Clasificación de texto de clase múltiple con aprendizaje profundo utilizando BERT

Daniel Oswaldo Mora Malaver

Universidad Nacional de Colombia

Facultad de Ciencias, Departamento de Estadística.

Bogotá, Colombia

Lista de Contenidos

1. Introducción

En Colombia actualmente existen 5 empresas comercializadoras y productoras de cemento: Argos, Cemex, Holcim, Cementos del Oriente, y Cementos Tequendama. Estas empresas localizadas cerca de los mercados plaza más representativos del país (Bogotá, Antioquia, Valle del Cauca, y la Costa Atlántica). Por el poco número de empresas en el sector actualmente se considera un mercado concentrado, con un modelo de integración vertical, barreras a la entrada de jugadores, altas inversiones iniciales, y un producto poco diferenciado.

Una de las empresas líderes del negocio del cemento en Colombia es HOLCREST S.A.S. Tiene presencia en unos 70 países y emplea a unos 72.000 empleados. Holcim opera cuatro segmentos de negocios: cemento , agregados , concreto premezclado y otros productos, incluyendo concreto prefabricado, asfalto, mortero y otros materiales de construcción.

Con sede en Suiza y cotizando en SIX Swiss Exchange y en Euronext Paris , Holcim Grupo ocupa posiciones de liderazgo en todas las regiones del mundo. El mercado de materiales de construcción está impulsado por el crecimiento masivo de la población mundial , el cambio hacia la ciudad y la vida urbana y la infraestructura, las carreteras , los puentes , los hospitales y las escuelas , que requieren las poblaciones en crecimiento. Tiene presencia a nivel nacional a través de 1 planta de cemento, 10 plantas de concreto premezclado, 1 plataforma de Geocycle, 1 planta de agregados, su propia red de ferreterías, Disensa, con más de 400 tiendas a nivel nacional; y ofrece servicios especializados de transporte de materiales o productos a través de Transcem.

En la empresa HOLCREST S.A.S. existe una inmensa cantidad de información disponible en formato electrónico. Toda esta información es improductiva si no se dispone con mecanismos apropiados para su acceso, clasificación y análisis. En particular, la clasificación automática de textos consiste en clasificar un documento dentro de un grupo de clases en este caso previamente definidas. El objetivo principal en esta pasantía fue desarrollar un modelo de clasificación de texto que permita a la compañía HOLCREST S.A.S clasificar sus correos en unas categorías establecidas previamente.

Cada actividad realizada dentro de la empresa ha sido con el objetivo de aplicar y adquirir conocimientos, mejorar el área de trabajo asignado. La implementación de estas actividades permitió

automatizar tareas y ayudar a hacer mas eficiente la empresa.

A continuación se expondrá un informe del proceso, actividades desarrolladas, aprendizaje adquirido, desempeño del trabajo y resultados finales.

2. Objetivos

2.1. Objetivo General

Desarrollar varios modelos de clasificación múltiple de texto para clasificar correos en unas categorías establecidas previamente.

2.2. Objetivos Específicos

- Comprender las diferentes técnicas y estrategias de clasificación de texto existentes en la literatura.
- Aprender sobre inteligencia artificial, en específico sobre Procesamiento de Lenguaje Natural para poder desarrollar la tarea deseada.
- Entender cómo funcionan las redes neuronales Transformer.
- Fortalecer los conocimientos teóricos asimilados durante los años de carrera universitaria, adquiriendo al mismo tiempo habilidades y destrezas para un mejor desempeño laboral en condiciones reales.

3. Problema propuesto por la empresa

La empresa HOLCREST S.A.S. tiene un problema y es que cada correo que llega a su poder llega sin clasificar y la tarea de clasificar estos correos es llevado manualmente y es una tarea que quieren automatizar para poder clasificar miles de correos en categorías específicas rápidamente.

4. Reporte de las actividades realizadas

La primera actividad de la pasantía fue la revisión fue entender el negocio y conocer más sobre la empresa Holcim, esto me permitió comprender la división de la empresa, cuáles son los elementos por considerar y como aplicar esta información a un modelo de clasificación.

