El valor de los datos estructurados y legibles por máquina en el entrenamiento de la IA generativa



Publicado el 23 de julio de 2023 por [**Editor**](https://www.xbrl.org/news/the-value-of-structured-machine-readable-data-in-training-generative-ai/)

En la era de la inteligencia artificial (IA), existe un interés creciente en el uso de modelos de IA como ChatGPT para mejorar los procesos de información financiera. Sin embargo, es crucial reconocer la importancia de los datos estructurados y legibles por máquina para entrenar modelos de IA de manera efectiva. Un nuevo trabajo de investigación utiliza divulgaciones textuales etiquetadas con XBRL para entrenar un modelo de lenguaje grande para clasificar temas contables, con beneficios adicionales obtenidos del uso de datos etiquetados con XBRL sobre datos no estructurados.

Los investigadores, Jenna Burke (Universidad de Colorado Denver), Rani Hoitash (Universidad de Bentley), Udi Hoitash y Summer Xiao (Universidad del Noreste) aprovechan el requisito en los EE. UU. de etiquetar cada nota de los estados financieros con una etiqueta estandarizada, asignándola a un concepto de contabilidad específico en los Estándares de Contabilidad FASB. El etiquetado XBRL proporcionó a los autores un amplio conjunto de datos estructurados (más de 350 000 etiquetas XBRL) para entrenar el modelo de lenguaje grande con mayor precisión, y sin la necesidad de intentar entrenar su sistema con humanos no expertos.

El documento se centra en las etiquetas de taxonomía más frecuentes de las notas de los estados financieros, que capturan el 92,5 por ciento de las etiquetas utilizadas en las presentaciones anuales de 10-K. Por ejemplo, la etiqueta TextBlock más frecuente es "IncomeTaxDisclosureTextBlock", que es utilizada por casi todas las empresas.

Los autores eligieron un modelo de lenguaje grande entrenado en una gran cantidad de datos financieros específicos. Utilizando los datos etiquetados con XBRL, los autores "enseñaron" el modelo a clasificar el texto en temas contables. Después de enseñar el modelo, los autores examinaron su rendimiento en datos fuera de muestra nunca vistos. El modelo clasificó con precisión el texto en temas el 95 % del tiempo. Después de esto, los investigadores pusieron el modelo a trabajar clasificando el área problemática de los párrafos sin etiquetar, por ejemplo, en la discusión y el análisis de gestión, con cierto éxito.

El estudio demuestra cómo los datos etiquetados pueden ser fundamentales para entrenar grandes modelos de lenguaje, superando la necesidad (y los riesgos de precisión) de la codificación manual por parte de los humanos. La combinación de datos estructurados y legibles por máquina con IA puede mitigar los riesgos asociados con ellos, lo que permite una capacitación, consistencia e interpretabilidad precisas.

Este tipo de investigación parecería sugerir que la inteligencia aumentada de próxima generación debería ser capaz de aumentar la precisión del etiquetado. Sin embargo, la dirección debe tener la última palabra en relación con lo que se publica para los inversores. Las taxonomías oficiales XBRL de alta calidad y bien pensadas seguirán creciendo en valor y relevancia.

Lea el documento  [aquí](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4484489) .

[Datos estructurados](https://www.xbrl.org/tag/structured-data/)[DE APRENDIZAJE AUTOMÁTICO](https://www.xbrl.org/tag/machine-learning/)[de IA](https://www.xbrl.org/tag/ai/)

Uso de un modelo de lenguaje extenso para la clasificación de temas contables

[*Northeastern U. D'Amore-McKim School of Business Research Paper No. 4484489*](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4484489)

[Jenna Burke](https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=2188033)

Universidad de Colorado en Denver

[Rani Hoitash](https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=112114)

Universidad de Bentley - Departamento de Contabilidad

[Udi Hoitash](https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=514198)

Universidad del Noreste - Grupo de Contabilidad

[Verano (Xia) Xiao](https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=2287932)

Universidad del Noreste - Grupo de Contabilidad

Fecha de redacción: junio de 2023



Uso de un modelo de lenguaje grande para la clasificación de temas contables

Extracto

Ajustamos un modelo de lenguaje grande para clasificar los temas contables dentro de las divulgaciones financieras. Esto permite la clasificación eficiente y precisa de temas de contabilidad en grandes volúmenes de texto sin etiquetar fuera de la muestra. Específicamente, nuestro modelo aprovecha las innovaciones en el aprendizaje automático supervisado y los modelos de lenguaje grande para superar los desafíos de etiquetar manualmente los datos para esta tarea y supera el método de clasificación de temas más frecuente en la investigación contable y financiera (LDA). Demostramos la importancia de estas innovaciones con varios ejemplos de divulgaciones sin etiquetar: notas personalizadas a los estados financieros, la sección MD&A y la sección de factores de riesgo, que pueden clasificarse en temas según nuestro modelo. Encontramos que estas divulgaciones contienen información significativa específica del tema, que anteriormente era difícil de descubrir y es predictiva de resultados contables específicos. Los investigadores y profesionales interesados en identificar información relevante y consistente sobre temas contables a partir de grandes volúmenes de datos textuales pueden usar nuestro modelo.

Uso de un modelo de lenguaje grande para la clasificación de temas contables

Introducción

Existe un creciente cuerpo de investigación sobre temas específicos en contabilidad, donde temas como ingresos o impuestos se identifican a partir de diversos datos textuales (por ejemplo, 10-Ks, 8-Ks, cartas de comentarios, etc.). La realización de esta investigación requiere un método preciso y eficiente para medir los temas contables dentro de las divulgaciones. Actualmente, los investigadores identifican más comúnmente los temas a través de agregadores de datos, clasificación manual o realizando la asignación de Dirichlet latente (LDA). Argumentamos que las innovaciones recientes en el aprendizaje automático supervisado y los modelos de lenguaje grande son más adecuados para satisfacer esta necesidad de análisis textual y superar varios desafíos de eficiencia, integridad y precisión de los enfoques comunes. Mejorar estos aspectos de la clasificación de temas es particularmente importante cuando se considera la abundancia de datos textuales no clasificados disponibles en las divulgaciones financieras. En este estudio, presentamos un nuevo método para clasificar los temas contables dentro de las revelaciones. Específicamente, ajustamos FinBERT (un modelo de lenguaje grande introducido por Huang et al. (2023)) para realizar esta nueva tarea y comparar su rendimiento con otros enfoques de clasificación de temas. Luego aplicamos nuestro modelo a tres divulgaciones no etiquetadas: notas etiquetadas a medida para los estados financieros, Discusión y Análisis de la Administración (MD&A) y la divulgación de factores de riesgo, y utilizamos los temas clasificados para responder preguntas de investigación específicas del tema.

