ANÁLISIS DE LITERATURA CIENTÍFICA DE MARKETING CON INTELIGENCIA ARTIFICIAL

Resumen

Presentamos una metodología de lectura de artículos y documentos científicos especializada para la disciplina académica de marketing basada en inteligencia artificial, adoptando los avances de lectura automatizada realizados en el campo científico de materiales. Describimos como se puede utilizar machine learning para la extracción, clasificación y etiquetado de términos y palabras del área de marketing. Eventualmente esta metodología en curso se podría utilizar para la formulación automatizada de hipótesis en marketing y para la transferencia de conocimiento científico a la práctica.

Palabras clave: inteligencia artificial; machine learning, ciencia del marketing

INTRODUCCIÓN

La aceleración de descubrimientos científicos es esencial para la prosperidad de la humanidad. Los avances científicos se difunden a través de publicaciones de investigación que están creciendo exponencialmente en número. Cada año se publican más de un millón de artículos. En campos de investigación como la educación, los materiales o el marketing, se publican miles. Los investigadores se han vuelto más eficaces en la generación de información que en su análisis, integración, interpretación y utilización para abordar cuestiones en las que se requiere conocimiento científico. La extracción manual de la información de investigación requiere de un esfuerzo humano sustancial. Como consecuencia de esta realidad, la sobrecarga y el crecimiento exponencial genera un cuello de botella en el avance de la ciencia. Mantenerse al día con las publicaciones es insostenible para los investigadores, incluso dentro de sus propios campos de especialización. Cada vez es más difícil asimilar y relacionar el conocimiento generado en la investigación, incluso dilucidar el rol de un fragmento de conocimiento. Crucialmente, los futuros descubrimientos en gran medida dependen del conocimiento publicado en investigaciones anteriores. Reconocer nuevas hipótesis podría ser cada vez más difícil, sin embargo, tan sólo una pequeña parte del conocimiento existente se utiliza para la formulación de nuevas hipótesis de investigación (Tshitoyan et al., 2019).

La extracción manual de la información es costosa. Aunque en la actualidad ya hay tecnologías disponibles con las que buscar y encontrar artículos relevantes, todavía no extraen ni organizan el contenido de estos documentos de manera automatizada. Tampoco todavía se formulan informáticamente nuevas hipótesis científicas a formular a partir de contenidos estructurados. El rápido aumento de publicaciones genera la necesidad de técnicas que puedan simplificar la utilización del conocimiento científico. Es aquí donde la inteligencia artificial (IA) puede ayudar.

FUNDAMENTOS TEÓRICOS

Literatura científica, minería de textos y procesamiento del lenguaje natural

Los documentos de investigación científica se presentan comúnmente en forma de artículos científicos y documentos de tesis publicados y disponibles a través de plataformas digitales. Estos artículos especializados, con terminologías propias de su dominio científico, están escritos con lenguaje natural en combinación con datos numéricos. Son difíciles de analizar

con análisis estadístico tradicional y también con aprendizaje automatizado moderno como el *machine learning (ML).*

Procesamiento natural de lenguaje (PNL) es un subcampo de la lingüística, informática, ingeniería de la información y de la inteligencia artificial, que se ocupa de las interacciones entre las computadoras y el lenguaje natural de los humanos, concretamente, cómo se programan las computadoras para procesar y analizar grandes cantidades de lenguaje natural.

La *minería de textos* se refiere al uso de algoritmos para la extracción de información de texto escrito. Asimismo, se utiliza en la literatura científica para identificar automáticamente referencias directas e indirectas a conocimientos específicos (ej. Srinivasan, 2004).

Minería de textos de literatura científica combinada con PNL

Varios estudios se han centrado en la minería de cadenas de texto y en la extracción de información contenida en literatura científica utilizando PNL supervisado. El conocimiento extraído puede ser eficientemente codificado como *word embeddings*, es decir, incrustaciones de palabras densas en información, esto es, representaciones vectoriales o mapas de palabras. Esta codificación se puede realizar sin etiquetado y sin supervisión humana (Tshitoyan et al., 2019).