Después de entender bien como se estructura la empresa, procedí a entender los datos con los que iba a trabajar, los datos son correos de siete países distintos y cada correo está clasificado en una categoría ya definida. Los países de los que Holcim recibe correos son Ecuador, Argentina, El salvador, Nicaragua, Colombia, Costa rica y México. Las siete categorías son: Aplicación de pago y pedido, No cliente, Datos maestro, Pedido, Otros, Aplicación de pago, y Logística. México tiene solo seis categorías, la categoría "logística" no existe en este país, los otros seis países si presentan las siete categorías presentadas. Para cada país tenía la siguiente cantidad de correos:

País	Correos en total
ecuador	27487
México	12963
nicaragua	10963
el Salvador	23983
argentina	29826
Colombia	19355
costa rica	18948

Gráfica 1: Numero de correos por pais

Para el caso específico de Colombia, tenemos 19355 correos en total, los cuales están distribuidos en estas siete categorías.

Categorias	Numero de correos
Aplicación de pago y pedido	150
Aplicación de pagos	4958
Datos maestro	368
Logistica	367
No cliente	4683
Otros	6206
Pedidos	2623
Total general	**19355**

Gráfica 2: Distribución de correos en Colombia.

Clasificación	Breve descripción	Correo electrónico	Descripción	País
Pedidos	Re: NUEVO PEDIDO DE CEMENTO	andaluciaconcretos@gmail.com	Por motivo de que el lunes es festivo solicito de manera respetuosa que el material llegue a la planta mañana sábado 13 de noviembre del año 2021.muchas gracias El jue., 11 de noviembre de 2021 6:06 p.m., Regional Customer Service <inf>	Colombia
Pedidos	Cancelación de pedidos	materiaprima.concremovil@gmail.com	Buen dia, Por favor cancelar los siguientes pedidos programados para el dia de mañana 31 de diciembre: 501684647501684650501685108 Quedo atenta. -- Cordialmente, Dina Luzeth Ramirez VegaAuxiliar Contable - Concremovil S.A.S.Carrera 8 # 15-10 Lot	Colombia
Pedidos	Pedido	info.colombia@lafargeholcim.com	Buenos dias Por favor me colaboran con la siguiente modificación en los pedidos Muchas gracias Piedad Pérez Secretaria Sika Colombia S.A.S Calle 3 Carrera 3 Duitama - Colombia IPhone: +58 7638120/21/22 · Mobile: 3115788010 · Fax: +5	Colombia

Gráfica 3: Ejemplos de correos.

Revisión Bibliográfica.

La siguiente actividad de la pasantía fue la revisión bibliográfica de los modelos de clasificación múltiple de texto.