Investigaciones recientes sobre temas específicos encuentran que los temas contables específicos son predictores importantes de resultados materiales específicos de temas, como la calidad de los informes financieros y las divulgaciones en áreas relacionadas (por ejemplo, Ahn et al. 2020; Bens et al. 2016; Christensen et al. 2014; Dechow et al. 2016; Drake et al. 2023; Dyer et al. 2017; Inger et al. 2018; Kubick et al. 2016). Estos estudios suelen utilizar datos clasificados por agregadores de datos o clasificación manual para identificar temas. Por ejemplo, Audit Analytics es un popular agregador de datos que identifica el tema de la inexactitud, la debilidad material, la carta de comentarios y las divulgaciones críticas de asuntos de auditoría. Para utilizar agregadores de datos, los investigadores deben tener acceso a los datos y, ocasionalmente, a módulos adicionales (por ejemplo, el módulo de Auditoría Analítica de Contabilidad y Supervisión), deben usar los temas que proporciona el agregador incluso si se desea una granularidad adicional, y deben esperar hasta que los datos se clasifiquen y publiquen. La clasificación manual tiene sus propios desafíos, donde los investigadores deben dedicar un tiempo significativo a la tarea y, a menudo, deberán limitar el volumen de divulgaciones examinadas. En relación con esta corriente específica del tema, varios estudios utilizan LDA para descubrir grupos de palabras relacionadas dentro de las divulgaciones (a las que los investigadores asignan etiquetas temáticas) y detectar asociaciones con resultados agregados (por ejemplo, Bao y Datta 2014; Brown et al. 2020; Brown et al. 2021; Hoberg y Lewis 2017; Huang et al. 2018; Ryans 2021). LDA es un método de aprendizaje automático no supervisado que descubre automáticamente temas en datos textuales (Blei et al. 2003). Por ejemplo, Brown et al. (2020) utilizan LDA para descubrir temas discutidos con frecuencia en el 10-K, como la disminución de los ingresos en comparación con períodos anteriores, los costos y compromisos de combustible y los costos de eliminación de desechos nucleares. LDA es útil para descubrir temas basados en la frecuencia de las palabras y cuando los datos etiquetados no están disponibles, pero no es ideal cuando el investigador está interesado en temas contables específicos, como ingresos e inventario, ya que no hay garantía de que estos temas se descubran o se distingan entre sí. En resumen, si bien los temas identificados por los agregadores de datos, la clasificación manual y LDA producen importantes conocimientos de investigación, cada enfoque tiene sus desafíos.

Proponemos que los avances en el aprendizaje automático supervisado y los modelos de lenguaje grandes, específicamente FinBERT, pueden superar estos desafíos (Bochkay et al. 2023). Los métodos supervisados de aprendizaje automático utilizan datos etiquetados para "aprender" a clasificar datos o predecir resultados en datos fuera de muestra. Hasta donde sabemos, los métodos supervisados no se han utilizado para la clasificación de temas contables. Para esta tarea, el aprendizaje supervisado es preferible al aprendizaje no supervisado (por ejemplo, LDA) ya que inherente al diseño es la capacidad de aprender a reconocer información específica en datos textuales y posteriormente identificar información similar en datos fuera de la muestra. Además, las innovaciones recientes en el aprendizaje supervisado que incorporan una comprensión profunda del lenguaje y el contexto (es decir, modelos de lenguaje grandes) son particularmente adecuadas para esta tarea. La representación de codificador bidireccional de Transformers (BERT) fue introducida por Google en 2018 y, al igual que otros modelos de lenguaje grandes, utiliza una arquitectura de red neuronal transformadora para construir una comprensión similar a la humana del lenguaje y el contexto. Los modelos de lenguaje grandes ahora están integrados en las funciones de búsqueda, lo que permite una mejor comprensión de la búsqueda del usuario y, por lo tanto, resultados más relevantes. FinBERT, desarrollado por Huang et al. (2023), es una adaptación de BERT para texto financiero, que garantiza que el modelo entienda palabras como "asignación" que pueden tener un significado diferente en el texto financiero que en el texto general. Siano y Wysocki (2021) usan BERT y Huang et al. (2023) usan FinBERT para clasificar el sentimiento de las divulgaciones y resaltar la precisión de clasificación de estos métodos sobre otras técnicas de clasificación, incluidos otros modelos supervisados de aprendizaje automático.

El objetivo principal de nuestro artículo es afinar FinBERT para realizar la nueva tarea de clasificar los temas de contabilidad. Esto implica tomar un modelo que ya fue entrenado para entender el lenguaje (FinBERT) y enseñarlo a realizar una nueva tarea. Dado que los métodos de aprendizaje supervisado aprenden de los datos etiquetados, debemos identificar un conjunto completo de datos textuales etiquetados por temas de contabilidad para ajustar FinBERT para clasificar los temas de contabilidad.3 Las notas de los estados financieros archivadas en formato Extensible Business Reporting Language (XBRL) brindan esta oportunidad. Específicamente, XBRL requiere que los preparadores de estados financieros "etiqueten" cada nota a los estados financieros. Estas etiquetas se asignan a conceptos contables específicos en la Codificación de Normas de Contabilidad FASB, que es una fuente integral para temas de contabilidad. Los requisitos XBRL, por lo tanto, brindan una oportunidad única para utilizar una gran muestra de datos etiquetados objetivamente para ajustar FinBERT para nuestra nueva tarea. Identificamos 470,238 notas de estados financieros en presentaciones 10-K para 7,415 empresas públicas de 2009 a 2021 con datos XBRL disponibles. Utilizamos las etiquetas XBRL de taxonomía estándar más frecuentes, que se asignan a 32 temas contables distintos en la Codificación FASB.6 Además, complementamos estos datos con 28,644 divulgaciones textuales de diferentes secciones del informe anual que no están relacionadas con los 32 temas para que el modelo pueda distinguir los 32 temas contables de información no relacionada.7 Dividimos estos datos en una muestra de capacitación/ajuste fino (357,058 notas y 19,941 divulgaciones no relacionadas) y una muestra de prueba fuera de la muestra (113,180 notas y 8,703 divulgaciones no relacionadas).

En la fase de ajuste, mostramos a FinBERT el texto mencionado anteriormente y sus etiquetas. FinBERT aprende a clasificar los 32 temas de contabilidad y distinguirlos de la información no relacionada de estos datos etiquetados. Durante esta fase, FinBERT desarrolla una comprensión del lenguaje y el contexto relacionados con la contabilidad, lo que ayuda a determinar a qué tema se debe asignar una palabra o frase. Por ejemplo, puede ser difícil separar los temas de propiedad, planta y equipo (PPE) y fondo de comercio e intangibles cuando ambos incluyen discusiones sobre deterioros, y FinBERT aprende esta distinción en la fase de ajuste.

Una vez afinado, evaluamos el desempeño de nuestro modelo en la tarea de clasificación de temas contables. Durante esta prueba, el modelo solo ve el texto de las empresas invisibles y asigna una de las 33 etiquetas (32 temas contables y una etiqueta no relacionada). Nos centramos en tres métricas de rendimiento comúnmente utilizadas en el dominio del aprendizaje automático (precisión, recuerdo y puntuación F1) para evaluar la clasificación de nuestro modelo en relación con las etiquetas de temas conocidos. Los promedios ponderados para precisión, recuerdo y puntajes F1 en todos los temas están por encima del 95 por ciento. Es importante destacar que, debido a su comprensión del lenguaje y el contexto, nuestro modelo funciona bien incluso para temas que serían difíciles de distinguir a través de otros métodos. Por ejemplo, tanto el PPE como el fondo de comercio y los intangibles tienen puntajes altos de F1 en 83.14 y 98.62 por ciento, respectivamente. También realizamos varias pruebas que demuestran la solidez de nuestro modelo para diferentes opciones en la fase de ajuste.