Estado del arte de la lectura automatizada en investigación científica

Tshitoyan et al. (2019) ya han dado pasos para tratar de lograr "una orientación generalizada de síntesis de literatura científica". Sus desarrollos se han centrado en el campo de los materiales, donde, en sus estudios, han realizado minería de literatura científica utilizando PNL supervisado. Estos autores sugieren que con su método no supervisado incluso se pueden recomendar nuevos materiales varios años antes de su descubrimiento.

Varios estudios previos ya han utilizado PNL supervisado (ej. Swain & Cole, 2016). En ellos fueron necesarios grandes bases de datos etiquetados de forma manual para el posterior entrenamiento usando machine learning. Por contra, Tshitoyan et al. afirmaron que, con su metodología, se puede codificar la literatura científica de manera eficiente sin etiquetado o sin supervisión humana usando word embeddings densas en información. Para ello, se usan algoritmos ML tales como Word2vec. Este algoritmo predice los contextos de palabras, y crea un espacio semántico, es decir, una representación matemática de un cuerpo de texto grande.

En esta línea, Kim et al. (2017) desarrollaron con PNL una metodología de aprendizaje estadística para la recopilación automatizada de parámetros de síntesis de materiales. Para ello utilizaron decenas de miles de artículos de investigación, tanto en formato PDF como en texto sin formato. En su método, eliminaron artículos no relevantes entrenando con ML un clasificador binario que etiquetaba el abstract de un artículo como "relevante" o "no relevante".

Inteligencia artificial en el campo académico de la ciencia del marketing

El ámbito académico de marketing se podría beneficiar de la utilización de IA como ocurre en otros campos de investigación. También se podrían incorporar técnicas IA de escritura predictiva. Sin embargo, a día de hoy todavía hay una gran falta de conocimiento de cómo se podrían utilizar las tecnologías IA en esta disciplina, tanto en el presente como en el futuro, y de manera eficaz.

La mayoría de los adelantos en la lectura de literatura de investigación sobre IA se han realizado en campos como la biología, la química y los materiales, pero también se han hecho progresos, por ejemplo, en defensa y educación. El campo académico de la ciencia del marketing se favorecería de la adopción de tecnologías basadas en IA para la minería de textos, su análisis, y la escritura predictiva automatizada.

En esta dirección, Mustak et al. (2020) recientemente sugirieron posibles líneas de investigación de IA en marketing. La ciencia del marketing se ocupa de campos de investigación como las marcas, la publicidad o el comportamiento del consumidor. Esta disciplina fue tradicionalmente trabajada por economistas, pero durante los últimos tiempos ha tenido que abrir los horizontes y adoptar tecnología, mucha de la cual proviene de la ingeniería. En consecuencia, se crearon nuevas disciplinas como el marketing online. De hecho, debido a esta interdisciplinariedad en constante evolución, la literatura se encuentra cada vez más fragmentada y es difícil para los académicos de marketing seguir los ritmos de publicación.

La aceleración de la ciencia con IA también podría facilitar la transferencia del conocimiento científico a la práctica. Reputados académicos del marketing han señalado durante mucho tiempo la necesidad de conectar la ciencia con la industria (ej., Lilien, 2011). En particular, empresas con proyectos novedosos como startups podrían beneficiarse de la adopción de modelos científicos testados para ayudarles a tomar decisiones basadas en evidencias, reducir

los esfuerzos derivados por el ensayo y error, así como obtener ventajas competitivas (ej. Lee & Kozar, 2009). De hecho, Isidro Laso (2020) directivo de I+D en la Comisión Europea (CE) afirmó que "las startups representan una parte sustancial de la economía de la UE, la creación de empleo y el bienestar futuro debido a su naturaleza enérgica". El último programa de fomento de I+D de la CE se llama Horizon Europe. En él se priorizan tanto el uso de la ciencia, como el emprendimiento hasta el año 2030, destacando la necesidad de toma de decisiones informadas gracias a la disponibilidad rápida y de fácil conocimiento, sintetizada para su utilización práctica. Muchas startups están basadas en tecnología y cuentan con una fuerza laboral especializada, sin embargo, el creciente conocimiento científico disponible en los artículos todavía está infrautilizado. Aquí es donde la IA podría ayudar tanto a sintetizar el conocimiento como hacerlo más manejable por parte de empresas.

