề xuất việc sử dụng thông tin phi cấu trúc trong việc dự báo KQTC như sau: “Đã tồn tại một lượng lớn các nghiên cứu xây dựng mô hình dự báo kiệt quệ tài chính (KQTC) sử dụng các dữ liệu kế toán cũng như dữ liệu thị trường. Dữ liệu dạng văn bản từ báo cáo tài chính (BCTC) và các nguồn khác, với khả năng hàm chứa các tín hiệu dự báo cho tình

trạng KQTC của công ty, là một đề tài hết sức thú vị và mang nhiều ý nghĩa trong việc phát triển các mô hình dự báo truyền thống trước đây. Nhiều bằng chứng cho thấy các dữ liệu dạng văn bản truyền tải nhiều thông điệp trong việc dự báo tình hình hoạt động tương lai của công ty, nhiều khả năng chúng sẽ mang các thông điệp ý nghĩa khác trong việc dự báo KQTC”. Với hướng đi này, các nhà nghiên cứu kinh tế đã bước đầu đề cao tầm quan trọng của dữ liệu phi cấu trúc so với dữ liệu cấu trúc trong việc xây dựng mô hình cảnh báo KQTC (Shirata và Sakagami, 2009; Cecchini, 2010; Lu, Shen và Wei, 2013).

Ở VN, những nghiên cứu xây dựng mô hình dự báo KQTC cho tới nay hầu hết chỉ tập trung vào mảng dữ liệu cấu trúc (Lê Đạt

Chí và Lê Tuấn Anh, 2012; Hay Sinh, 2013, Nguyễn Trà Ngọc Vy và Nguyễn Văn Công, 2013). Cùng với sự gia tăng vượt trội của số lượng thông tin phi cấu trúc nói chung, việc xây dựng mô hình dự báo KQTC không chỉ đơn thuần dựa trên nguồn dữ liệu cấu trúc truyền thống mà cần thiết phải mở rộng sang cả nguồn dữ liệu phi cấu trúc.

Nghiên cứu này được thực hiện trong bối cảnh nền kinh tế và điều kiện kinh doanh tại VN còn nhiều khó khăn, tình trạng KQTC trong các doanh nghiệp VN đang là vấn đề đáng quan ngại đối với các nhà đầu tư cũng như các nhà hoạch định chính sách. Gánh chịu tàn dư của khủng hoảng tài chính toàn cầu 2008, toàn cảnh kinh tế VN nói chung cũng như khu vực doanh nghiệp nói riêng rơi vào trạng thái khó khăn và dấu hiệu phục hồi vẫn chưa rõ nét. Theo đánh giá của Ủy ban Giám sát Tài chính Quốc gia (2013) cho giai đoạn 2008-2012, tồn tại nhiều dấu hiệu cho thấy tình trạng KQTC vẫn đang đe dọa khu vực doanh nghiệp phi tài chính như: Tốc độ tăng trưởng tổng tài sản và doanh thu hằng năm giảm sút, ROA và ROE ở mức thấp, hàng tồn kho cao, nợ đọng lớn và kéo dài, chỉ số đòn bẩy cao cùng với quản trị công ty đại chúng hạn chế. Nghiên cứu này được phát triển theo hướng tiếp cận mới trên nguồn dữ liệu phi cấu trúc, hơn nữa là kết hợp giữa hai nguồn dữ liệu cấu trúc và phi cấu trúc, do vậy các mô hình dự báo KQTC được kì vọng sẽ mang lại những kết quả dự báo có giá trị, không chỉ giúp cho việc cảnh báo sớm tình trạng KQTC mà còn khởi đầu cho những nghiên cứu sử dụng thông tin phi cấu trúc ở nhiều lĩnh

vực phân tích và dự báo tài chính khác trong tương lai tại thị trường VN.

2. Phân tích thông điệp văn bản trong nghiên cứu KQTC trên thế giới

Theo Petersen (2004), dữ liệu được chia thành dữ liệu cấu trúc (dữ liệu cứng) và dữ liệu phi cấu trúc (dữ liệu mềm). Dữ liệu có cấu trúc, thường là số liệu kế toán, giá cổ phiếu, dòng tiền, chỉ số thị trường, là dạng thông tin có khả năng lưu trữ trong các cơ sở dữ liệu quan hệ (Relational Database), trong đó các thực thể và thuộc tính đã được định nghĩa sẵn, thường được thu thập trên Báo cáo tài chính (BCTC). Ngược lại với dữ liệu cấu trúc, dữ liệu phi cấu trúc được định nghĩa là dạng dữ liệu không có cấu trúc định nghĩa sẵn và không thể biểu diễn dưới dạng bảng số liệu quan hệ. Phổ biến nhất là các dạng dữ liệu văn bản dạng chữ (Text), ngoài ra còn có các dạng khác như tập tin video, tập tin ảnh, tập tin âm thanh. Theo Healy và Palepu (2001), thông tin văn bản phi cấu trúc của doanh nghiệp được chia thành hai nguồn thông tin nội bộ và đại chúng. Nguồn thông tin nội bộ của doanh nghiệp được phân định thành các công bố phi tài chính bằng văn bản dạng bắt buộc theo quy định hoặc tự nguyện, hầu hết thuộc Báo cáo thường niên (BCTN). Nguồn thông tin phi cấu trúc bên ngoài doanh nghiệp tồn tại đa dạng hơn, bao gồm các nguồn như tin đồn, các tin văn công bố về cổ phiếu trên các Sở giao dịch và các diễn đàn cổ phiếu, các phân tích tài chính độc lập, bình luận của các chuyên gia kinh tế, tạp chí tài chính, công bố của bên thứ ba (công ty kiểm toán) hoặc các bên

trung gian (ngân hàng), thông tin văn bản về thị trường, khách hàng, nhà cung cấp và đối tác....