Si bien las ventajas de eficiencia para los investigadores de usar nuestro modelo sobre los métodos no automatizados de clasificación de temas (por ejemplo, agregadores de datos y clasificación manual) son más claras, LDA es el método de aprendizaje automático más utilizado en la investigación contable y financiera, y las ventajas de nuestro método pueden ser menos evidentes. Por lo tanto, utilizamos un enfoque LDA para identificar temas de contabilidad en el mismo conjunto de datos y comparar su rendimiento con nuestro modelo. LDA es un método no supervisado para descubrir grupos de palabras relacionadas, que los investigadores luego etiquetan manualmente como temas, en divulgaciones no etiquetadas. A priori, esperamos que nuestro modelo supere a LDA por tres razones. En primer lugar, dado que LDA utiliza datos no etiquetados, puede pasar por alto temas importantes e identificar temas que no están relacionados con el objetivo de investigación de identificar temas de contabilidad. En segundo lugar, LDA utiliza un enfoque de "bolsa de palabras" que ignora el contexto y, en cambio, trata las palabras de forma independiente y se centra en su frecuencia. Por lo tanto, los temas producidos pueden ser difíciles de interpretar o etiquetar y distinguir entre sí. Por último, LDA requiere la subjetividad del investigador en varios pasos (por ejemplo, número de agrupaciones de palabras / temas para descubrir, qué etiquetas asignar a las agrupaciones de palabras resultantes y si combinar temas ex post).

Para realizar nuestra comparación con LDA, ejecutamos el algoritmo LDA en los mismos datos de entrenamiento que nuestro modelo y le pedimos que identifique 33 temas. LDA produce 33 grupos de palabras relacionadas basadas en la co-aparición en el texto de las notas. Después de etiquetar manualmente las agrupaciones de palabras y colapsar las etiquetas relacionadas, quedan 22 temas, lo que significa que 11 de los 33 temas necesarios para nuestro objetivo de investigación no están representados en la clasificación final de LDA (por ejemplo, inventario y cuentas por cobrar). Estos resultados de LDA resaltan las desventajas esperadas de LDA para la clasificación de temas de contabilidad, donde sin datos etiquetados y una comprensión profunda del lenguaje, se pueden devolver temas no relacionados, se pueden pasar por alto temas y puede ser difícil diferenciar entre temas. Para otras preguntas de investigación que requieren descubrir temas y patrones ocultos en datos no etiquetados, LDA aún puede ser preferible. Para garantizar que esta conclusión sea correcta, también usamos las agrupaciones de palabras producidas a partir de LDA para predecir temas en los datos fuera de la muestra para comparar sus métricas de rendimiento con nuestro modelo. Estas métricas varían mucho entre los temas de contabilidad y dan como resultado promedio ponderados para puntajes de precisión, recuerdo y F1 de 72.9, 71.2 y 72.0 por ciento, respectivamente. Por lo tanto, podemos concluir que nuestro modelo tiene ventajas de completitud (es decir, identifica los 33 temas necesarios) y precisión (puntaje F1 de 95.0 por ciento vs. 72.0 por ciento) sobre el uso de LDA para la tarea de clasificación de temas contables.

A continuación, para demostrar las oportunidades que presenta nuestro modelo, lo usamos para clasificar los temas contables en tres divulgaciones sin etiquetar: notas etiquetadas a medida para los estados financieros, la sección MD&A del 10-K y la sección de factores de riesgo del 10-K. Estas divulgaciones no están etiquetadas (con XBRL o de otro tipo) pero contienen información significativa específica del tema que nuestro modelo puede identificar de manera eficiente y precisa. Esto es particularmente importante ya que los agregadores de datos no cubren estas divulgaciones, la clasificación manual de estas divulgaciones requiere mucho tiempo y es algo subjetiva, y LDA no puede identificar temas específicos dentro de las divulgaciones.

En primer lugar, utilizamos nuestro modelo para identificar temas contables de notas etiquetadas a medida para los estados financieros, que están representadas por extensiones de etiquetas XBRL. Hasta donde sabemos, el tema contable de estas revelaciones no ha sido estudiado. Las extensiones de etiqueta son personalizaciones específicas de la empresa que se usan cuando un preparador de estados financieros decide que no hay una etiqueta estándar relevante. Por lo tanto, estas notas son diferentes de las utilizadas para entrenar nuestro modelo y sirven como un conjunto de datos sin etiquetar porque no hay ningún vínculo a un tema de contabilidad en la codificación. La personalización es frecuente (alrededor del 62 por ciento de las observaciones del año de la empresa revelan al menos una nota etiquetada a medida) y puede ser problemática. Usando nuestro modelo, podemos clasificar el 74.2 por ciento de las notas etiquetadas a medida como relacionadas con uno de los 32 temas contables. Los temas de contabilidad más comúnmente personalizados incluyen patrimonio, deuda y cuentas por cobrar.

En segundo y tercer lugar, utilizamos nuestro modelo para identificar temas de contabilidad dentro de las secciones de MD&A y factores de riesgo del 10-K. Estas son revelaciones narrativas importantes hechas por la gerencia que contienen información contable, sin embargo, la información no es utilizada a menudo por los investigadores, ya que está en forma de texto no estructurado y, por lo tanto, es difícil y subjetiva de recopilar. En la sección MD&A, la administración proporciona su explicación narrativa de los estados financieros y la perspectiva sobre el desempeño financiero y las perspectivas de la empresa. En la sección de factores de riesgo, se requiere que la gerencia resalte los factores más significativos que hacen que la empresa sea especulativa o arriesgada. A diferencia de las notas etiquetadas a medida, los temas más comunes de MD&A son impuestos, deuda, capital e ingresos, y los temas de factores de riesgo más comunes son compromisos y contingencias, capital e impuestos. Estos temas representan cuentas importantes para los inversores, otros responsables de la toma de decisiones e investigadores. Además, la variación en el enfoque del tema a través de estas divulgaciones es interesante.