Campos científicos que avanzan con la lectura automatizada utilizando IA

La sobrecarga de investigación ya sucede en muchos campos académicos. Por ejemplo, en la investigación sobre educación, el tiempo entre un hallazgo, su publicación y su citación en una revisión de la literatura varía entre 2,5 y 6,5 años. Como respuesta, Crues (2017) propuso el desarrollo de un proceso "vivo y sistemático de revisión de la literatura" donde los hallazgos más recientes y publicados se agregan de manera automatizada para que tanto investigadores como profesionales puedan están mejor informados. En el campo de defensa, DARPA (2018) está desarrollando máquinas que entienden y razonan en contexto, con el objetivo de que, incrementalmente, las máquinas se conviertan en socias de los profesionales. Para que esto suceda, DARPA destaca que expertos en IA y expertos en dominios concretos deben trabajar juntos.

Gran parte de los recientes progresos en la lectura científica automatizada se ha hecho en el campo de los materiales, donde los estudios se han centrado en la asimilación de textos e información utilizando PNL supervisado. En este contexto, Tshitoyan et al. (2019) desarrollaron un método no supervisado capaz de extraer conocimiento y generar relaciones entre conceptos presentes en cuerpos masivos de literatura científica utilizando CrossRef API. Este interfaz de programación de aplicaciones (API) usa para obtener grandes listas de identificadores los objetos digitales (DOI) de los artículos de investigación. Editoriales como Elsevier (https://dev.elsevier.com) y Springer

Nature (https://dev.springernature.com) utilizan esta API para la descarga de artículos de revistas de texto completo, utilizando el servicio clic-through que proporciona CrossRef.

El trabajo de Tshitoyan et al. 2019. se llevó a cabo en el campo de la ciencia de los materiales. Con ML, entrenaron embeddings con 3.3 millones de abstracts científicos de los años 1922 a 2018 de más de 1000 journals principalmente de los campos de ciencia de materiales y física, así como artículos de journals de química que probablemente contengan investigación relacionada con materiales. Los artículos fueron obtenidos de las bases de datos científicas de Elsevier y Science Direct conjuntamente con web scraping. El rendimiento del algoritmo mejoró sustancialmente cuando se eliminaron resúmenes irrelevantes y en otros idiomas. Los 1,5 millones de abstracts restantes fueron clasificados como relevantes y etiquetados utilizando ChemDataExtractor (Swain & Cole, 2016) para producir las palabras individuales, lo que dio como resultado un vocabulario de aproximadamente 500.000 palabras.

Otro trabajo destacado en el campo de materiales es el de Kim et al. (2017) quienes desarrollaron un enfoque de aprendizaje estadístico específico para materiales sintetizados. El trabajo estaba centrado en como extraer datos de síntesis de materiales. Usando PNL compilaron parámetros de síntesis de manera automatizada de decenas de miles de artículos de investigación, tanto en formato PDF como en texto. Los artículos fuera del área de la literatura relevante fueron eliminados con un clasificador binario que etiquetaba los abstracts como “relevantes” o “no relevantes”. También Leaman, Wei, & Lu (2015) desarrollaron un reconocedor de términos químicos creado con dos modelos ML bautizados como "tmChem system".

Extracción de conocimiento a partir de bases de datos

Las principales fuentes de datos interpretables por máquinas provienen bases de datos estructuradas. Kim et al. (2017) presentaron su plataforma automatizada aprovechando la gran cantidad de fórmulas de síntesis publicadas a través de PNL, y utilizaron estas fórmulas para entrenar sus modelos ML. Su idea clave fue que, debido a que palabras con significados parecidos aparecen a menudo en contextos similares, las correspondientes embeddings también serán similares. Su plataforma lee artículos de forma automatizada y luego extrae y codifica