Với khả năng hàm chứa các thông điệp có giá trị, thông tin phi cấu trúc có thể được sử dụng trong nhiều lĩnh vực nghiên cứu bao gồm cả việc cảnh báo sớm KQTC cũng như các sự kiện tiêu cực khác. Quá trình phát hiện tri thức từ nguồn dữ liệu văn bản này được gọi là Phân tích thông điệp văn bản hoặc Phân tích văn bản (Textual analysis), tương ứng với kỹ thuật khai thác văn bản (Text mining). Trong đó, nhiều tác giả sử dụng quy trình này trên nguồn thông tin phi cấu trúc nội bộ của doanh nghiệp trong xây dựng mô hình dự báo như Shirata và Sakagami (2009) dự báo KQTC theo quá trình phân tích thông tin phi cấu trúc dựa trên Báo cáo thường niên của các doanh nghiệp Nhật Bản ứng dụng kỹ thuật phân tích hình thái học cho bộ ký tự Kanji. Nghiên cứu cũng đưa ra những nhóm từ mới có tính đại diện cao cho nhóm công ty không kiệt quệ như “R&D”, “đầu tư vốn”, “kinh doanh mới”... Cecchini (2010) phát triển phương pháp phân tích nội dung thông điệp dạng tự điển từ nhằm dự báo KQTC trên phân thông tin văn bản trong mục Thảo luận và Phân tích của Ban quản trị thuộc BCTN. Tác giả tìm thấy việc dùng riêng các thông tin phi cấu trúc nhằm dự báo các sự kiện tài chính đạt được những kết quả tích cực (dự báo chính xác 75% cho nhóm công ty kiệt quệ), đồng thời trích lập được nhóm từ khóa đặc trưng cho phân thông tin mềm của nhóm công ty KQTC như “giảm lợi nhuận gộp”, “ảnh hưởng lạm phát”, “lợi ích”, “nỗ lực công ty”. Ngoài ra, Lu, Shen và Wei (2013) sử dụng nguồn thông tin mềm đại

chúng loại tin tức trực tuyến dạng Hán tự để xây dựng biến thông tin trong mô hình dự báo KQTC và nhận thấy nó có ý nghĩa trong mô hình dự báo KQTC bên cạnh các yếu tố truyền thống khác, cũng như trích lập được danh sách 110 từ khóa đặc trưng cho nhóm KQTC và 129 từ khóa cho nhóm không KQTC.

3. Phương pháp nghiên cứu

3.1. Lựa chọn nguồn dữ liệu phi cấu trúc

Nguồn dữ liệu văn bản phi cấu trúc nội bộ được thu thập trong các phần phản ánh các nội dung về kế hoạch, triển lược phát triển hoặc dự báo tương lai của doanh nghiệp trên BCTN. Các phần chi tiết được thu thập bao gồm:

- Những thông tin phải công bố theo quy định của Thông tư 52/2012/TT-BTC bao gồm:

- Triển vọng và kế hoạch phát triển tương lai thuộc báo cáo của HĐQT

- Kế hoạch phát triển trong tương lai thuộc báo cáo của Ban giám đốc.

- Những thông tin công bố tự nguyện gồm các phát biểu dự báo tương lai của công ty trong các BCTN được lựa chọn theo tiêu chí của Vu, Tower và Scully (2011) bao gồm:

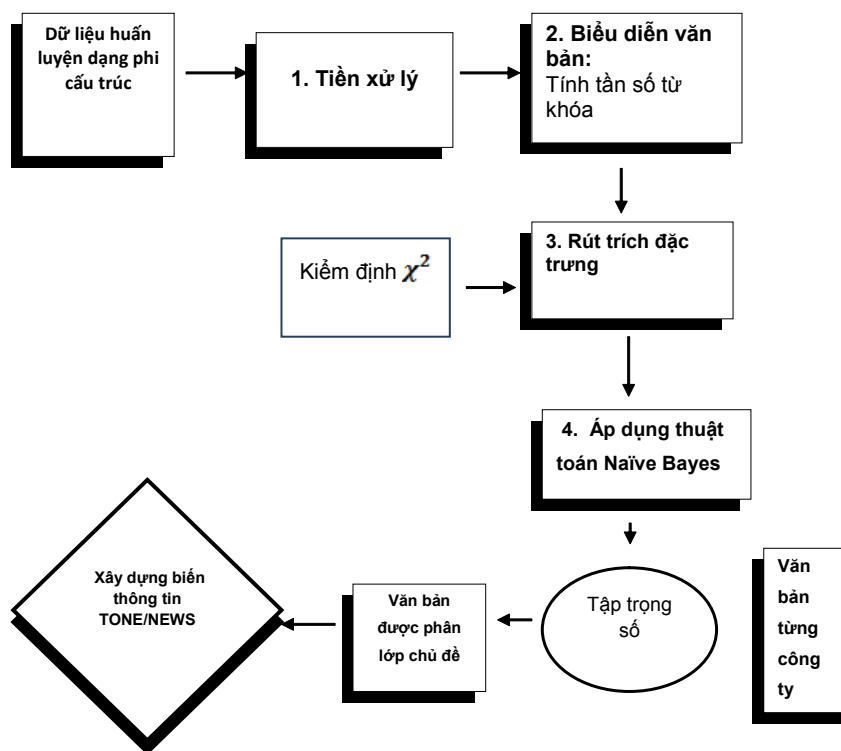
- Dự báo doanh thu, phần diễn giải văn bản định tính.

- Dự báo thu nhập, phần diễn giải văn bản định tính.

- Thảo luận về các nhân tố bên ngoài ảnh hưởng đến tình trạng hoạt động tương lai của công ty (kinh tế, chính trị và công nghệ).