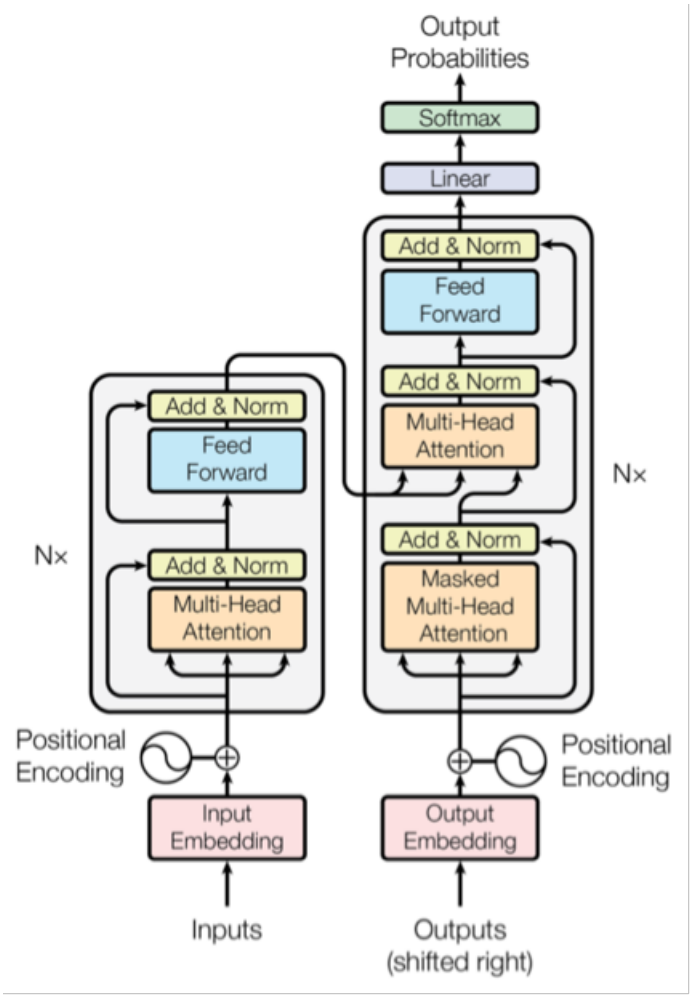
Clasificación automática de Textos

El objetivo de la clasificación automática de texto es categorizar documentos dentro de un número fijo de categorías predefinidas en función de su contenido. Un documento puede pertenecer a una de las categorías dadas. Cuando se utiliza aprendizaje automático, el objetivo es aprender a clasificar a partir de ejemplos que permitan hacer la asignación a la categoría automáticamente.

El modelo que se implementó para la clasificación de texto multiclase fue un modelo Bert (BERT es un Transformer codificador bidireccional). Antes de hablar del modelo Bert mas ampliamente es necesario definir un tipo de redes neuronales, las redes transformer.

Redes transformer

La arquitectura que utiliza Bert es la denominada arquitectura transformer, esta fue presentada por Google a finales del 2017, en una presentación conocida como "Attention is all you need", la aparición de esta arquitectura trajo consigo las denominadas como capas de autoatención. Estas redes son especiales para el tratamiento de procesamiento de lenguaje natural. Recientemente han sido aplicadas a series de tiempo e imágenes. Se basan en el concepto de auto-atención.



Gráfica 4: Los modelos Transformer

El Transformer recibe una oración de entrada y la convierte en dos secuencias: una secuencia de vectores de palabras y una secuencia de codificaciones posicionales. Ambos vectores son escritos usando representaciones numéricas del texto para que la red neuronal pueda procesarlas. Cada palabra del diccionario se representa como un vector. Las codificaciones posicionales son una representación vectorial de la posición de la palabra en la oración original.

El transformer junta ambas secuencias y pasa el resultado a través de una serie de codificadores, seguidos de una serie de decodificadores. A diferencia de las RNN, el input no es alimentado en la red de forma secuencial sino que se pasa todo de una vez.

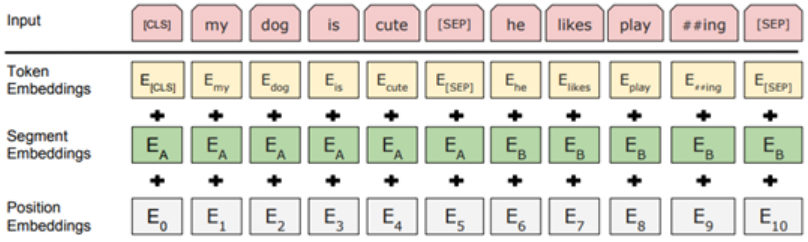
Cada uno de los codificadores convierte su entrada en otra secuencia de vectores llamados codificaciones. Los decodificadores hacen lo contrario: vuelven a convertir las codificaciones en una secuencia de probabilidades de diferentes palabras de salida. Cada codificador y decodificador contiene un componente llamado mecanismo de atención, que le permite generar un contexto a cada palabra que procesa, y cómo los mismos son colocados en forma paralela, la red tiene la capacidad de procesar todas las palabras en simultáneo, lo que le brinda su ventaja.

Finalmente las probabilidades numéricas en la capa de salida se pueden convertir en otra oración en lenguaje natural usando la función softmax.

Modelo Bert

BERT (Bidirectional Encoder Representations from Transformers) o Representación de Codificador Bidireccional de Transformadores es una técnica basada en redes neuronales para el preentrenamiento del procesamiento del lenguaje natural (PLN) desarrollada por Google.

- BERT es un modelo con incrustaciones de posición absoluta, por lo que generalmente se recomienda rellenar (padding) las entradas a la derecha en lugar de a la izquierda.
- BERT tiene la estructura básica del codificador del transformer, con más cabezas, más capas y tres embedding posicionales que son aprendidos en el entrenamiento.
- BERT fue entrenado con el modelado de lenguaje enmascarado (MLM) y los objetivos de predicción de la siguiente oración (NSP). Es eficiente para predecir tokens enmascarados y en NLU en general, pero no es óptimo para la generación de texto.