las condiciones de síntesis de materiales, conjuntamente con parámetros encontrados en el texto. Esto permite desarrollar nuevo conocimiento basado en parámetros fundamentales que son relevantes para impulsar la síntesis y creación de materiales específicos, tecnológicamente aplicables, y con un alto nivel de automatización. Al combinar los parámetros extraídos de texto a gran escala, esta base de datos de síntesis se utiliza para descubrir relaciones subyacentes. También Eltyeb & Salim (2014) desarrollaron un método de extracción de moléculas y sus propiedades, con el fin de minar textos, extraer datos y determinar relaciones contextuales útiles para investigadores. Para ello utilizaron técnicas basadas en diccionarios, en reglas, y también en ML. Utilizaron sistemas de reconocimiento de químicos híbridos que tratan con la búsqueda y clasificación de menciones de información específica dentro de documentos de texto. En su trabajo concretaron que los algoritmos más habituales en la lectura automática de artículos científicos son los algoritmos de aprendizaje supervisados; los no supervisados que utilizan representaciones a partir de datos; y los semi-supervisados que utilizan tanto datos etiquetados como no etiquetados. Eltyeb & Salim concluyeron que los sistemas basados en diccionarios son más adecuados y eficientes cuando se utilizan palabras del vocabulario bien definidas y actualizadas así cuando se escriben las palabras correctamente en los documentos.

Generación de hipótesis con IA

La generación automatizada de hipótesis también es un campo incipiente y en desarrollo. En el trabajo de Spangler et al. (2014) se fijó como objetivo combinar minería de textos, visualización y analítica, con la idea de integrar todo contenido disponible, identificar evidencias que son relevantes para una consulta determinada y, a partir de estas evidencias, sugerir hipótesis que son nuevas, interesantes, que se puedan contrastar y probablemente se confirmen. Para que esto suceda, se requieren tanto datos suficientes como expertos en dominios complejos para poder acelerar el progreso científico y lograr descubrimientos relevantes.

El proceso de Spangler et al. tiene tres fases: 1. Exploración, donde se examina la información no estructurada relevante, se diseñan consultas de texto y se extrae información relevante a utilizar; 2. Interpretación, donde se crea un gráfico de relaciones de similitudes para ayudar a los expertos de dominio a visualizar conexiones ocultas; 3. Análisis, esto es, la clasificación de conceptos que se ordenan según sean mejores candidatas para experimentar con ellas y realizar predicciones novedosas. El experto en el dominio elige los conceptos candidato cuya relación

sea más probable desde el punto de vista analítico, sean comprobables experimentalmente, y de relevante interés para el problema general a tratar.

Escritura predictiva. BERT y ELMo

La escritura automatizada de textos es un campo ya avanzado, aunque su aplicación en la investigación todavía está en desarrollo. BERT y ELMo son las dos técnicas más avanzadas. BERT (Devlin, Chang, Lee & Toutanova, 2018) es un codificador bidireccional de texto, que trabaja tanto para adelante como para detrás, y que permite el aprendizaje automático en el contexto de una palabra o término en función de todo su entorno, es decir, a la izquierda y a la derecha de la palabra. Puede ser utilizado para generar modelos de lenguaje. Su innovación técnica clave es la aplicación bidireccional del popular modelo de atención "Transformer" (https://talktotransformer.com) al modelado de lenguaje. Este sistema aprende las relaciones contextuales entre palabras o subpalabras en un texto usando un codificador que lee la entrada de texto y un decodificador que produce una predicción. BERT puede ser afinado con pequeñas cantidades de datos, aunque se logra una precisión más elevada cuanto mayores sean los datos de entrenamiento. ELMo (Peters el al., 2018) se utiliza para producir embeddings de palabras contextuales, ya que una palabra puede tener un significado diferente según las palabras que la rodean.

OBJETIVOS DE INVESTIGACIÓN

El objetivo de este trabajo es presentar una metodología de lectura y análisis automatizada de textos científicos basada en machine learning específica para el campo académico de marketing. Para ello nos basamos en las metodologías desarrolladas por Tshitoyan et al. (2019) y Kim et al. (2017) en el campo de investigación científica de materiales.