- Thảo luận về các nguồn quỹ và kinh phí của công ty trong tương lai.

Hình 1: Tóm lược quy trình xây dựng biến thông tin phi cấu trúc áp dụng phân tích thông điệp văn bản



Nguồn: Tóm lược theo Li (2010) và Lu, Shen và Wei (2013).

- Thảo luận về các ảnh hưởng của tỷ giá hối đoái lên các hoạt động kinh doanh của công ty trong tương lai.

- Dự báo dòng tiền, phần văn bản diễn giải định tính.

Nguồn dữ liệu dạng tin tức trực tuyến được thu thập trên các sở giao dịch HOSE và HNX trong mục công bố thông tin của các doanh nghiệp niêm yết giai đoạn 2008 -2013. Việc lựa chọn thông tin trên các sở giao dịch nhằm đảm bảo mức độ công khai, minh bạch và độ tin cậy của thông tin.

3.2. Xây dựng biến đại diện thông tin phi cấu trúc

Biến đại diện thông tin phi cấu trúc nội bộ trong BCTN của doanh nghiệp được gọi là TONE và biến đại diện thông tin phi cấu trúc dạng tin tức trực tuyến gọi là NEWS. Các biến thông tin này được xây dựng thông qua quy trình Phân tích

thông điệp văn bản do Li (2010) và Lu, Shen và Wei (2013) đề xuất và được tóm lược trong Hình 1.

- Xác định tập trọng số

Có 20% mẫu dữ liệu ở mỗi dạng thông tin phi cấu trúc được sử dụng cho việc xác định tập trọng số để chuẩn bị cho việc xây dựng biến thông tin trên toàn bộ dữ liệu. Phần dữ liệu huấn luyện này đã được phân loại thủ công thành lớp tài liệu tương ứng với chủ đề phân lớp. Đó là phần tài liệu văn bản ứng với lớp thứ c ($c = \overline{1, n}$), n là số lớp chủ đề phân lớp văn bản tương ứng.

Dữ liệu huấn luyện thông tin phi cấu trúc dạng tin tức trực tuyến sử dụng để dự báo KQTC được phân thành hai lớp chủ đề KQTC và không KQTC. Chủ đề KQTC gồm 500 câu tin tức trực tuyến của nhóm công ty xảy ra KQTC vào năm sau, chủ đề không KQTC gồm 500 câu

tin tức trực tuyến của nhóm công ty khỏe mạnh trong năm sau.

Dữ liệu huấn luyện thông tin phi cấu trúc trên BCTN, được phân thành ba lớp chủ đề văn bản là Tích cực, Tiêu cực và Trung tính. Nguồn thông tin huấn luyện gồm 400 câu được phân loại thủ công vào ba lớp chủ đề này dựa theo nghiên cứu của Li (2010). Một câu phản ánh nội dung Tích cực khi nhà lãnh đạo đưa ra những phát biểu lạc quan, tự tin, quả quyết, nhiều khi cả tự kiêu và độc đoán về tình hình công ty tương lai. Ngược lại, một câu phản ánh nội dung Tiêu cực khi nhà lãnh đạo đưa ra các tuyên bố với tính chất bi quan, tự ti, e dè, nhiều lúc lo lắng, tức giận và buồn rầu về tình hình công ty. Các thông tin còn lại được xếp vào nội dung Trung tính.

Tiền xử lý: Tin tức dạng thô sơ được trải qua quá trình tiền xử lý (loại bỏ dấu câu, xử lý từ sai chính tả, phân tách câu rõ ràng...) nhằm đạt được dạng văn bản thống nhất và toàn vẹn theo yêu cầu của kỹ thuật khai phá văn bản và được lưu trữ trong cơ sở dữ liệu.

Biểu diễn văn bản: Tính tần số từ khóa (term frequency). Gọi $tf(c, j)$ là tần số của từ khóa t thứ j thuộc lớp tài liệu thứ c (D_c), được tính bằng số lần xuất hiện của từ khóa t trong tài liệu D_c .

Trích chọn đặc trưng: Đặc trưng được chọn là các từ khóa có hai từ. Phương pháp kiểm tra Chi bình phương χ^2 được sử dụng để giữ lại những từ có khả năng hàm chứa thông điệp và loại bỏ những từ vô nghĩa khỏi tài liệu nhằm nâng cao hiệu suất phân loại văn bản:

$$\chi^2(c, j) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

Trong đó

A: $tf(c, j)$ đại diện cho tần số từ khóa t thứ j trong lớp tài liệu D_c

B: tổng tần số của các từ khóa (không kể từ khóa t) trong lớp tài liệu D_c .

C: $tf(c', j)$, đại diện cho tần số từ khóa t thứ j trong các lớp tài liệu còn lại.

D: tổng tần số của các từ khóa (không kể từ khóa t) trong các lớp tài liệu còn lại.

N: Tổng tần số của từ khóa thứ j trong tất cả các lớp tài liệu.

Phương pháp kiểm định χ^2 nhằm kiểm tra mối tương quan giữa từ khóa thứ j và lớp tài liệu D_c . Nếu giá trị $\chi^2(c, j)$ cao hơn giá trị tới hạn, khi đó từ khóa j được xếp vào danh sách từ khóa mang tính đại diện cao cho phân lớp D_c và ngược lại, từ khóa sẽ không mang tính đại diện cho phân lớp D_c . Những từ khóa nào không mang tính đại diện cho một phân lớp nào sẽ bị loại bỏ khỏi tập văn bản.

Áp dụng thuật toán Naïve Bayes phân loại văn bản: Phương pháp Naïve Bayes được áp dụng vào quá trình phân loại. Giả định rằng sự xuất hiện của các từ trong văn bản đều độc lập với nhau.