Para demostrar las preguntas de investigación habilitadas al conocer los temas prevalentes de estas divulgaciones no etiquetadas, examinamos si el enfoque en un tema determinado está asociado con resultados materiales específicos del tema. Específicamente, examinamos la asociación entre la cobertura de un tema en notas etiquetadas a medida y el alcance de la discusión del tema en las secciones MD&A y factores de riesgo y tres resultados específicos del tema. Estos resultados incluyen (1) la probabilidad de recibir una carta de comentarios de la SEC sobre ese tema, (2) la divulgación posterior de un asunto de auditoría crítico sobre ese tema por parte del auditor externo de la compañía, y (3) re-expresiones posteriores relacionadas con el tema. Todos estos resultados tienen consecuencias negativas y, por lo tanto, es importante comprender si el enfoque de la gerencia en el tema en sus divulgaciones está asociado con los resultados. En los análisis multivariantes, encontramos consistentemente que los temas contables cubiertos en las tres divulgaciones proporcionadas por la gerencia predicen la recepción de una carta de comentarios relacionada y la posterior divulgación de un asunto de auditoría crítico relacionado. Los resultados de las declaraciones erróneas son mixtos, tal vez debido a la menor frecuencia de declaraciones erróneas sobre temas específicos, pero no obstante revelan ciertas asociaciones significativas para cuentas importantes para los inversores (por ejemplo, una mayor cobertura de ingresos e impuestos en el MD&A se asocia con declaraciones erróneas de ingresos e impuestos). Estos hallazgos de temas específicos económicamente significativos son nuevos en la literatura, y los análisis no serían posibles sin un método para clasificar de manera eficiente y precisa las grandes muestras de divulgaciones no etiquetadas en temas contables.

Nuestro estudio pertenece a la creciente investigación que utiliza grandes modelos de lenguaje para comprender el texto de una manera similar a la humana (por ejemplo, Kim et al. 2023). A medida que esta investigación se desarrolla y las divulgaciones financieras se vuelven más complejas, los métodos de análisis textual con una fuerte validez interna son cruciales. Creamos un modelo para identificar temas contables en grandes volúmenes de divulgaciones y demostrar que funciona de manera eficiente y precisa y, lo que es más importante, supera los desafíos de usar temas identificados por agregadores de datos, clasificación manual y LDA. De esta manera, extendemos el aprendizaje supervisado y las aplicaciones de modelos de lenguaje grandes, particularmente BERT y FinBERT (Huang et al. 2023), a una nueva tarea. Antes de que estos métodos estuvieran disponibles, los primeros estudios como Brown et al. (2020) utilizan LDA para identificar el tema de los datos textuales en el 10-K y examinar las asociaciones con los resultados agregados. Construimos y ampliamos esta literatura aprovechando las innovaciones recientes para identificar temas contables específicos en el 10-K y examinar las asociaciones con los resultados específicos del tema (por ejemplo, la medida en que los ingresos están cubiertos en el MD&A y las cartas de comentarios relacionados con los ingresos, asuntos críticos de auditoría y declaraciones erróneas).

Nuestros análisis multivariados documentan que los temas contables cubiertos en divulgaciones no etiquetadas son útiles para predecir la identificación de la SEC de deficiencias de divulgación (cartas de comentarios) y áreas desafiantes de la auditoría (asuntos críticos de auditoría y declaraciones erróneas), lo que demuestra que nuestro modelo puede usarse como una herramienta para mejorar la identificación de preguntas de investigación específicas del tema. Los futuros investigadores también pueden usar nuestro modelo para identificar temas contables en nuevas divulgaciones, muestras y períodos de tiempo. Por ejemplo, además de las divulgaciones que utilizamos en nuestras aplicaciones de investigación, las revelaciones fuera de los estados financieros y otros textos no estructurados se pueden clasificar en temas. Los inversores y otros usuarios de estados financieros también pueden usar nuestro método para clasificar la información del tema en las divulgaciones a medida que se publican e incorporar esta información específica del tema en sus decisiones.

Antecedentes

Investigación temática específica en contabilidad

La investigación sobre temas específicos es cada vez más frecuente en contabilidad, y los investigadores examinan temas contables específicos, como ingresos o impuestos, que están cubiertos en divulgaciones más amplias. Por ejemplo, la investigación encuentra que los temas de contabilidad son predictores importantes de los resultados materiales. Específicamente, los estudios examinan la relevancia del valor de cuentas específicas (Barth et al. 2023) y temas específicos de las cartas de comentarios de la SEC y los resultados relacionados (Bens et al. 2016; Dechow et al. 2016; Kubick et al. 2016). Los estudios también encuentran que la experiencia del auditor en áreas contables específicas está asociada con la calidad de la información financiera y otros resultados en esas áreas (Ahn et al. 2020; Christensen et al. 2014; McGuire et al. 2012). Además, los estudios encuentran que cuando los auditores destacan ciertos temas en sus divulgaciones, los gerentes realizan cambios en sus divulgaciones en estas áreas (Burke et al. 2023; Drake et al. 2023), y los inversores cambian su percepción de las noticias relacionadas (Abbott y Buslepp 2022).

Estos estudios generalmente identifican temas de contabilidad con clasificación manual o a través de un agregador de datos como Audit Analytics, que identifica el tema de errores erróneos, debilidades materiales, cartas de comentarios y divulgaciones críticas de asuntos de auditoría. Si bien estos métodos pueden identificar con precisión los temas de contabilidad, ambos pueden tener desventajas. Específicamente, la clasificación manual requiere un tiempo significativo, y los investigadores a menudo necesitarán limitar el volumen de divulgaciones examinadas. Los datos proporcionados por los agregadores pueden ser costosos, especialmente cuando se necesitan módulos adicionales. Además, los investigadores, inversores y otras partes interesadas deben esperar hasta que se clasifiquen los datos, lo que puede ser meses o años después de que se publique la divulgación. Los usuarios también deben usar los temas proporcionados incluso si se desea una granularidad adicional o una clasificación diferente (por ejemplo, ingresos e ingresos diferidos, deuda y capital se combinan en la codificación de Audit Analytics).

Análisis textual para reconocer temas contables

Dados los avances en la potencia informática, los datos textuales son la forma de datos de más rápido crecimiento en la investigación en ciencias sociales (Bochkay et al. 2023; Ke et al. 2019). En lugar de la clasificación manual o el uso de agregadores de datos, los investigadores y otras partes interesadas pueden utilizar el análisis textual para clasificar los temas de contabilidad en los datos textuales con la granularidad y la línea de tiempo necesarias para su análisis. Además, el análisis textual puede permitir que varias partes interesadas accedan a divulgaciones que los agregadores de datos no cubren. La clasificación de temas de manera eficiente y precisa y con una fuerte validez interna depende del método de análisis textual seleccionado.

Asignación de Dirichlet latente (LDA)

La asignación latente de Dirichlet (LDA), introducida por Blei et al. (2003), es el método más utilizado para el análisis de temas en la investigación contable y financiera. LDA es un método de aprendizaje automático no supervisado, lo que significa que no está expuesto a datos etiquetados y es el "trabajo de la máquina" descubrir atributos latentes (es decir, "temas") en los datos. Los "temas" producidos por LDA son grupos de palabras relacionadas (Loughran y McDonald 2020). Los investigadores asignan etiquetas temáticas basadas en las palabras que ocurren con mayor frecuencia en el grupo y su interpretación de las conexiones entre las palabras en el grupo.

