Desarrollo de un Modelo de Aprendizaje Automático para la Asignación de Códigos de Producto por Sociedades Comisionistas de la Bolsa Mercantil de Colombia a partir de Descripciones de Productos en Supermercados

S. Collantes Zuluaga[1], J. P. García Cifuentes[2], J. Gil González[3]

8 de julio de 2024

[1]*Pontificia Universidad Javeriana Cali, Colombia*
[2]*Pontificia Universidad Javeriana Cali, Colombia*
[3]*Pontificia Universidad Javeriana Cali, Colombia*

Resumen

Este proyecto presenta el desarrollo de un modelo de aprendizaje automático para la asignación automática de códigos de productos en la Bolsa Mercantil de Colombia (BMC) a partir de descripciones proporcionadas por sus Sociedades Comisionistas (SC). Utilizando técnicas avanzadas de procesamiento de lenguaje natural (NLP) y aprendizaje profundo, se busca mejorar la precisión y eficiencia del proceso actual de asignación manual.

1. Introducción

En la era actual de la digitalización y el avance tecnológico, la eficiencia en el procesamiento de grandes volúmenes de datos se ha convertido en una ventaja competitiva crucial para diversas industrias. El sector financiero, en particular, enfrenta el desafío de manejar millones de transacciones diarias con precisión y rapidez. La Bolsa Mercantil de Colombia (BMC), una entidad vital en el sistema financiero colombiano, no es ajena a estos retos. Un proceso crítico dentro de las operaciones de la BMC es la asignación de códigos de productos estandarizados basados en las descripciones proporcionadas por supermercados. Esta tarea, actualmente realizada de forma manual por las Sociedades Comisionistas (SC), es laboriosa y propensa a errores, consumiendo una cantidad significativa de tiempo y recursos humanos. Se estima que la conversión de un solo producto del código interno al código del BMC requiere entre 2 y 3 minutos, lo que se traduce en aproximadamente 450 horas de trabajo para una empresa que maneja 9000 productos. La automatización de este proceso mediante técnicas de inteligencia artificial, específicamente el procesamiento de lenguaje natural (NLP) y el aprendizaje automático, se presenta como una solución prometedora para optimizar esta operación crítica. Este estudio se centra en el desarrollo de un modelo de aprendizaje automático capaz de asignar automáticamente códigos de productos a partir de sus descripciones, utilizando técnicas avanzadas de NLP.

1.1. Objetivos

El objetivo general de esta investigación es desarrollar un modelo de aprendizaje automático que utilice técnicas de procesamiento de lenguaje natural para la asignación eficiente y precisa de códigos de productos en la Bolsa Mercantil de Colombia. Los objetivos específicos incluyen:

1. Preparar un conjunto de datos que contenga descripciones de productos de supermercados y sus

correspondientes códigos de la BMC.

2. Implementar dos modelos de procesamiento de lenguaje natural, basados en técnicas de aprendizaje de máquina, para mapear automáticamente las descripciones de productos a los códigos de la BMC y comparar sus resultados.

3. Evaluar los modelos desarrollados utilizando métricas de rendimiento apropiadas, como top-k accuracy, para garantizar su eficacia y precisión en la asignación de códigos.

1.2. Relevancia

Este estudio no solo busca mejorar la eficiencia operativa y reducir los costos para las Sociedades Comisionistas, sino que también aspira a sentar un precedente sobre cómo las tecnologías emergentes pueden ser aplicadas en el sector financiero colombiano, potenciando su competitividad y adaptabilidad en un mundo cada vez más digitalizado.

2. Fundamentación Teórica

El Procesamiento de Lenguaje Natural (NLP) es una disciplina que se centra en la interacción entre las máquinas y el lenguaje humano, permitiendo a los sistemas inteligentes entender, interpretar y generar lenguaje de manera efectiva. Desde sus inicios en la década de 1950, el NLP ha experimentado una evolución significativa, impulsada por avances en técnicas y tecnologías, mejorando considerablemente en términos de eficacia y aplicabilidad [1].