Gráfica 5: BERT-Input representacion

Tokenizador de Bert

La tokenización es un proceso que se encarga de dividir un texto en elementos más sencillos con significado propio, además se realiza un proceso de limpieza, eliminando del texto los caracteres que no interesan como signos de puntuación o símbolos especiales. Bert incluye dos modelos de tokenización:

- BERT Uncased: Se pasan a minúsculas todas las palabras y se eliminan todos los acentos.
- BERT Cased: Se realiza la tokenización manteniendo los caracteres en mayúscula y los acentos

5. Desarrollo del modelo.

Cuando utilizas un modelo pre-entrenado, lo entrenas con un dataset específico para tu tarea. Esto se conoce como fine-tuning, una técnica de entrenamiento increíblemente poderosa.

Entrene un modelo Bert para cada país, en este caso utilice el modelo "bert-base-spanish-wwm-uncased", un modelo pre-entrenado más pequeño. Como los datos estaban desbalanceados, había categorías que tenían muchos datos y otras tenían una baja cantidad de datos, utilice como métrica F1 score. Los datos con los que entrene el modelo Bert ya estaban estructurados y estaban listas para entrenar el modelo.

Para entrenar los modelos primero utilice mi computador pero el proceso se demoraba mucho, casi 24 horas duraba el entrenamiento usando 3 ciclos, pero me recomendaron usar Google colab, una herramienta que permite utilizar una GPU, lo cual me permitió entrenar los modelos muy rápidamente.

Debido a que las clases estaban desequilibradas, dividimos el conjunto de datos de forma estratificada, usándo las etiquetas de clase. Dividimos los datos de la siguiente manera: el 85 % de los datos para entrenar el modelo y el 15 % de los datos para la validación del modelo.

Clasificacion	label	data_type	
Aplicación de pago y pedido	6	train	128
		val	22
Aplicación de pagos	2	train	4214
		val	744
Datos maestro	4	train	313
		val	55
Logistica	1	train	312
		val	55
No cliente	5	train	3980
		val	703
Otros	3	train	5275
		val	931
Pedidos	0	train	2229
		val	394

Gráfica 6: Distribución de correos en Colombia

Para fines de ajuste, probé con los siguientes valores:

- Batch size: 16, 32
- Learning rate (Adam): 2e-5, 1e-5
- Number of epochs: 2, 3, 4, 5

Además use el optimizador AdamW. AdamW es una variante del optimizador Adam que tiene una implementación mejorada de disminución de peso. El uso de la disminución del peso es una forma de regularización para reducir la posibilidad de sobreajuste.

5.1. Resultados Obtenidos

Use la puntuación F1 score y la precisión por clase como métricas de rendimiento. He decidido presentar el porcentaje de aciertos sobre el conjunto de validación, que en este caso es el 15 % de los

datos disponibles para entrenar el modelo. Además, incluimos el "F1 score" obtenido al clasificar los datos con el modelo:

País	F1 Score	Precisión validación
Ecuador	0.92	0.93
México	0.88	0.87
Nicaragua	0.88	0.89
El salvador	0.81	0.81
Argentina	0.90	0.90
Colombia	0.89	0.89
Costa rica	0.92	0.90

Los resultados por país fueron buenos a excepción de El Salvador donde el F1 Score fue de 0.81, para los demás países el F1 Score fue superior a 0.87.

Luego cree una función para clasificar nuevos correos, esta función nos arroja la categoría a la que pertenece el nuevo correo, esta función puede calcular las probabilidades del correo de pertenecer a cada categoría, la categoría con la probabilidad más alta es a dónde se clasifica el correo.

