Diseño de una estrategia de limpieza y estandarización de direcciones postales a través de redes neurales recurrentes tipo LSTM

Santiago Ceballos Gallego

Universidad Nacional de Colombia
Facultad de ingeniería, Departamento de ingeniería de Sistemas e Industrial)
Bogotá, Colombia
2020

Diseño de una estrategia de limpieza y estandarización de direcciones postales a través de redes neurales recurrentes tipo LSTM

Santiago Ceballos Gallego

Trabajo final de maestría presentado como requisito parcial para optar al título de:
Magister en Ingeniería de Sistemas y Computación

Director(a):
Ph.D. Fabio Augusto Gonzaláez

Línea de Investigación:
Machine Learning
Grupo de Investigación:
MindLab

Universidad Nacional de Colombia
Facultad de ingeniería, Departamento de ingeniería de Sistemas e Industrial)
Bogotá, Colombia
2020

Agradecimientos

A mi tutor Ph.D. Fabio Augusto Gonzaláez, por creer en el valor del proyecto desde el principio, por las múltiples ideas para obtener mejores resultados, la guía, los consejos y sobretodo, por recordarme, durante el desarrollo de este trabajo el valor de la frase "Trabajo y Rectitud", lema que me acompaña desde el pregrado en la Facultad de Minas de la Universidad Nacional de Colombia.

A mi organización y equipo de trabajo en el ámito laboral, quienes no sólo me dieron la oportunidad de aplicar los conocimientos adquiridos en en análisis de datos, durante el desarrollo de mi maestría, sino que además me permitió expandir dichos conocimientos liderando el área de Data Analytics para Colombia, Chile y Perú. Fueron los recursos, la flexibilidad y el apoyo de esta empresa para convertir las aplicaciones en valor real de negocio, las que permitieron realizar el presente trabajo aplicado.

Por último, a mi familia, quienes creyeron incondicionalmente en mi y en mis capacidades para sacar la maestría adelante, en especial en momentos de múltiples responsabilidades, Y que estuvieron para mantenerme motivado y recordarme el porqué del esfuerzo realizado. Sin ustedes no hubiera posible finalizar este documento.

Resumen

Las direcciones geográficas son uno de los elementos más comunes en las bases de datos de diferentes tipos de organizaciones. Sin embargo, el registro de dichas direcciones se realiza, a menudo, de forma manual y sin un formato de referencia, lo que da lugar a múltiples representaciones de los elementos que componen la dirección. Esto, a su vez, genera que el registro sea usualmente inutilizable para fines de geolocalización automática, un área cada vez más relevante en los principales sectores de la economía.

En el presente documento se propone una metodología para la limpieza y estandarización de direcciones geográficas, basada en redes neuronales recurrentes tipo LSTM, como solución a este problema. Dicha metodología, incluye la estrategia de generación de un conjunto de datos sintético, para el entrenamiento de la red, que está compuesto por direcciones no estructuradas y las direcciones equivalentes en formato estandar. El desempeño del modelo se mide en dos conjuntos de datos diferentes: El primero contiene 10000 direcciones sintéticas sucias y su equivalente limpio, contra el cual se compara la dirección genearada utilizando los indices de Jaccard, Jaro y Levenshtein, como medidas de similitud; el segundo, contiene 5000 direcciones reales de establecimientos comerciales en las tres principales ciudades de Colombia, para los cuales se cuenta con la geolocalización exacta. Esta ubicación real se compara con la obtenida tras geolocalizar la dirección resultante del proceso de estandarización.

Al aplicar esta estrategia, se evidencia una mejora significativa tanto en la precisión del formato estandar obtenido, como en la geolocalización de la dirección resultante, cuando se compara contra los dos modelos base más utilizados en este campo: el modelo basado en reglas de limpieza y el modelo basado en cadenas de Markov ocultas.

Por último, se muestran aplicaciones de la metodología para limpieza y geolocalización de direcciones tomadas de una base de datos real, en ámbitos como la optimización de fuerza de ventas, la atención al cliente y el mercadeo digital.

Conceptos clave: (Address standardization, Recurrent Neural Networks, Long-Short Term Memory, Hidden Markov models(HMM), Geocoding, Text similarity).