Theo định lý Bayes:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Theo tính chất độc lập điều kiện:

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$$

Với

$P(C_i|X)$: Xác suất thuộc phân

lớp i khi biết trước câu X .

$P(C_i)$: xác suất phân lớp i .

$P(x_k|C_i)$ là xác suất thuộc tính thứ k mang giá trị x_k khi đã biết câu X thuộc phân lớp i .

Như vậy, tập trọng số $P(C_i)$ và $P(x_k|C_i)$ được xác định.

- Phân lớp văn bản trên thông tin văn bản của từng công ty

Việc phân lớp này được áp dụng lần lượt cho từng phần thông tin phi cấu trúc của mỗi công ty trong mẫu. Từ các tập trọng số được xác định ở trên, từng câu văn bản trong phần dữ liệu phi cấu trúc của một công ty sẽ được gán vào lớp chủ đề có xác suất lớn nhất theo công thức:

$$\max_{C_i \in C} \left(P(C_i) \prod_{k=1}^n P(x_k|C_i) \right)$$

- Tính biến đại diện thông tin

$$TONE_{ij} = \frac{1}{q} \sum_{q=1}^q tone_{ijq}$$

q là số câu văn bản thuộc tài liệu của công ty i .

Trong đó $tone_{ijq}$ là giá trị thông điệp trích xuất của câu văn bản thứ q trong phần dữ liệu phi cấu trúc trên BCTN của công ty i trong năm j , $tone_{ijq} = 1$ khi câu văn bản k được gán vào chủ đề Tích cực, $tone_{ijq} = 0$ khi câu văn bản k được gán vào chủ đề Trung tính và $tone_{ijq} = -1$ khi câu văn bản q được gán vào chủ đề Tiêu cực.

$TONE_{ij}$ là biến đại diện thông tin phi cấu trúc nội bộ trên BCTN của công ty i tính trung bình trong toàn năm j , thể hiện các sắc thái trong nội dung các phát biểu về tương lai của công ty của nhà quản trị.

Giá trị của $TONE$ kỳ vọng

ngược biến với tình trạng KQTC. Giá trị TONE dương và càng gần 1 chứng tỏ các phát biểu của nhà quản lý càng hàm chứa nhiều thông điệp (từ ngữ) mang màu sắc tích cực về tình trạng tương lai của công ty,

$$NEWS_{ij} = \frac{1}{q} \sum_{q=1}^Q news_{ijq}$$

q là số câu văn bản thuộc tài liệu của công ty i.

Trong đó $news_{ijq}$ là giá trị thông điệp trích xuất của câu văn bản thứ k trong phần dữ liệu phi cấu trúc dạng tin tức trực tuyến của công ty i trong năm j, $news_{ijq} = 1$ khi câu văn bản q được gán vào chủ đề KQTC, $news_{ijq} = -1$ khi câu văn bản q được gán vào chủ đề không KQTC.

$NEWS_{ij}$ là biến đại diện thông tin phi cấu trúc dạng tin tức trực tuyến của công ty i tính trung bình trong toàn năm j, $NEWS$ có giá trị dương và càng gần 1 chứng tỏ tin tức của công ty đó càng hàm chứa nhiều thông điệp (các cụm từ và câu) phản ánh nội dung về KQTC và ngược lại $NEWS$ có giá trị âm và càng gần -1 chứng tỏ tin tức của công ty đó hàm chứa nhiều thông điệp phản ánh về tình trạng tài chính khỏe mạnh của công ty. $NEWS$ kỳ vọng có mối quan hệ đồng biến với tình trạng KQTC.

3.3. Các biến đại diện thông tin cấu trúc

Ba nhóm biến số dạng cấu trúc lần lượt được đề xuất nhằm đại diện cho ba yếu tố tác động đến tình trạng KQTC, bao gồm:

- Nhóm biến tỷ số tài chính (được sử dụng trong các mô hình Beaver, 1967; Altman, 1968): Các biến số kế toán ở nghiên cứu này chia làm 4 nhóm cụ thể, đại diện

Bảng 1: Tóm tắt cách xây dựng biến thông tin phi cấu trúc và mối tương quan dự báo

Biến phụ thuộc đại diện chủ đề dự báo	Biến đại diện thông tin	Thông tin phi cấu trúc	Chủ đề phân lớp và giá trị văn bản trích xuất	Mối tương quan chủ đề dự báo
DISTRESS (= 1 khi là công ty KQTC và =0 nếu là công ty khỏe mạnh)	NEWS	Tin tức trực tuyến	2 chủ đề: KQTC (nhận giá trị là 1) và không KQTC (nhận giá trị 0)	+
	TONE	Thông tin nội bộ trên BCTN	3 chủ đề: Tích cực (nhận giá trị 1), Tiêu cực (nhận giá trị -1) và Trung tính (nhận giá trị 0)	-

Bảng 2: Các nhóm biến cấu trúc truyền thống được sử dụng trong mô hình dự báo KQTC