Identificamos que cuando un correo queda mal clasificado, la segunda probabilidad más alta, es un valor muy elevado en comparación a cuando el correo queda bien clasificado. Por eso establecimos un umbral para cada país, todo correo cuya segunda probabilidad más alta, supere este valor será considerado como indefinido, ese correo se clasifica en una nueva categoría llamada "indefinido", este correo deberá ser clasificado manualmente.

```python
MAX_LEN = 300
######################################### funcion que sirve para clasificar nuevos correos
def clasficador_correos(review_text):
  encoding_review = tokenizer.encode_plus(
      review_text,
      max_length = MAX_LEN,
      truncation = True,
      add_special_tokens = True,
      return_token_type_ids = False,
      padding='max_length',
      return_attention_mask = True,
      return_tensors = 'pt'
      )

  input_ids = encoding_review['input_ids'].to(device)
  attention_mask = encoding_review['attention_mask'].to(device)
  output = model(input_ids, attention_mask)
  predictions = torch.sigmoid(output.logits).cpu().detach().numpy().tolist()
  #return possible_labels[int(np.argmax(predictions, axis=1))]
  if sorted(predictions[0])[-2] > 0.85:
          print('indefinido')
  else:
      print(possible_labels[int(np.argmax(predictions, axis=1))])

#AQUI ENTRE COMILLAS SE PONE EL NUEVO CORREO
clasficador_correos("Buenos días Amigos de Info.Por Favor anular totalmente el pedido No. 501647457 Por 640 Bolsas de Cemento")
```

Gráfica 7: Función de clasificación de Colombia

Este umbral se calculó teniendo en cuenta que el número de correos indefinidos no fuera demasiado grande, ya que entre menor el límite, mayor es el número de correos indefinidos, y se buscó atrapar el mayor número de correos mal clasificados, Por ejemplo, para Colombia concluimos que el mejor umbral era del 15 %, quiere decir que los correos cuales segundas probabilidades sean superiores a 0.85, serán clasificados en la categoría "indefinido", con este umbral, el 10 % de los correos de validación de colombia fueron clasificados como indefinidos y la precisión es ahora del 0,93.

Pais	correos en total	F1 score	Precisión validación	Umbral	Precision validación despues del umbral	% de correos indefinidos
ecuador	27487	0,92	0,93	0,10	0,94	11%
mexico	12963	0,88	0,87	0,30	0,91	13%
nicaragua	10963	0,88	0,89	0,15	0,91	10%
elsalvador	23983	0,81	0,81	0,07	0,84	21%
argentina	29826	0,90	0,90	0,15	0,94	9%
colombia	19355	0,89	0,89	0,15	0,93	10%
costa rica	18948	0,92	0,93	0,11	0,96	10%
Union 7 paises	143525	0,90	0,90	0,04	0,91	12%
Union sin el salvador	119542	0,90	0,90	0,05	0,92	9%

Gráfica 8: Desempeños de los modelos

Entrene un modelo para cada país, pero después de ver los resultados de de cada país también decidimos entrenar un modelo con todos los datos de los siete países para comparar resultados. pero

como se ve en la gráfica 8, no hay gran diferencia.

Como el desempeño del modelo de El Salvador no fue muy bueno, entrene un modelo con los datos de seis países sin los datos de El Salvador, pero no hubo mucha diferencia con los demás modelos.

6. Conclusiones

A través de esta experiencia logre afianzar los conocimientos relacionados con inteligencia artificial, Transformers, procesamiento del lenguaje natural, modelos Bert. Se reforzaron los conocimientos sobre como entrenar un modelo de clasificación múltiple de texto al implementar un modelo Bert y pudimos automatizar el proceso de clasificación de los correos.

Como se ha revisado en la pasantía, Bert es una herramienta muy útil para las tareas relacionadas con el análisis de texto, permite relacionar las palabras con el contexto que las rodea, gracias a esto, es capaz de generar una representación consistente del texto que está analizando.

El entrenamiento y evaluación de los modelos para la clasificación de texto multiclase permitió obtener resultados superiores al 89 % en las medidas de precisión y F1-score de los modelos entrenados de casi todos los países.

7. Bibliografia

[1] D. Bikel and I. Zitouni, Multilingual natural language processing applications: from theory to practice, IBM Press, 2012.

[2] Dr. Dataman, 2018, Looking into Natural Languaje Processing (NLP). https://medium.com/dataman-in-ai/natural-language-processing-nlp-forelectronic-health-record-ehr-part-i-4cb1d4c2f24b

[3] Rémi Louf, 2019, Encoder-decoders in Transformers: a hybrid pre-trained architecture for seq2seq. https://medium.com/huggingface/encoder-decoders-intransformers-a-hybrid-pre-trained-architecture-for-seq2seq-af4d7bf14bb8

[4] J. Howard and S. Ruder, Universal language model fine-tuning for text classification, arXiv preprint arXiv:1801.06146, (2018).