Este Trabajo Final de maestría fue calificado en noviembre de 2020 por el siguiente evaluador:

Elizabeth León Guzmán Phd.
Profesora Departamento de Ingeniería de Sistemas e Industrial
Facultad de Ingeniería
Universidad Nacional de Colombia

Contenido

1. Introducción

Las direcciones postales están presentes en muchas bases de datos pertenecientes a diferentes tipos de organizaciones, sobre todo aquellas enfocadas a atender a una gran cantidad de consumidores: Tal es el caso de bancos, universidades, hospitales, empresas de comercio electrónico, servicio postal y/o empresas dedicadas a la compra y venta de productos de consumo masivo [1][2]. Sea cual sea el ámbito, este tipo de organizaciones utiliza la información presente en dichas bases de datos para obtener información relevante sobre clientes, personas, hábitos de consumo, lugares y otras organizaciones [3].

Sin embargo, debido a la naturaleza de una dirección postal, su registro en una base de datos determinada suele hacerse de forma manual, y según el criterio de la entidad que brinda la dirección y de aquella que lo registra [1]. Esto hace que dichos registros estén sujetos a variaciones en la forma de escribir cada uno de los componentes de la dirección, el orden de esos componentes, los separadores utilizados y, en los casos en los que se hace de forma completamente manual, a errores de tipeo y ortográficos. Lo anterior, sumado a que en varios de los países de Latinoamérica no existe un formato estándar [4], hace que, en muchos casos, estas direcciones sean inutilizables dentro de las posibles aplicaciones que podría tener. La razón es que todas ellas implican transformar cada una de esas direcciones a una ubicación geográfica (ej. Latitud y longitud), lo cual a su vez implica que estén en un formato determinado [5].

En el presente trabajo, se propone diseñar una estrategia que permita, a partir de una base de datos de direcciones, identificar estas diferencias en el texto y llevarlas a un formato único que pueda ser leído por los principales aplicaciones de geolocalización.

1.1. Definición del problema

Las direcciones postales son comunes en diversas bases de datos y múltiples usos en el sector financiero, comercial, logístico, áreas de la salud y seguridad pública etc. (Ver sección 6.2). Sin embargo, en una gran mayoría de casos el registro de esta dirección se hace de manera manual y de modo tal que el usuario que los registra no está restringido por ningún formato que deba seguir [6]. Los marcos de referencia utilizados para dicho registro de direcciones en una base de datos, son diferentes para cada persona, lo cual a su vez resulta en el uso de

Figura 1-1.: Ejemplos de escritura para una misma dirección geográfica.

definiciones y términos diferentes para el mismo elemento de la dirección [7].

La figura **1-1** muestra diversas formas posibles de escritura para una misma dirección en el formato colombiano. Adicionalmente, se pueden encontrar ejemplos variados de este tipo de diferencias entre los formatos, en los estudios presentes en [8],[9], [10], [11] realizados en tailandia, india y turquía respectivamente.

Las cuatro direcciones presentes en la figura 4 corresponden a una misma ubicación geográfica de un local comercial en una determinada ciudad de Colombia. Sin embargo, dadas las variantes en los componentes y el orden de los elementos es fácil obtener desde cuatro fuentes diferentes, formatos totalmente diferentes para la dirección. De ahí precisamente la dificultad técnica de llevar estas direcciones a un formato estándar, pues el método utilizado no sólo tiene que identificar a qué componente corresponde cada grupo de caracteres, sino también la ausencia de caracteres como lo es el complemento de barrio (Laureles) o la falta de separadores entre 50A y 16, en la segunda opción.

En resumen, se podría decir que las diferentes variaciones de una misma dirección están dadas por los siguientes cinco elementos [12]:

- Hay múltiples convenciones para las direcciones que el método debe manejar, de acuerdo al dominio específico de la geografía en la que se esté aplicando.
- Existen múltiples formas de ordenar y escribir una dirección.
- Puede haber varias interpretaciones para una cadena de texto correspondiente a una dirección (Por ejemplo, la dirección de la figura 4 se podría encontrar en Bello Antioquia o en Medellín siendo municipios que limitan el uno con el otro.