Nhóm biến		BIẾN	Chỉ số đại diện	Kỳ vọng dấu
Nhóm biến cấu trúc D ₁ đại diện cho chỉ số kế toán	Tỷ số thanh khoản	VLC/DT	Vốn luân chuyển trên doanh thu	-
		TM/TTS	Tiền mặt trên tổng tài sản	-
		TSNH/TTS	Tài sản ngắn hạn trên tổng tài sản	-
	Tỷ số lợi nhuận và tỷ số hiệu quả hoạt động	TNR/TTS	Thu nhập ròng trên tài sản	-
		DT/TTS	Doanh thu trên tổng tài sản	-
	Tỷ số về dòng tiền	CF/TTS	Dòng tiền trên tổng tài sản	-
	Nhóm chỉ số đòn bẩy	TN/TTS	Tổng nợ trên tổng tài sản	+
Nhóm biến cấu trúc D ₂ đại diện cho Quản trị công ty	Cấu trúc sở hữu	SDIC	Tỷ lệ sở hữu của Ban giám đốc	+/-
		SOUT	Tỷ lệ sở hữu của cổ đông đại chúng	+/-
		SGOV	Tỷ lệ sở hữu của Nhà nước	+
		SBIG	Tỷ lệ sở hữu của cổ đông lớn	+
	Đặc điểm Ban quản trị	OUTSIDE	% Số lượng giám đốc thuê ngoài	-
		BOSS	Mức độ tập quyền công ty (= 1 khi Chủ tịch HĐQT vừa là Tổng giám đốc và = 0 trong trường hợp ngược lại)	+
Nhóm biến cấu trúc D ₃ đại diện cho tình hình vĩ mô		INF	Tỉ lệ lạm phát	+
		GDPG	Tỉ lệ tăng trưởng GDP	-
		IR	Lãi suất cho vay của ngân hàng	+

cho: Tính thanh khoản, Lợi nhuận và Hiệu quả hoạt động, Dòng tiền, Chỉ số đòn bẩy. Nhóm biến này được thu thập từ BCTC đã kiểm toán của doanh nghiệp.

- Nhóm biến đại diện cho Quản trị công ty đại chúng của doanh nghiệp, được đề xuất trên hai khía cạnh chính yếu của nhân tố này bao gồm Cấu trúc sở hữu và Đặc điểm Ban giám đốc theo Woidtke, Yeh và Lee (2001), Mitton (2002), Wang và Deng, 2006; Lu, Yang-Cheng và Shu-Lien, 2009. Nhóm biến này được thu thập từ BCTN của doanh nghiệp.

- Nhóm biến đại diện cho các điều kiện kinh tế vĩ mô (theo Shumway, 2001) được thu thập từ cơ sở dữ liệu của IMF.

3.4. Xây dựng mô hình dự báo KQTC

Để đánh giá vai trò riêng lẻ và vai trò kết hợp của thông tin phi cấu trúc và thông tin cấu trúc trong việc dự báo xác suất KQTC, nghiên cứu này thực hiện ba hồi quy logistic sau:

- Mô hình 1 (chỉ bao gồm dữ liệu phi cấu trúc)

$$DISTRESS_{t+1} = \alpha_1 + \beta_1 TONE_t + \delta_1 NEWS_t + e_1$$

- Mô hình 2 (chỉ bao gồm dữ liệu cấu trúc)

$$DISTRESS_{t+1} = \alpha_2 + \sum_{i=1}^{k_1} b_i D_{i_1,t} + \sum_{i=1}^{k_2} b_i D_{i_2,t} + \sum_{i=1}^{k_3} b_i D_{i_3,t} + e_2$$

- Mô hình 3 (kết hợp dữ liệu phi cấu trúc và cấu trúc)

$$DISTRESS_{t+1} = \alpha_3 + \theta TONE_t + \alpha NEWS_t + \sum_{i=1}^{k_1} b_i D_{i_1,t} + \sum_{i=1}^{k_2} b_i D_{i_2,t} + \sum_{i=1}^{k_3} b_i D_{i_3,t} + e_3$$

Trong đó

Biến D_{i1} là biến thứ i trong nhóm biến cấu trúc D_1 đại diện cho các chỉ số tài chính được đề xuất dự báo KQTC.

Biến D_{i2} là biến thứ i trong nhóm biến cấu trúc D_2 đại diện cho Quản trị công ty đại chúng được đề xuất dự báo KQTC.

Biến D_{i3} là biến thứ i trong nhóm biến cấu trúc D_3 đại diện cho các chỉ số vĩ mô thị trường được đề xuất dự báo KQTC.

$TONE_t$ là biến sắc thái thông điệp đại diện thông tin phi cấu trúc trên nguồn BCTN, kỳ vọng có mối quan hệ âm với tình trạng KQTC của doanh nghiệp trong năm sau, tức thông điệp phi cấu trúc dự báo tương lai trong năm nay càng mang tính tích cực thì khả năng công ty rơi vào tình trạng KQTC sẽ giảm đi.

$NEWS_t$ là biến đại diện cho tin tức trực tuyến của doanh nghiệp phản ánh các sự kiện đã xảy ra trong năm t , kỳ vọng có mối quan hệ dương với tình trạng KQTC năm sau $t+1$. Tức thông tin tin tức trực tuyến càng hàm chứa nhiều thông điệp đại diện cho KQTC thì tình trạng KQTC của công ty sẽ tăng.

$DISTRESS_{t+1}$: tình trạng KQTC của doanh nghiệp vào năm sau $t+1$ có giá trị bằng 1 nếu công ty rơi vào KQTC vào năm $t+1$ và có giá trị bằng 0 nếu tình trạng doanh nghiệp là khỏe mạnh vào năm $t+1$.

Mẫu nghiên cứu

Mẫu nghiên cứu được thu thập từ danh sách các công ty niêm yết trên hai Sở giao dịch HOSE và HNX trong khoảng thời gian từ năm 2008 đến năm 2013. Tiêu chí lựa chọn các công ty KQTC là các công ty nằm trong danh sách kiểm soát và hủy niêm yết trên hai Sở giao dịch HOSE và HNX. Sau quá trình này, mẫu cuối cùng để sử dụng cho xây dựng mô hình dự báo KQTC bao gồm 199 công ty niêm yết, được phân vào hai nhóm KQTC (60 công ty) và công ty không KQTC 139 công ty

4. Kết quả thực nghiệm

4.1. Trích xuất các từ khóa đặc trưng

Bảng 3 trình bày danh sách một số từ khóa đặc trưng cho các chủ đề thông tin phi cấu trúc, danh

sách này được trích xuất dựa vào tính đại diện cao nhất của chúng cho một phân lớp chủ đề tương ứng (tức có chỉ số χ^2 cao nhất và chỉ thuộc duy nhất một phân lớp chủ đề đó, thu được từ kết quả trích chọn đặc trưng trong quy trình xây dựng biến thông tin phi cấu trúc tại thị trường VN.