METODOLOGÍA

El primer paso de esta metodología en curso consiste en recopilar cuerpos consolidados de literatura científica de marketing sobre temas concretos. Para ello hemos elegido como punto de partida los tres temas de marketing offline branding, retail y publicidad, siguiendo el método de Tshitoyan et al. (2019) de extracción de conocimiento y relaciones para el manejo de grandes volúmenes de literatura científica. Esto se realiza con CrossRef Application Programming Interface (API) con el fin de obtener grandes listas de Digital Object Identifiers (DOI) de artículos. Este sistema es utilizado por las APIs de editoriales como Elsevier (https://dev.elsevier.com) y Springer Nature (https://dev.springernature.com) para la descarga

de artículos de revistas de texto completo utilizando el servicio clic-through proporcionado por CrossRef. A continuación, se codifican los textos como word embeddings, o mapas vectoriales de palabras, densas en información, utilizando etiquetado humano y entrenamiento machine learning Word2vec.

En segundo lugar, se entrenarán las embeddings con los abstracts de cada uno de los temas offline elegidos, esto es, branding, retail o publicidad. Para ello, utilizamos abstracts de artículos de 1975 a 2021 de más de mil journals científicos y también de artículos que probablemente contengan investigación relacionada con el marketing, obtenidos directamente de las bases de datos científicas anteriormente mencionadas, en combinación con web scraping. Prevemos que el rendimiento del algoritmo mejorará cuando se eliminen abstracts irrelevantes. Los abstracts restantes serán clasificados como relevantes y serán etiquetados mediante ChemDataExtractor (Swain & Cole, 2016), lo que debería dar como resultado un vocabulario de miles de palabras individuales. Esperamos que los resultados mejorarán con un pre-procesamiento correcto, especialmente con la selección de frases, que serán incluidas como palabras individuales.

En el tercer y último paso, repetiremos los dos primeros integrando los temas de marketing offline seleccionados anteriormente, con los tres equivalentes online, es decir, branding online, retail online y publicidad online. Las lecciones aprendidas durante el tratamiento de los temas de marketing offline nos pueden orientar sobre cómo tratar mejor los análisis realizados con los equivalentes online. En este punto, evaluaremos qué ocurre al combinar branding offline con el online, retail offline y online así como la publicidad offline y online. La integración de las seis áreas combinadas nos podría informar sobre fenómenos cross-channel potenciales.

RESULTADOS Y DISCUSIÓN

El objetivo de este trabajo en curso es dilucidar si la inteligencia artificial será capaz de analizar la literatura científica de marketing haciendo uso de tecnologías de IA basado en los progresos realizados en otros campos científicos. Durante la revisión de la literatura, se encontró que avances como la *minería de textos, la síntesis de conocimientos* y la *formulación de hipótesis novedosas* ya están ocurriendo en ciertas disciplinas científicas. Bien es sabido, como es la posición de DARPA, que en IA los avances se producen cuando hay colaboración entre expertos en IA y expertos en el dominio a tratar, como sería el campo científico de marketing.

Implicando a startups en el proceso, se debería mejorar la síntesis de investigación automatizada para la transferencia de conocimiento de marketing basado en evidencias y el empoderamiento de las mismas. La necesidad de conectar la investigación con la práctica ha sido señalada por los mejores académicos de marketing (ej. Lilien, 2011), pero todavía no ha habido avances implicando IA. Además, es bien sabido que, en IA, la mayoría de los avances se han obtenido con una gran e inevitable frustración. El experimentado investigador de IA Eric Saund (2020) afirmó que, debido a esta situación habitual, tener objetivos claros, paciencia, tiempo y recursos son fundamentales para poder avanzar en la consolidación de metodologías como la que hemos expuesto, sabiendo que nos encontraremos con obstáculos por el camino. Por ello, tendremos siempre presentes los siguiente objetivos generales de nuestro trabajo en curso:

- Resolver cuellos de botella científicos con IA por la sobrecarga en la investigación y el previsible crecimiento exponencial.
- Integrar e interpretar conocimiento en la disciplina de investigación en marketing.
- Ser conscientes que los descubrimientos futuros dependen en gran medida del conocimiento documentado en publicaciones anteriores.
- La aceleración del descubrimiento científico debe estar cuantificada.
- Mantener localizadas bases de datos estructuradas identificadas por el camino, ya que eventualmente serán necesarias para entrenamiento con ML.
- Co-crear los desarrollos con personal de marketing de startups.
- Poner en marcha una revisión sistemática de la literatura de nuevos artículos de revistas.