- Existen ambigüedades tanto semánticas como sintácticas dentro de la escritura de direcciones. Adicionalmente, la persona que registra la dirección puede tener errores de tipeo u ortografía que dificultan el uso de la misma.
- Los elementos presentes en una dirección postal pueden variar significativamente de un país a otro (e incluso de una ciudad a otra): Mientras que el identificador de avenida en ciertos países es un número con cierta cantidad de dígitos máxima, en otros corresponde a un nombre específico.

Además del hecho de que las direcciones se registran en las diferentes bases de datos de forma, generalmente, manual. Existen algunos otros elementos que contribuyen a que la complejidad del problema de estandarización de direcciones sea alta.

1. La falta de un estándar de direcciones en determinados mercados.
   Muchas de las aplicaciones encontradas en la actualidad para estandarización de direcciones se basan en diccionarios y reglas predefinidas que sirven para determinados mercados en los que se cuenta con un estándar claro de direcciones y por lo tanto es el formato final en el que debe salir la direcció postal, luego de pasar por el proceso de estandarización. Tal es el caso de Estados Unidos que cuenta con varias aplicaciones [13], [14] Sin embargo, países como India, Turquía, pakistan, Tailandia o los países de Latinoamérica [15], [6],[8], no cuentan con este tipo de sistemas estándar lo que dificultad llevar dichas direcciones a un único formato. En este sentido hay evidencia de propuestas para crear este tipo de estándar en dichos mercados como la que se puede observar en [4].

2. Cómo obtener el conjunto de datos de entrenamiento o referencia para aplicar en modelos de estandarización de direcciones.

   La gran mayoría de modelos aplicados a la limpieza de información no estructurada requiere de un conjunto de datos de entrenamiento del que se puedan extraer las características o parámetros a tener en cuenta para llegar a una estructura definida. En el caso de las direcciones, esto implicaría limpiar y etiquetar manualmente una gran cantidad de direcciones lo cual es costoso en términos de recursos y no se puede garantizar que este etiquetado se haga correctamente [12]. Adicionalmente, esta es precisamente la tarea que se quiere evitar al desarrollar un programa para la limpieza de direcciones. De allí que es necesario pensar en técnicas que permitan transferir conocimiento desde otros ámbitos para los que ya haya modelos entrenados, como se puede ver en [12]. O bien, diseñar estrategias que permitan aprender desde información que no esté completamente estándar, o generar un conjunto de datos sintético para el entrenamiento

del modelo.

Las dos características acá mencionadas serán tomadas como desafíos para la elaboración de esta propuesta y serán considerados en las secciones siguientes.

1.2. Objetivos

Objetivo Principal

- Diseñar un método para limpiar direcciones postales de manera automática, transformándolas desde formatos no estructurados a un formato estándar.

Objetivos Específicos

- Construir un set de datos de entrenamiento de direcciones, incluyendo las variaciones más comunes registradas en las bases de datos a las que se tenga acceso.
- Diseñar e implementar un modelo basado en redes neuronales recurrentes que esté al nivel del estado del arte para estandarización de direcciones postales utilizando las direcciones generadas como set de entrenamiento.
- Evaluar el desempeño del modelo propuesto versus modelos de línea base.

1.3. Esquema

Este documento está organizado en 5 capítulos. En el capítulo 1 se presenta la definición del problema y motivación del estudio, así como los objetivos principales y específicos. En el capítulo 2 se presenta una revisión del estado del arte en temas relacionados con limpieza y estandarización de direcciones, partiendo de las aplicaciones recientes e incluyendo las principales metodologías utilizadas en temas relacionados, contribuciones de los diferentes autores y sus resultados. En el capítulo 3 se muestra el detalle de la elaboración y codificación de dos modelos base: Uno basado en reglas de estandarización y el segundo basado en cadenas de Markov Ocultas. En el capítulo 4 se explica en detalle la metodología propuesta para la limpieza y estandarización de direcciones utilizando redes neuronales recurrentes de tipo LSTM. En el capítulo 5, se muestra la comparación de resultados al aplicar los diferentes métodos sobre diversos sets de datos de direcciones pertenecientes a establecimientos comerciales. Por último, en el capítulo 6 se muestran las conclusiones y trabajo futuro en el tema de estudio.