En el contexto de este estudio, dos técnicas de NLP son particularmente relevantes: los *word embeddings* y los *transformers*. Los *word embeddings* son representaciones vectoriales de palabras que capturan relaciones semánticas entre ellas. Técnicas como Word2Vec, GloVe y FastText han demostrado ser eficaces en la transformación de texto en un espacio vectorial, facilitando diversas tareas de NLP [2].

Por otro lado, los *transformers*, introducidos por Vaswani et al. [3], han revolucionado el campo del NLP. Estos modelos, que se basan en mecanismos de atención auto-regresiva, han establecido nuevos estándares en múltiples tareas de NLP, desde traducción hasta generación de texto. Modelos como BERT (Bidirectional Encoder Representations from Transformers) han demostrado ser particularmente efectivos en tareas de clasificación y similitud de texto [4].

Sin embargo, uno de los desafíos con BERT es la sobrecarga computacional al comparar similitud entre oraciones. Para abordar este problema, Reimers y Gurevych [5] introdujeron Sentence-BERT (SBERT), una modificación de BERT que utiliza estructuras de redes siamesas para generar embeddings de oraciones que capturan su significado semántico. SBERT permite comparar eficientemente oraciones utilizando similitud coseno, reduciendo significativamente el tiempo y la sobrecarga computacional.

La similitud de oraciones, una aplicación crucial de estas técnicas para nuestro estudio, se mide comúnmente utilizando la similitud coseno. Esta métrica, definida como el producto punto normalizado de dos vectores, proporciona un valor entre -1 y 1, donde 1 indica vectores idénticos y -1 indica vectores diametralmente opuestos [6].

Para evaluar el rendimiento de los modelos en tareas de clasificación multiclase (como en nuestro caso, con la asignación de códigos de productos) se utiliza comúnmente la métrica de precisión top-k (*top-k accuracy*). Esta métrica evalúa la capacidad del modelo para incluir la respuesta correcta entre sus k predicciones más probables [7].

Estudios recientes han demostrado la eficacia de estas técnicas en diversos campos. Por ejemplo, Chakrabarty [8] demostró cómo los embeddings de oraciones basados en *transformers* pueden ahorrar tiempo y esfuerzo en el etiquetado de datos para clasificación de textos, logrando una precisión del 90 %. Asimismo, Liu et al. [9] desarrollaron "Ticket-BERT", un modelo BERT afinado específicamente para etiquetar incidentes en sistemas de gestión de tickets, logrando un rendimiento superior al de los clasificadores tradicionales.

Estas técnicas y estudios previos proporcionan una base sólida para nuestro enfoque de asignación automática de códigos de productos en la Bolsa Mercantil de Colombia, prometiendo mejorar significativamente la eficiencia y precisión del proceso actual.

3. Resultados

Se preparó el corpus de datos y se evaluaron diversos modelos y técnicas para la alineación automática de descripciones de productos, utilizando tanto datos reales como sintéticos. A continuación, se presentan los procedimientos de preparación, cálculo y los resultados correspondientes para cada enfoque.

3.1. Preparación y Enriquecimiento del Conjunto de Datos

Antes de aplicar los modelos, se realizó un extenso trabajo de preparación y enriquecimiento del conjunto de datos:

- **Limpieza y Estandarización** Se llevó a cabo una limpieza exhaustiva de los datos, que incluyó la eliminación de duplicados, la estandarización de formatos y la corrección de errores ortográficos. Este proceso redujo el ruido en los datos y mejoró la calidad general del conjunto.
- **Enriquecimiento Semántico** Se aplicaron técnicas de procesamiento de lenguaje natural para enriquecer semánticamente las descripciones. Esto incluyó la tokenización, lematización y la extracción de características lingüísticas clave, lo que mejoró la capacidad de los modelos para capturar el significado subyacente de las descripciones.

3.2. Modelo Word2Vec

Se entrenó el modelo utilizando tokens de las descripciones de productos y se calculó la similitud del coseno entre los vectores resultantes.

Métrica	**Top-1**	**Top-3**
Precisión de alineación correcta	0.30 %	0.82 %

Cuadro 1: Resultados del modelo Word2Vec

3.3. Modelo Preentrenado spaCy

Se generaron embeddings de las descripciones utilizando spaCy y se utilizó el modelo preentrenado `es_core_news_md` de spaCy para generar embeddings de las descripciones, posterior a esto se calculó la similitud del coseno entre ellos.