Varios estudios contables utilizan LDA, más comúnmente para descubrir temas de divulgaciones completas o partes de divulgaciones. Este método se utiliza a menudo porque la mayoría de las divulgaciones contables no están etiquetadas. Utilizando temas identificados por LDA, los estudios han detectado asociaciones con resultados agregados como la reacción y la atención del mercado (Bao y Datta 2014; Huang et al. 2018; Ryans 2021), declaraciones erróneas (Brown et al. 2020; Hoberg y Lewis 2017; Ryans 2021), y relevancia del valor (Brown et al. 2021). Relacionado con nuestro enfoque en temas de contabilidad está Ryans (2021), que utiliza LDA para identificar temas de cartas de comentarios. Si bien muchos de los temas identificados son consistentes con el objetivo de identificar temas de información financiera, otros no lo son.16 Del mismo modo, Brown et al. (2020) utilizan LDA para descubrir temas en las presentaciones 10-K, que incluyen cuestiones contables y no contables, como la disminución de los ingresos en comparación con períodos anteriores, los costos y compromisos de combustible, los costos de eliminación de desechos nucleares, etc.

Si bien el uso de LDA en la divulgación de información contable ha producido muchas ideas importantes, existen varios desafíos al usarlo para la tarea de clasificación de temas contables. En primer lugar, dado que, por definición, el aprendizaje no supervisado no aprende de un conjunto de datos etiquetados, LDA es útil cuando se buscan temas ocultos, pero menos ideal cuando se buscan temas específicos, ya que puede pasar por alto temas y devolver temas no relacionados con el objetivo de la investigación. En segundo lugar, el algoritmo LDA utiliza un enfoque de "bolsa de palabras" para analizar el texto, lo que significa que ignora el contexto y el orden de las palabras y, en cambio, trata una colección de palabras individuales de forma independiente y se centra en la frecuencia (Loughran y McDonald 2016; Huang et al. 2023). Una palabra ambigua como "deterioro" podría estar discutiendo buena voluntad, EPP o reestructuración, y LDA no podría distinguir según el contexto, lo que puede llevar a la clasificación errónea de los temas. Por último, se requiere la subjetividad del investigador en varios puntos del proceso LDA, lo que puede afectar su producción y replicabilidad en todos los estudios.

Innovaciones adecuadas para la tarea de clasificación de temas contables

En general, si bien LDA ha tenido un impacto y ha avanzado en el uso de datos textuales en la investigación contable, las innovaciones en el aprendizaje automático supervisado y los modelos de lenguaje grande brindan nuevas oportunidades para una validez interna más sólida para la tarea de clasificación de temas contables. Esto es particularmente cierto ya que la creciente complejidad de las revelaciones contables y la prevalencia de las revelaciones no estructuradas y no etiquetadas requieren métodos más sofisticados y exigentes de análisis de texto / clasificación de temas.

Los métodos supervisados de aprendizaje automático entrenan un modelo en datos etiquetados, donde el modelo "aprende" de un conjunto de datos etiquetado para clasificar datos o predecir resultados con precisión. Hasta ahora, el aprendizaje supervisado se ha utilizado para tareas de predicción en la investigación contable.18 Sin embargo, también es útil cuando se buscan construcciones específicas en datos textuales (es decir, tareas de clasificación). Específicamente, un modelo puede ser entrenado de manera supervisada en texto con etiquetas definidas por expertos y luego aplicado a datos no etiquetados fuera de la muestra. A diferencia de LDA y otros métodos no supervisados, esto asegura que los temas deseados serán devueltos.

**Modelos de lenguaje grandes.**  Los modelos de lenguaje grandes son innovaciones recientes que incorporan una comprensión profunda del lenguaje y el contexto para construir una comprensión humana del texto. Los modelos de lenguaje grandes comúnmente usan una arquitectura de red neuronal transformadora, que se basa en un modelo de lenguaje enmascarado y la predicción de la siguiente oración, para desarrollar esta comprensión similar a la humana. Al entrenar en el modelo de lenguaje enmascarado, estos modelos ocultan una palabra en una oración y hacen que el programa prediga qué palabra se ha ocultado (enmascarado) en función del contexto de la palabra. Esto ayuda al modelo a entender el contexto dentro de una oración. Al entrenar en la predicción de la siguiente oración, el modelo desarrolla una comprensión del orden y el contexto a través de diferentes oraciones al predecir si dos oraciones están conectadas lógica o aleatoriamente.

Dentro de la categoría de modelos de lenguaje grande, Google introdujo la representación de codificadores bidireccionales de Transformers (BERT) en 2018. Siano y Wysocki (2020) utilizan BERT para clasificar el sentimiento (positivo / negativo) de los anuncios de ganancias de las empresas y resaltar su precisión en comparación con otras técnicas de clasificación. Más recientemente, Huang et al. (2023) introdujeron FinBERT, que es una adaptación de BERT para el dominio financiero. Los autores entrenan previamente su modelo para comprender el lenguaje en varios textos financieros, incluidas las presentaciones anuales y trimestrales, los informes de analistas financieros y las transcripciones de conferencias telefónicas de ganancias.20 Luego usan FinBERT para reconocer el sentimiento con una muestra de oraciones etiquetadas por investigadores de informes de analistas y demostrar su desempeño superior en la identificación del sentimiento de texto fuera de muestra.

Los modelos de lenguaje grandes como BERT y FinBERT tienen aplicaciones sin explotar en la investigación contable más allá de la clasificación de sentimientos. Específicamente, estos métodos se pueden enseñar a reconocer temas contables dentro de las divulgaciones que, lo que es más importante, superan las trampas de la clasificación manual, los agregadores de datos o LDA para esta tarea. La ventaja de ajustar FinBERT para la clasificación de temas contables incluye el uso de datos etiquetados para entrenar el modelo, lo que garantiza que se identifiquen temas relevantes y significativos, en lugar de aleatorios. El modelo también se puede aplicar a otras divulgaciones no etiquetadas y siempre identificará los temas preestablecidos. Esta ventaja es particularmente útil para identificar temas de contabilidad, ya que existe un conjunto conocido de temas. Además, la comprensión similar a la humana del contexto y el orden de las palabras es importante para reconocer temas de contabilidad. El contexto puede determinar a qué tema se debe asignar una palabra o frase. Por ejemplo, las revelaciones tanto sobre combinaciones de negocios como sobre derivados y coberturas implican discusiones sobre el valor razonable. El uso de un modelo de lenguaje grande como FinBERT para separar los dos temas, incluso si se usan palabras similares, debería producir resultados más precisos.

Método y resultados

Conjunto de datos etiquetado: Temas de contabilidad de XBRL y FASB Codificación

Para ajustar FinBERT para realizar nuestra nueva tarea de reconocer temas de contabilidad, le mostramos datos textuales y etiquetas de temas correspondientes. La obtención de datos etiquetados suele ser costosa porque requiere la clasificación manual del texto. Para lograr su tarea de identificar el sentimiento, Huang et al. (2023) etiquetan manualmente 10,000 oraciones de los informes de los analistas financieros como positivas, negativas o neutrales. Superamos los desafíos de la clasificación manual aprovechando Extensible Business Reporting Language (XBRL) para identificar texto que se refiere a temas de contabilidad.