Từ khóa đặc trưng cho chủ đề Tích cực, Tiêu cực và Trung tính trên thông tin BCTN

Từ khóa đặc trưng cho chủ đề Tích cực bao gồm 32 từ khóa, các từ thuộc vị trí cao nhất bao gồm: “lợi nhuận”, “doanh thu”, “đạt được”, “duy trì”, “phát triển”. Nhóm từ khóa đặc trưng cho chủ đề Tiêu cực bao gồm 22 từ khóa, các từ thuộc vị trí cao nhất là “rủi ro”, “chi phí”, “bất ổn”, “không tăng”. Nhóm từ trung tính gồm một số từ đại diện

Bảng 3: Bảng kết quả hồi quy logistic các mô hình dự báo tình trạng KQTC sử dụng thông tin phi cấu trúc, thông tin cấu trúc và kết hợp hai loại thông tin trên

Chủ đề phân lớp	KQTC			Không KQTC		
	Từ	Chi ²	Pvalue	Từ	Chi ²	P value
1	Lợi nhuận	6.038143	0.0141	Lợi nhuận	6.822827	0.0090
2	Chi phí	5.802874	0.0161	Doanh thu	5.709292	0.0332
3	Thua lỗ	5.596149	0.0181	Đầu tư	4.709292	0.0312
4	Khó khăn	5.168505	0.0232	Kế hoạch	4.397105	0.0360
5	Giảm sản	3.909986	0.0481	Chi phí	4.260868	0.0390

Chủ đề phân lớp	Tiêu cực			Tích cực			Trung tính		
	Từ	Chi ²	P value	Từ	Chi ²	P value	Từ	Chi ²	P value
1	Chi phí	6.922827	0.0095	Lợi nhuận	6.038143	0.0141	Khoảng chừng	5.411894	0.0203
2	Đổi thủ	6.822827	0.0090	Doanh thu	5.709292	0.0332	Thay đổi	4.828067	0.0280
3	Bất lợi	5.411894	0.0203	Phát triển	4.956019	0.0263	Biến đổi	4.397105	0.0360
4	Rủi ro	4.828067	0.0280	Duy trì	4.916468	0.0271	Ước tính	4.217885	0.0423
5	Khắc phục	4.350412	0.0371	Đạt được	3.875353	0.0492	Có thể	4.217885	0.0422

***, **, * đại diện lần lượt cho mức ý nghĩa thống kê 1%, 5% và 10%.

“khoảng”, “thay đổi”, “biến đổi”, “ước tính”. Các từ khóa này được tìm thấy có một số từ khóa tương đồng với Tự điển thuật ngữ kinh tế của Loughran và McDonald (2010). Danh sách các từ khóa này có vai trò quan trọng trong việc xây dựng Tự điển từ kinh tế của VN giành cho Phân tích thông điệp phi cấu trúc trong tương lai.

Từ khóa đặc trưng cho chủ đề KQTC và không KQTC

Kết quả thu được 38 từ đặc trưng cho nhóm KQTC và 78 từ đặc trưng cho nhóm không kiệt quệ tài chính. Một số từ có trọng số cao nhất trong nhóm công ty KQTC như “Chi phí”, “Thua lỗ”, “Khó khăn”, “Giảm sản”, “Lợi nhuận”, “Vi phạm”, “Tồn đọng”, “Giám đốc” và nhóm công ty không kiệt quệ như “Lợi nhuận”, “Doanh thu”, “Đầu tư”, “Kế hoạch”, “Chi phí”, “Hoạt động”, “Cổ tức”. Các

từ này đa phần được tìm thấy trên các nhóm tin về Tình hình sản xuất kinh doanh, Trả cổ tức và một số mục khác như Tình hình tuân thủ pháp luật và quy chế, Tình hình thay đổi nhân sự. Một số từ được tìm thấy có mức tương đồng cao với các từ khóa được trích xuất thuộc thông tin phi cấu trúc của các thị trường khác trên thế giới, ví dụ đối với nhóm công ty kiệt quệ tài chính là thua lỗ, chi phí và lợi nhuận (Lu, Shen và Wei, 2013 tại thị trường Đài Loan, Shirata và Sakagami, 2009 tại thị trường Nhật Bản). Một số từ mới được tìm thấy khác biệt tại thị trường VN như “chậm nộp”, “giải trình”.

Vì giới hạn trình bày nên chúng tôi chỉ liệt kê danh sách 5 từ khóa tương ứng với từng nhóm chủ đề phân lớp. Chủ đề phân lớp KQTC và không KQTC được trích xuất trên nguồn tin tức

trực tuyến. Chủ đề Tích cực, Tiêu cực và Trung tính được trích xuất trên Báo cáo thường niên của doanh nghiệp. Chỉ số Chi bình phương, ứng với mức ý nghĩa P (p-value) < 5% sẽ được chọn làm từ khóa đại diện. Kết quả thu được từ việc trích chọn đặc trưng trong quy trình xây dựng biến thông tin phi cấu trúc.

Kết quả hồi quy

Mô hình 1 dự báo kiệt quệ tài chính chỉ sử dụng các biến số phi cấu trúc (TONE, NEWS).