Métrica	**Top-1**	**Top-3**
Precisión de alineación correcta	2.60 %	4.64 %

Cuadro 2: Resultados del modelo preentrenado spaCy

3.4. Combinación de Similitud Coseno y Jaccard con spaCy

Se aplicaron tanto la similitud del coseno como la similitud de Jaccard a los embeddings generados por spaCy oara observar como diferentes metricas de similitud podían afectar la correcta alineación de las descripciones.

Métrica	**Top-1**	**Top-3**
Precisión de alineación correcta	11.71 %	27.41 %

Cuadro 3: Resultados de la combinación de similitud coseno y Jaccard con spaCy

3.5. Modelo SBERT (Sentence-BERT) Fine-tuned

Se realizó fine-tuning del modelo SBERT con 3 épocas y ejemplos negativos, calculando posteriormente la similitud del coseno entre los embeddings generados.

Tipo de Descripción	**Top-1**	**Top-3**
Completas	34.84 %	59.52 %
Subyacentes	47.14 %	69.71 %

Cuadro 4: Resultados del modelo SBERT fine-tuned

3.6. Evaluación con Datos Sintéticos

Se generaron 1750 descripciones sintéticas mediante back-translation y modelos generativos para una evaluación más exhaustiva.

Modelo	Top-1	Top-3
spaCy con Jaccard + Coseno	16.46 %	25.49 %
SBERT fine-tuned	44.63 %	70.57 %

Cuadro 5: Resultados de la evaluación con datos sintéticos

3.7. Análisis Comparativo

El modelo SBERT fine-tuned demostró el mejor rendimiento en todas las pruebas, especialmente al utilizar descripciones subyacentes simplificadas. La combinación de similitud coseno y Jaccard mejoró consistentemente los resultados en comparación con el uso de una sola métrica. La evaluación con datos sintéticos confirmó la superioridad del modelo SBERT fine-tuned en términos de precisión y robustez para la tarea de alineación de descripciones de productos.

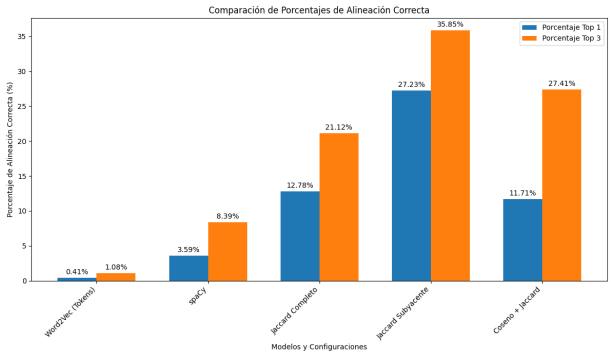

Figura 1: Comparación de Porcentajes de Alineación Correcta entre Word2Vec y spacy.

4. Discusión y Conclusiones

Este estudio presenta un avance significativo en la aplicación de técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático para la asignación automática de códigos de productos en el contexto de la Bolsa Mercantil de Colombia (BMC). Los resultados obtenidos demuestran el potencial de estas tecnologías para mejorar la eficiencia y precisión

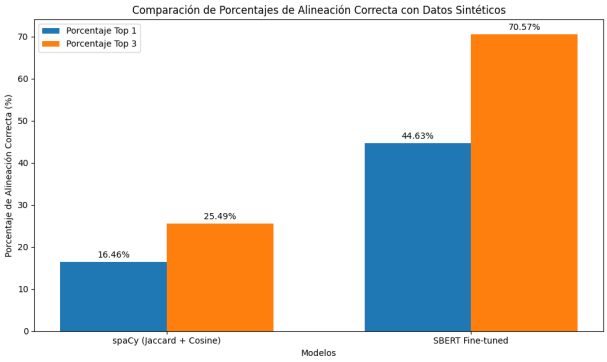

Figura 2: Comparación de Porcentajes de Alineación con datos sintéticos.

en procesos tradicionalmente manuales y propensos a errores.