Las empresas públicas están obligadas a producir informes financieros en XBRL y "etiquetar" cada valor numérico y nota de estado financiero. Este último es particularmente adecuado para nuestra tarea por varias razones. En primer lugar, las notas de los estados financieros cubren un amplio conjunto de temas contables, que es un requisito esencial para nuestro objetivo de investigación. En segundo lugar, se requiere que los preparadores etiqueten cada nota con una etiqueta estandarizada (es decir, una etiqueta de taxonomía) que asigna la nota a un concepto de contabilidad específico en la Codificación de Normas de Contabilidad del FASB, que es la fuente más completa de temas de contabilidad.21 Esto nos proporciona un amplio conjunto de datos etiquetados (más de 350,000 etiquetas XBRL en el conjunto de entrenamiento), que normalmente no está disponible en otras aplicaciones de aprendizaje automático. La utilización de la asignación a la codificación también elimina la subjetividad y el error humano del proceso de etiquetado. Por último, las notas incluyen una cantidad grande y diversa de texto específico del tema, lo que permite a FinBERT aprender qué texto está asociado con qué temas de contabilidad con mayor precisión. Utilizamos este gran conjunto de datos etiquetado para desarrollar nuestro modelo y luego para probar su rendimiento.

Para recopilar notas de estados financieros, utilizamos Python para descargar los datos XBRL de todas las empresas públicas de 2009 a 2021. Limitamos la muestra a empresas con cobertura en Compustat, lo que resulta en 7,415 empresas. Analizamos todas las etiquetas XBRL y mantenemos solo las etiquetas de las notas de los estados financieros. Nos centramos en las etiquetas de taxonomía con la frecuencia más alta, lo que resulta en 85 etiquetas distintas que juntas representan el 92.5 por ciento de todas las etiquetas de taxonomía divulgadas en las presentaciones anuales de 10-K. Por ejemplo, la etiqueta más frecuente es "IncomeTaxDisclosureTextBlock", que es utilizada por casi todas las empresas y aparece 43.848 veces en nuestro conjunto de datos final. Cada etiqueta se asigna a un concepto de contabilidad en la codificación FASB, y puede haber varias etiquetas asignadas a cada concepto contable y, por lo tanto, recibir la misma etiqueta. Por ejemplo, las etiquetas "BusinessCombinationDisclosureTextBlock", "MergersAcquisitionsAndDispositionsDisclosuresTextBlock" y "BusinessAcquisitionIntegrationRestructuringAndOtherRelatedCostsTextBlock", se asignan al tema de combinaciones de negocios de la codificación. La lista completa de las 85 etiquetas XBRL y su asignación a temas de contabilidad se muestra en el Apéndice A.

Después de procesar los datos de texto de todas las notas y eliminar las notas y empresas atípicas,25 el resultado son 470.238 notas que cubren 32 temas contables. Dividimos aleatoriamente esta muestra en 70 por ciento de las empresas (357,058 notas) en la muestra de ajuste fino y 30 por ciento de las empresas (113,180 notas) en la muestra de prueba. Es importante que obliguemos a la muestra de prueba a pertenecer a un conjunto distinto de empresas que no se incluyeron en la muestra de ajuste fino porque la descripción textual de las notas es pegajosa con el tiempo y el uso de las notas de la misma empresa en los datos de ajuste fino y prueba proporcionaría una ventaja injusta a los modelos.

Complementamos los datos de las notas XBRL con 28.644 observaciones no relacionadas con los 32 temas para que el modelo pueda distinguir el texto relacionado con los temas contables del texto no relacionado en las divulgaciones financieras. Específicamente, para las empresas de nuestra muestra, obtenemos las secciones de descripción comercial (Ítem 1), MD&A (Ítem 7) y factor de riesgo (Ítem 1A) del 10-K de Calcbench. Luego utilizamos un proceso de dos pasos para identificar párrafos de estas divulgaciones que no están relacionados con ninguno de los 32 temas contables. Primero, utilizamos un diccionario de palabras contables generado por el investigador (por ejemplo, crédito, acciones comunes, deteriorado) para contar el número de palabras contables en cada párrafo. Los párrafos sin palabras contables se codifican automáticamente como no relacionados.26 En segundo lugar, seleccionamos aleatoriamente 1,500 párrafos de cada divulgación (es decir, MD&A, factor de riesgo y descripción comercial) con al menos 30 palabras, entre una y cinco palabras contables de nuestro diccionario, y que no contienen signos de dólar. Estos criterios tienen como objetivo capturar las divulgaciones de área gris que podrían estar relacionadas o no con un tema contable. Dos investigadores codificaron manualmente estas 4.500 observaciones como relacionadas o no relacionadas. Combinados, el 78.4 por ciento de los párrafos del área gris se clasifican como no relacionados. Por ejemplo, en la sección MD&A, la gerencia puede discutir la perspectiva económica general, el mercado objetivo o la importancia del equipo de alta gerencia, y estas discusiones no están relacionadas con los 32 temas a pesar del uso de palabras contables. Este proceso de dos pasos garantiza que nuestro modelo aprenderá a identificar texto no relacionado y, por lo tanto, no se ajuste demasiado al forzar todo el texto en uno de los 32 temas de contabilidad. Seleccionamos aleatoriamente el 70 por ciento de estas divulgaciones no relacionadas para la muestra de ajuste fino y el 30 por ciento para la muestra de prueba.

Creación de un modelo para la clasificación de temas contables

FinBERT personaliza el modelo de lenguaje grande Bidirectional Encoder Representation from Transformers (BERT) de Google para comprender textos financieros. BERT desarrolla una comprensión del lenguaje y el contexto a partir de un gran volumen de texto, incluidos Wikipedia y BookCorpus (Devlin et al. 2019), y FinBERT expone a BERT a presentaciones corporativas, informes de analistas y transcripciones de conferencias telefónicas de ganancias (Huang et al. 2023). Como se discutió en nuestra revisión de la literatura, ambos son modelos de lenguaje grandes que utilizan una arquitectura de red neuronal transformadora para construir una comprensión similar a la humana del lenguaje y el contexto.

Descargamos el modelo FinBERT proporcionado por Huang et al. (2023) en HuggingFace y lo ajustamos para realizar nuestra nueva tarea de reconocer temas de contabilidad en la muestra de ajuste fino. El proceso de ajuste fino en el aprendizaje automático implica tomar un modelo que ya fue entrenado para entender el lenguaje y hacer ajustes para que pueda aprender una tarea específica. Huang et al. (2023) ajustan FinBERT para clasificar el sentimiento de las divulgaciones. En contraste, ajustamos FinBERT para clasificar el texto financiero en uno de los 33 temas. Dado que el modelo FinBERT ya está capacitado para comprender el lenguaje y el contexto financiero, proporcionamos al modelo el texto completo de las divulgaciones y no realizamos una reprocesamiento adicional del texto, como la lematización o la eliminación de palabras vacías. También proporcionamos las etiquetas de los temas y el modelo aprende a clasificar los 32 temas de contabilidad y el tema no relacionado de estos datos etiquetados.