Dựa trên kết quả từ Bảng 4, tác giả nhận thấy cả hai biến số phi cấu trúc TONE và NEWS đều có ý nghĩa thống kê ở mức 1%. Điều này cho thấy các thông tin phi cấu trúc bên trong hoặc bên ngoài doanh nghiệp đều hàm chứa những thông điệp dự báo giá trị cho khả năng KQTC của công ty. Biến số TONE (-0.502) có mối tương quan âm

Bảng 6: Danh sách các từ khóa đại diện cho các chủ đề phân lớp tương ứng

		Mô hình 1	Mô hình 2	Mô hình 3
	Biến	DISTRESS (t+1)	DISTRESS (t+1)	DISTRESS (t+1)
Phi cấu trúc	TONE _t	-0.502***		-0.281***
	NEWS _t	0.173***		0.099***
	VLC/DT _t		-0.013**	-0.079**
	TM/TTS _t		-0.126	-0.117
	TSHH/TTS _t		-0.007	-0.015
	TNR/TTS _t		-0.103***	-0.037*
Cấu trúc	DT/TTS _t		-0.093*	-0.064*
	CF/DT _t		-0.221	-0.257
	TN/TTS _t		-0.015***	-0.061***
	SDIC _t		0.162	0.079
	SOUT _t		0.001	0.001
	SGOV _t		0.194***	0.203***
	SBIG _t		0.008*	0.039*
	OUTSIDE _t		-0.036**	-0.055**
	BOSS _t		0.025**	0.016**
	INF _t		0.001*	0.001*
	GDPG _t		-0.056	-0.042
	IR _t		0.030	0.031
	CONSTANT	0.579*	0.606*	0.634*
	-2 Log likelihood	100.621	152.438	287.266
	Cox& Snell R square	0.399	0.451	0.596
	Nagelkerke R Square	0.473	0.502	0.688
	Độ chính xác dự báo	60.601%	78.159%	89.424%
	Sai số loại I	0.475	0.449	0.365

như kỳ vọng với tình trạng KQTC, cho thấy khi các thảo luận của Ban quản trị về tương lai của doanh nghiệp mang thiên hướng Tích cực thì xác suất KQTC của doanh nghiệp trong năm sau đó sẽ giảm xuống và ngược lại. Biến số NEWS (0.173) cũng phù hợp với kỳ vọng về dấu, cho thấy nếu tin tức trực tuyến của doanh nghiệp càng hàm chứa nhiều từ khóa và nội dung đại diện cho chủ đề KQTC, sẽ làm gia tăng xác suất công ty rơi vào tình trạng KQTC trong tương lai. Cuối cùng, khi xem xét mức ý nghĩa dự báo của mô hình, tác giả thấy rằng chỉ với hai nguồn thông điệp phi cấu trúc từ bên trong và bên ngoài

doanh nghiệp có thể dự báo chính xác lên đến hơn 60% các công ty kiệt quệ hay khỏe mạnh, bước đầu khẳng định được vai trò của nguồn dữ liệu phi cấu trúc trong việc dự báo KQTC.

4.2. Mô hình 2 (chỉ sử dụng dữ liệu cấu trúc)

Kết quả mô hình 2 cho thấy khả năng dự báo KQTC của các biến số cấu trúc truyền thống. Tác giả nhận thấy có 4 trong tổng số 7 biến kế toán có ý nghĩa thống kê bao gồm biến vốn luân chuyển trên doanh thu (VLC/DT), thu nhập ròng trên tổng tài sản (TNR/TTS), doanh thu trên tổng tài sản (DT/TTS), thu nhập trên tổng tài sản (TN/

TTS). Tiếp theo, tác giả xem xét kết quả ở nhóm biến quản trị công ty. Với 6 biến quản trị được sử dụng, 4 biến có ý nghĩa thống kê bao gồm SDIC (có ý nghĩa thống kê ở mức 10%), OUTSIDE và BOSS (có ý nghĩa thống kê ở mức 5%) và SGOV (có ý nghĩa thống kê ở mức 1 %). Khi xem xét hệ số hồi quy của các biến số này, Biến số SGOV có hệ số hồi quy cao nhất (0.194) so với các biến số còn lại SDIC (0.008), OUTSIDE (-0.036), BOSS (0.025). Giá trị cao của hệ số hồi quy của biến SGOV cho thấy sở hữu nhà nước tác động đáng kể đến xác suất KQTC của các công ty. Kết quả này thống nhất với các nghiên cứu trên thế giới (như Lu, Shen, Wei, 2013). Những công ty có mức sở hữu nhà nước càng cao thì càng có xác suất KQTC cao hơn so với những công ty có mức sở hữu nhà nước thấp. Sở hữu của ban giám đốc và sở hữu cổ đông lớn càng cao thì xác suất KQTC của công ty cũng tăng cao. Ngược lại, xác suất KQTC có tương quan âm với là tỉ lệ giám đốc thuê ngoài (OUTSIDE), nghĩa là công ty càng có số lượng giám đốc thuê ngoài lớn thì càng có xác suất lâm vào tình trạng KQTC thấp. Cuối cùng, khi xét vai trò của các nhóm biến đại diện cho điều kiện vĩ mô trong việc dự báo KQTC, 3 biến số tiêu biểu được sử dụng là Lạm phát (INF), Tốc độ tăng trưởng GDP và Lãi suất (IR). Kết quả tìm thấy khá tương đồng với nhiều nghiên cứu trên thế giới chỉ có biến lạm phát có ý nghĩa thống kê nhưng chỉ ở mức ý nghĩa 10%.