La comparación de diferentes modelos y técnicas revela la superioridad de los enfoques basados en transformers, particularmente el modelo SBERT (Sentence-BERT) con fine-tuning. Este modelo demostró una precisión significativamente mayor en comparación con técnicas más tradicionales como Word2Vec o el uso de modelos preentrenados sin ajuste fino. La capacidad del modelo SBERT para capturar relaciones semánticas complejas en descripciones de productos técnicos subraya su potencial para aplicaciones en dominios especializados.

Un hallazgo particularmente relevante fue la mejora en la precisión al utilizar descripciones subyacentes simplificadas. Esto sugiere que la categorización más general de productos puede ser una estrategia efectiva para mejorar la asignación automática de códigos, especialmente en contextos donde las descripciones detalladas pueden variar significativamente entre diferentes fuentes.

La combinación de métricas de similitud, como la similitud coseno y Jaccard, demostró ser una estrategia efectiva para mejorar la precisión de la alineación. Este enfoque híbrido podría ser valioso en otros escenarios de NLP donde se requiere una comparación robusta de textos cortos o técnicos.

Sin embargo, es importante reconocer las limitaciones del estudio, particularmente en relación con la calidad y cantidad de datos disponibles. La variabili-

dad en las descripciones de productos y la presencia de errores ortográficos representan desafíos significativos que futuros estudios deberán abordar para mejorar aún más la precisión de los modelos.

Además, el enfoque desarrollado en este estudio tiene el potencial de ser adaptado y aplicado a otros sectores que enfrentan desafíos similares en la clasificación y categorización de productos o servicios basados en descripciones textuales. Esto podría incluir aplicaciones en comercio electrónico, gestión de inventarios, o incluso en sectores como la salud, donde la clasificación precisa de información textual es crucial.

En conclusión, este estudio no solo proporciona una solución práctica a un problema específico de la BMC, sino que también contribuye al cuerpo de conocimiento en la aplicación de técnicas de NLP en contextos financieros y técnicos. Los resultados obtenidos abren camino para futuras investigaciones en la optimización de procesos basados en texto en diversos sectores, destacando el potencial de las tecnologías de IA para transformar operaciones tradicionalmente manuales en procesos más eficientes y precisos.

Referencias

[1] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, "Natural Language Processing: History, Evolution, Application, and Future Work", in *Proceedings of 3rd International Conference on Computing Informatics and Networks*, pp. 365-375, 2021. DOI: 10.1007/978-981-15-9712-1_31.

[2] Turing, *Word embeddings in NLP: A Complete Guide* , Turing.com. [En línea]. Disponible: `https://www.turing.com/kb/guide-on-word-embeddings-in-nlp`. [Accedido: 26-Nov-2023].

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin,"Attention Is All You Need" en *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017. [Online]. Disponible: arXiv:1706.03762v7 [cs.CL]

[4] C. Wang, M. Qiu, C. Shi, T. Zhang, T. Liu, L. Li, J. Wang, M. Wang, J. Huang, y W. Lin, "EasyNLP: A Comprehensive and Easy-to-use Toolkit for Natural Language Processing", arXiv:2205.00258v2 [cs.CL], 13 Mar 2023

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", Ubiquitous Knowledge Processing Lab (UKP-TUDA), Department of Computer Science, Technische Universität Darmstadt, 2019. [Online]. Disponible: `https://github.com/UKPLab/sentence-transformers`

[6] A. Huang, "Similarity measures for text document clustering," *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC)*, Christchurch, New Zealand, 2008.

[7] scikit-learn, *sklearn.metrics.top_k_accuracy_score,* Available: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.top_k_accuracy_score.html`, Accessed on: May 20, 2024

[8] A. A. Chakrabarty, "Text Data Labelling Using Transformer Based Sentence Embeddings and Text Similarity for Text Classification", en International Journal on Natural Language Computing (IJNLC), vol. 11, no. 2, abril de 2022

[9] Z. Liu, C. Benge, S. Jiang "Ticket-BERT: Labeling Incident Management Tickets with Language Models", arXiv:2307.00108, 2023.