Prueba del rendimiento del modelo con datos fuera de la muestra

A continuación, evaluamos el rendimiento utilizando nuestro modelo ahora ajustado para predecir temas en la muestra de prueba. En esta etapa de prueba, nuestro modelo ve solo el texto de las empresas invisibles y asigna una de las 33 etiquetas temáticas. A continuación, podemos evaluar el rendimiento de la clasificación en relación con el tema conocido real.

La Tabla 2 muestra una matriz de la clasificación de nuestro modelo de cada observación (columnas) en relación con el tema real de cada observación (filas). Por ejemplo, la muestra de prueba incluye 4.130 observaciones en el tema de inversiones (la suma de valores en la fila 17). De las 4.197 observaciones identificadas por el modelo como inversiones (la suma de los valores en la Columna 17), 3.925 están realmente relacionadas con el tema de las inversiones (la intersección de la Fila 17 y la Columna 17) capturando el 95,04 por ciento de las notas reales, y 272 están clasificadas erróneamente como el tema de inversión. En general, la matriz sigue un patrón diagonal claro que representa una clasificación precisa.

Tres métricas de rendimiento se usan comúnmente en el dominio del aprendizaje automático: precisión, recuperación y la puntuación F1. La figura A ilustra una matriz de confusión, que es una técnica común para resumir el rendimiento de un algoritmo de clasificación en el aprendizaje automático. Nuevamente usamos el tema Inversiones como ejemplo en esta Figura. En primer lugar, la precisión es un porcentaje calculado como el número de observaciones correctamente clasificadas en un tema (Verdadero Positivo) escalado por el número total de observaciones clasificadas en ese tema (Verdaderos y Falso Positivos). La precisión de nuestro modelo para las inversiones es del 93,52 por ciento, calculada como 3.925 verdaderos positivos divididos por la suma de verdaderos positivos y falsos positivos (3.925+272). Esto ilustra que al clasificar el tema Inversiones, nuestro modelo tiene una alta precisión y, por lo tanto, devuelve clasificaciones significativamente más correctas en relación con las clasificaciones incorrectas. La segunda métrica, recordar, es un porcentaje calculado como el número de observaciones correctamente clasificadas en un tema (True Positive) escalado por el número real de observaciones en ese tema (True Positives y False Negatives). El retiro de inversiones de nuestro modelo es del 95.04 por ciento, calculado como 3,925 verdaderos positivos divididos por la suma de verdaderos positivos y falsos negativos (3,925 + 205). Por último, la puntuación F1 combina estas dos métricas por su media armónica.30 La puntuación F1 para inversiones para nuestro modelo es 94.

La Tabla 3 presenta estas métricas de rendimiento para cada uno de los temas. La precisión varía de 77.3 a 100.0 por ciento, y el recuerdo varía de 70.2 a 100.0 por ciento. Nuestro modelo funciona particularmente bien en la identificación de temas comunes, así como temas que son importantes para los investigadores e inversores (por ejemplo, los ingresos y los impuestos tienen puntajes F1 de 96.3 y 100.0 por ciento, respectivamente). Además, debido a su comprensión del contexto, nuestro modelo tiene éxito en distinguir entre temas que pueden contener lenguaje relacionado. Por ejemplo, aprende qué discusiones sobre el deterioro se relacionan con el fondo de comercio y los intangibles, el PPE y la reestructuración durante el ajuste fino y se desempeña con una puntuación F1 de 98.6, 83.1 y 95.9 por ciento para estos temas en las pruebas.

En general, el promedio ponderado de puntuación F1 de nuestro modelo es del 95,0 por ciento, que es incluso más alto que el rendimiento de BERT o FinBERT en la clasificación de sentimientos en textos financieros (tasas de precisión del 77,0 y 88,2 por ciento, respectivamente, Huang et al. 2023). El rendimiento de nuestro modelo demuestra que es muy adecuado para la tarea de clasificar temas contables.

Ensayos de robustez

Realizamos varias pruebas para examinar la robustez de nuestro modelo a diferentes opciones en la fase de ajuste fino. Primero, evaluamos el rendimiento de nuestro modelo cuando usamos muestras de entrenamiento más pequeñas. Si bien nuestra muestra de entrenamiento primario es de 376,999 observaciones, usamos el 5 por ciento y el 20 por ciento de esta muestra en estas pruebas alternativas.31 La Tabla 4 muestra el promedio ponderado fuera de la muestra para precisión, recuerdo y puntaje F1 para cada una de estas muestras. Estas estadísticas son solo ligeramente más bajas que cuando se usan muestras más grandes, lo que sugiere que nuestro modelo conserva una alta precisión incluso cuando se entrena con muestras más pequeñas.

En segundo lugar, utilizamos ventanas rodantes para examinar el rendimiento de nuestro modelo cuando se entrena solo en años específicos y se prueba en años no vistos de datos para empresas no vistas. Específicamente, entrenamos nuestro modelo en un período de tres años y lo probamos en el cuarto año invisible (por ejemplo, 2011-2013 para el ajuste y 2014 para las pruebas), donde la muestra de prueba incluye compañías distintas de la muestra de capacitación. Repetimos esto ocho veces para que las ventanas enrollables cubran nuestro período de muestra. El promedio ponderado fuera de la muestra para la precisión, el recuerdo y la puntuación F1 para cada una de estas ventanas rodantes también se muestra en la Tabla 4. Los resultados muestran nuevamente que el modelo sigue siendo altamente preciso, con una puntuación F1 de más del 94 por ciento para cada una de las ventanas rodantes.

Comparando nuestro modelo con LDA

LDA también es un método automatizado de aprendizaje automático y es el método más utilizado para la clasificación / identificación de temas en la investigación financiera y contable. En esta sección, usamos LDA para identificar temas de contabilidad en el mismo conjunto de datos de ajuste fino y comparar su rendimiento con nuestro modelo en el conjunto de datos de prueba.

LDA es un enfoque no supervisado que utiliza datos no etiquetados para descubrir temas en las divulgaciones. Para comenzar, los investigadores deben ingresar el número de temas a descubrir. Por lo general, el número de temas es desconocido y se determina a través de varias iteraciones y requiere juicio humano durante todo el proceso. En nuestro entorno, sabemos a priori que el texto pertenece a 32 temas de contabilidad y 1 tema no relacionado. Por lo tanto, configuramos el modelo LDA para descubrir 33 temas.

LDA se muestra solo el texto de las notas33 (es decir, no las etiquetas) y genera 33 "temas", que son grupos no etiquetados de palabras relacionadas (se usan términos de una, dos y tres palabras). Las 20 palabras más frecuentes en cada tema se muestran en la Tabla 5. Siguiendo la literatura previa (Brown et al. 2020; Dyer et al. 2017), luego usamos el juicio para etiquetar cada uno de los 33 temas identificados por la LDA en temas interpretables.34 El proceso de etiquetar temas es difícil. A pesar de que se conocen las etiquetas temáticas, los investigadores tuvieron un desacuerdo del 36 por ciento en el etiquetado porque solo se observan palabras específicas y su frecuencia, no su contexto. Por ejemplo, el Tema 25 se etiquetó inicialmente como deuda y capital. Este desacuerdo subraya la subjetividad involucrada en el método LDA.