4.3. Mô hình 3 (kết hợp dữ liệu phi cấu trúc và dữ liệu cấu trúc truyền thống)

Kết quả hồi quy của mô hình 3 cho thấy khả năng kết hợp của dữ

liệu cấu trúc và phi cấu trúc so với các mô hình 1 hoặc 2, chỉ sử dụng từng dạng dữ liệu riêng biệt. Các biến số phi cấu trúc TONE ($\alpha = -0.281$) và NEWS ($\alpha = 0.099$) đều có ý nghĩa ở mức 1%, dấu các hệ số vẫn như kỳ vọng và không thay đổi so với khi được hồi quy riêng lẻ ở mô hình 1. Nguồn thông tin từ bên trong doanh nghiệp (TONE) có tác động nhiều đến xác suất KQTC hơn là nguồn thông tin từ bên ngoài doanh nghiệp. Các biến số kế toán, quản trị và vĩ mô cũng có dấu và mức ý nghĩa khá tương đồng với kết quả hồi quy khi chúng được xem xét riêng (mô hình 2). Điều này cho thấy hai mảng thông tin cấu trúc và phi cấu trúc có tác dụng bổ sung cho nhau chứ không thay thế lẫn nhau trong việc dự báo xác suất KQTC. Hơn nữa, độ chính xác của mô hình kết hợp hai mảng thông tin này là cao nhất (89.424%) so với mô hình đơn lẻ (60,1% và 78.159%), cho thấy vai trò của biến số phi cấu trúc trong việc cải thiện khả năng dự báo của mô hình so với dạng mô hình dự báo trên dữ liệu cấu trúc truyền thống.

5. Kết luận và hàm ý chính sách

5.1. Kết luận

Bên cạnh các biến số tài chính luôn mang vai trò trọng yếu trong dự báo tình trạng KQTC của doanh nghiệp, bằng chứng thực nghiệm tại thị trường VN cho thấy quản trị công ty đại chúng là một nhân tố quan trọng tác động đến tình trạng KQTC của doanh nghiệp VN, điều này là phù hợp với các bằng chứng thực nghiệm tại các quốc gia châu Á khác như Trung Quốc (Wang và Deng, 2006; Lu, Yang-Cheng và Shu-Lien, 2009; Lu, Shen và Wei, 2013). Các chỉ số vĩ mô không có đóng góp nhiều trong dự báo KQTC tại VN, trừ chỉ số lạm phát

có một đóng góp nhỏ trong mô hình dự báo.

Quan trọng hơn, nguồn thông tin phi cấu trúc góp phần đáng kể trong việc dự báo KQTC. Nghiên cứu này đã xác định được danh sách các từ khóa mang đặc trưng của VN có vai trò quan trọng trong phân tích thông điệp phi cấu trúc trong tương lai, bên cạnh những từ khóa đã được tìm thấy từ các học giả trên thế giới. Nguồn thông tin mềm nội bộ của các công ty niêm yết được thu thập trên các phần thảo luận và dự báo tình hình tương lai của công ty trên BCTN đã thể hiện vai trò quan trọng trong dự báo tình trạng KQTC của doanh nghiệp VN. Khi nhà quản lý công bố những thông tin dự báo mang sắc thái tích cực về tình hình tương lai của doanh nghiệp, doanh nghiệp này sẽ có xác suất rơi vào tình trạng KQTC giảm đi. Nguồn thông tin phi cấu trúc dạng tin tức trực tuyến đăng trên các sở giao dịch chứng khoán cũng chứng tỏ khả năng dự báo xác suất KQTC của doanh nghiệp trong tương lai. Đồng thời, khi kết hợp cả hai nguồn thông tin phi cấu trúc và cấu trúc thì hiệu quả dự báo KQTC được cải thiện đáng kể.

5.2. Hàm ý chính sách

Bằng chứng thực nghiệm ở thị trường VN đã cho thấy khả năng kết hợp của hai nguồn thông tin phi cấu trúc và cấu trúc trong việc dự báo KQTC vượt trội hơn so với chỉ sử dụng nguồn dữ liệu cấu trúc truyền thống. Tuy nhiên, để khai thác được nguồn thông tin phi cấu trúc, đòi hỏi phải có những quy định nhằm đảm bảo sự thống nhất trong bình luận nội dung thông tin. Chẳng hạn xét riêng hệ thống thông tin phi cấu trúc trên Báo cáo thường niên tại VN, việc không

đồng nhất về trình bày mẫu báo cáo này hiện nay đang gây không ít khó khăn trong việc xây dựng biến số đại diện cho thông tin phi cấu trúc ở mảng này.

Về hệ thống thông tin phi cấu trúc bên ngoài doanh nghiệp, chúng tôi cho rằng việc mở rộng và phát triển một hệ thống tập trung dữ liệu tin tức trực tuyến cần được chú trọng, chúng ta có thể tham khảo các hệ thống dữ liệu thông minh ở các quốc gia trên thế giới, trong đó dữ liệu được công bố một cách công khai, minh bạch nhất đến nhà đầu tư

Kết quả nghiên cứu cũng hàm ý rằng nhà đầu tư, thị trường và các nghiên cứu tài chính trong tương lai cần chú ý tới các thông tin phi cấu trúc trong việc đánh giá, xây dựng các mô hình dự báo tài chính. Sẽ là thiếu sót nếu các mô hình dự báo bỏ qua nguồn thông tin này●

TÀI LIỆU THAM KHẢO

- Altman, E. I. (1968), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *Journal of Finance*, Vol 23 (4), 589-609.
- Beaver (1967), "Financial Ratios as predictors of Failure", *Journal of Accounting Research*, 123-154.
- Cecchini, M. (2010), "Making words work: Using financial text as a predictor of financial events", *Decision Support Systems*, Vol. 50, 164-175.
- Hay Sinh (2013), "Ước tính xác suất phá sản trong thẩm định giá trị doanh nghiệp", *Tạp chí Phát triển và hội nhập*, số 8(18), tr.52-57.
- Healy, P. M. and Palepu, K. (2000), "Information Asymmetry, Corporate Disclosure and the Capital Markets: A review of Empirical Disclosure Literature", *Journal of Accounting and Economics*, Volume 31 (1), 405-440.

(Xem tiếp trang 51)