CONCLUSIONES

Esperamos que nuestro método en curso sea eventualmente capaz de analizar literatura científica de marketing utilizando machine learning. Iremos incorporando desarrollos en inteligencia artificial pertinentes que se realicen en los campos más avanzados como materiales según vayamos haciendo progreso. Si los procesos descritos se hacen correctamente, los avances podrían facilitar el desarrollo de la ciencia del marketing así como la transferencia de conocimiento a la práctica especialmente a startups, muchas de las cuales operan en tecnología, son inicialmente pequeñas, generalmente tienen recursos limitados, y requieren de conocimientos adecuados en los momentos necesarios con el fin de reducir sus errores y riesgos.

En sucesivos pasos, podríamos incorporar investigación de marketing sobre producto, precios, promociones y distribución, tanto en versiones offline como online. Asimismo, las lecciones aprendidas durante los tres pasos de la metodología expuesta, podrían repetirse al incorporar otras áreas especializadas como el marketing turístico online, el marketing educativo online o el marketing internacional, que se sumarían a los campos anteriormente mencionados. Asimismo, como IA también es capaz de escribir textos utilizando escritura predictiva con los codificadores bidireccionales BERT y ELMo (Devlin et al., 2018; Peters et al., 2018) sería posible la formulación automatizada de hipótesis en contextos de marketing como está ocurriendo en ciencia de materiales (Spangler et al., 2014).

Nuestro método propuesto no está libre de limitaciones. Aunque la metodología expuesta está avanzando en áreas de las ciencias básicas y en las tecnologías de la información, en la etapa temprana en la que estamos todavía no sabemos si va a funcionar bien, incluso valorado, en la investigación en ciencias sociales. Asimismo, si bien en IA la focalización en un campo de investigación concreto puede conducir a avances, la especialización también limita inherentemente las oportunidades de encontrar conexiones entre campos. Esto podría ocurrir en áreas tales como el branding online.

BIBLIOGRAFÍA

1. Crues, W. (2017). Automated Extraction of Results from Full Text Journal Articles in Proceedings of the 10th *International Conference on Educational Data Mining.*

2. DARPA, Defence Advanced Research Projects Agency (2018). "AI Next Campaign". https://www.darpa.mil/work-with-us/ai-next-campaign

3. Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*

4. Eltyeb, S., & Salim, N. (2014). Chemical named entities recognition: a review on approaches and applications. *Journal of cheminformatics*, *6*(1), 1-12.

5. Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., & Olivetti, E. (2017). Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, *29* (21), 9436-9444.

6. Laso, I. (2020). European Commission Directorate-General for Research and Development. EUvsVirus livestream. https://www.facebook.com/EUvsVirus/videos/2635593133382903

7. Leaman, R., Wei, C.H., & Lu, Z. (2015). tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, *7*(1), 1-10.

8. Lee, Y., & Kozar, K. A. (2009). Designing usable online stores: A landscape preference perspective. *Information & Management*, *46* (1), 31-41.

9. Lilien, G. L. (2011). Bridging the academic–practitioner divide in marketing decision models. *Journal of Marketing*, *75* (4), 196-210.

10. Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, *124*, 389-404.

11. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*

12. Saund, E. (2020). Algorithmic Bias and the Confusion Matrix. In Nag, R. SCI-52 Artificial Intelligence: Deep Learning, Human-Centered AI, and Beyond. Stanford University.

13. Spangler S., Wilkins A.D., Bachman B.J., Nagarajan M., Dayaram T., Haas P., Regenbogen S., Pickering C.R., Comer A., Myers J.N., Stanoi I., Kato, L., Lelescu, A, Labrie, J.J., Parikh N., Lisewski, A.M., Donehower, L., Chen, Y., & Lichtarge, O. (2014). Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining,* 877-1886.

14. Srinivasan, P. (2004). Text mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, *55* (5), 396-413.

15. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G. & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571* (7763), 95-98.