Luego colapsamos las etiquetas superpuestas en el mismo tema. Por ejemplo, los temas 6 y 28 están etiquetados como compromiso y contingencias y se contraen en uno. El LDA colapsado produce 22 temas únicos. Como resultado, LDA no identificó 11 de los 32 temas contables reales cubiertos en las notas. Las categorías contables no identificadas son: retiro de activos, efectivo, ingresos diferidos, seguros, inventario, pasivos, otros activos, cuentas por cobrar, investigación y desarrollo, transferencias y servicios, y garantías y garantías. La falta de 11 temas subraya un inconveniente de usar LDA cuando se conocen los temas.

Prueba del rendimiento de LDA

A continuación, evaluamos el rendimiento de LDA en datos fuera de muestra. Específicamente, usamos las agrupaciones de palabras etiquetadas de la Tabla 5 para predecir temas en la muestra de prueba. LDA busca estas agrupaciones de palabras en el texto y asigna la etiqueta del tema en función de la frecuencia de las palabras que coinciden.

Siguiendo el mismo formato que la Tabla 2, la Tabla 6 muestra una matriz de la clasificación de LDA de cada observación (columnas) en relación con el tema real de cada observación (filas).35 Por ejemplo, de las 3.405 observaciones identificadas bajo el tema de inversiones ahora etiquetado (columna "Inversiones"), 2.272 están realmente relacionadas con el tema de inversiones y 1.133 están clasificadas erróneamente como el tema de inversión. Por lo tanto, la precisión para las inversiones es del 66,73 por ciento (frente al 93,52 por ciento de nuestro modelo). De los 4,130 temas de inversiones reales (fila "Inversiones"), 1,858 se clasifican en otros temas, lo que lleva a un retiro del 55.01 por ciento (frente al 95.04 por ciento de nuestro modelo). La puntuación F1 para el tema de inversiones es del 60,31 por ciento (frente al 94,27 por ciento de nuestro modelo). En conjunto, la Tabla 6 muestra la variabilidad en el éxito de LDA en la clasificación de notas en temas contables.

Aplicaciones de investigación

La capacidad de nuestro modelo para clasificar de manera precisa y eficiente los temas de contabilidad se puede utilizar para responder preguntas de investigación específicas del tema. Para demostrar este potencial, utilizamos nuestro modelo para clasificar los temas contables en tres divulgaciones sin etiquetar. Los agregadores de datos no recopilan información temática de estas divulgaciones y la clasificación manual llevaría mucho tiempo y sería algo subjetiva. Utilizamos los temas identificados para examinar si la cobertura de temas contables específicos es indicativa de futuros problemas de divulgación y / o desafíos para la auditoría externa. Esta pregunta de investigación sería difícil sin tal método de clasificación de temas.

Específicamente, examinamos tres resultados específicos del tema que son indicativos de futuros problemas de divulgación y / o desafíos para la auditoría externa. Primero, la literatura previa encuentra que las cartas de comentarios son más probables cuando las cuentas son más riesgosas y complejas (por ejemplo, Ahn et al. 2020; Cassell et al. 2013). Por lo tanto, la forma en que se cubre un tema de contabilidad específico puede estar asociada con la probabilidad de recibir una carta de comentarios sobre ese tema. En segundo lugar, examinamos si los auditores externos divulgan posteriormente asuntos críticos de auditoría (CAM) sobre el mismo tema. Las CAM representan la visión del auditor de las áreas materiales y riesgosas de la auditoría. Además, la literatura reciente ha encontrado que, al cumplir con los nuevos requisitos de divulgación,36 los auditores de las empresas públicas evitan proporcionar nueva información al garantizar que sus divulgaciones se alineen con las divulgaciones de la gerencia (Burke et al. 2023; Griffith et al. 2022). Por lo tanto, la forma en que se cubre un tema contable específico puede predecir el tema de CAM (s) divulgado en el año siguiente. En tercer lugar, examinamos las re-expresiones posteriores relacionadas. Brown et al. (2020) encuentran que ciertos temas cubiertos en el 10-K están asociados con errores agregados. Ampliamos esto identificando el tema de la declaración errónea para permitir una identificación más sólida. Las declaraciones erróneas representan errores financieros y, por lo tanto, comprender cómo se cubren temas contables específicos puede predecir estas revelaciones de períodos futuros es importante para muchas partes interesadas.

Conclusión

A medida que los datos y análisis textuales se vuelven cada vez más importantes para la investigación contable (Bochkay et al. 2023), las innovaciones recientes en modelos de lenguaje grandes y el aprendizaje supervisado pueden proporcionar una identificación más sólida y validez interna para diversas tareas de clasificación de textos. Estos métodos de aprendizaje automático se han utilizado hasta ahora para reconocer el sentimiento en grandes volúmenes de divulgaciones financieras no etiquetadas (Frankel et al. 2022; Huang et al. 2023; Siano y Wysocki 2021). En este estudio, ajustamos FinBERT para reconocer temas contables dentro de las divulgaciones.

Una gran cantidad de literatura encuentra que los temas específicos de contabilidad están significativamente asociados con determinantes y / o resultados en las mismas áreas temáticas. Estos estudios utilizan datos clasificados por agregadores de datos, datos clasificados manualmente o un método de aprendizaje automático no supervisado (LDA) para clasificar los temas. Demostramos que nuestro modelo puede clasificar de manera más eficiente, completa y precisa los temas contables en grandes volúmenes de divulgaciones y, lo que es más importante, supera a la LDA comúnmente utilizada. Después de demostrar este desempeño, proporcionamos tres ejemplos de divulgaciones que se pueden clasificar en temas contables con nuestro modelo. Específicamente, encontramos que las notas etiquetadas a medida para los estados financieros, la sección MD&A del 10-K y la sección de factores de riesgo del 10-K contienen información significativa específica del tema que es predictiva de resultados contables específicos. Cada configuración es muy importante, ya que el 62 por ciento de las observaciones del año de la empresa utilizan al menos una etiqueta XBRL personalizada, y todas las empresas divulgan la sección MD&A y factor de riesgo.

Al igual que otros estudios que proponen un nuevo modelo, nuestro estudio está sujeto a varias limitaciones. Para ajustar un modelo de lenguaje grande para realizar una nueva tarea posterior, los investigadores necesitan algunos conocimientos de programación, así como recursos informáticos significativos. Sin embargo, una vez que el modelo se ajusta como el modelo de Huang et al. (2023) es para el sentimiento y nuestro modelo es para temas de contabilidad, estos recursos son menos limitantes. Además, inherente a un enfoque de aprendizaje automático supervisado es el uso de datos etiquetados para reconocer las mismas etiquetas en nuevos datos textuales. Por lo tanto, nuestro enfoque no identificará temas nuevos u ocultos a menos que esté capacitado para hacerlo. A pesar de estas limitaciones, creemos que nuestro modelo puede ayudar a los investigadores y profesionales interesados en identificar información relevante y consistente sobre temas contables a partir de grandes volúmenes de datos textuales.