2. Antecedentes y Trabajo Relacionado

2.1. Aplicaciones reales de estandarización de direcciones postales

Son varios los casos de uso encontrados, tanto en la literatura como en la industria, para las direcciones postales y en ámbitos muy variados. Uno de los más comunes es la geolocalizacién de personas, clientes, locales comerciales o eventos, entendida como el proceso de obtener una representacién geográfica (como las coordenadas de latitud y longitud) a partir de una descripcién de dicha ubicación, que generalmente es una dirección [5]. Este proceso de geolocalizacién es crítico para análisis espaciales a su vez aplicados en diferentes áreas del conocimiento como lo puede ser la compra, envío y entrega de mercancías y mails, análisis de carácter social y demográfico, estudios ambientales o en el área de salud [5]. A continuacién, se describen algunos de dichos estudios, ámbitos y lugares en los que se han aplicado.

2.1.1. Aplicaciones en el área comercial y logística

En la industria, uno de los ejemplos más comunes tiene que ver con la optimización de la cadena de suplemento y la red de distribución de las empresas a través de la geolocalización de proveedores y clientes a partir de sus direcciones, lo cual a su vez permite la optimización en todo el proceso de planeación de la demanda [16]. La ubicación de clientes y locales comerciales permite también la reducción de los costos asociados a la entrega de correo y mercancías, a la vez que se asegura que el producto o servicio es entregado en la dirección correcta [13]. Lo anterior, toma cada vez más relevancia en un medio social en el que el e-commerce, aplicaciones de servicio de entrega a domicilio y las ventas por internet, muestran crecimientos cada vez más acelerados. [17]

Otros casos, encontrados tanto en la industria como en la literatura, se concentran en el tener la dirección postal con buena calidad en las bases de datos, como medida base para análisis estadísticos, como lo puede ser el análisis de información relacionada con los clientes combinado con información demográfica adicional, que le permite a las diferentes organizaciones mejorar sus planes de mercadeo, planear expansiones futuras, la identificación de patrones de consumo y la división de sus clientes reales o potenciales en pequeños clusters utilizados en análisis estadísticos adicionales [18].

2.1.2. Aplicaciones en el área de la salud

El área de la salud es una de las que más estudios presenta en temas relacionados con el uso de direcciones estándar y posterior geolocalización en diferentes ámbitos, tales como el estudio detallado de las epidemias, que permite identificar áreas de exposición potencial que pueden ser causa o consecuencia de dichas epidemias. Dos ejemplos de esos estudios se pueden encontrar en [19],[20].
En este mismo sentido, el estudio de direcciones postales permite identificar grupos locales donde se concentra una determinada enfermedad y ubicar a las personas que viven cerca de dichas áreas [18] [21]. Esto a su vez facilita la planeación de tratamientos y centros de atención al permitir ubicar centros de atención y emergencias en aquellas ubicaciones con una mayor necesidad del servicio. Otro caso de uso importante tiene que ver con los servicios de atención de emergencias, en el que es común recibir direcciones incompletas o ambiguas de aquellos que reportan la emergencia pero es crítico contar con una ubicación exacta para atenderla correctamente [22].

2.1.3. Aplicaciones en el ámbito social

Dado que las direcciones postales son uno de los referentes más comunes a eventos y fenómenos que ocurren en áreas urbanas [22], también se convierten en referentes para el estudio de varios de dichos fenómenos, como lo es el estudio del tránsito y el transporte público, la seguridad pública, la recolección de impuestos, entre otros [2, 22].
Dos de los eventos que más se asocian con direcciones postales, son los crímenes y accidentes cuya ocurrencia se registra en un lugar geográfico determinado. En [23], Ratcliffe muestra la importancia del éxito en la geolocalización para el análisis de este tipo de eventos y demuestra, utilizando simulación Montecarlo, que la precisión mínima en la geolocalización de los eventos para obtener un análisis acertado es de 85 %. Dicha geolocalización se obtiene precisamente a partir de direcciones postales.
Lo anterior también aplica para realizar análisis de riesgo y crear programas de prevención y atención de desastres [24] de acuerdo a medidas socioeconómicas y el nivel de vulnerabilidad de una determinada ubicación. Esto a su vez permite que la limpieza de direcciones sea susceptible de ser usada en el desarrollo de políticas públicas como se demuestra en [24].

2.2. Estrategias utilizadas para segmentación y estandarización de información no estructurada.

A nivel mundial existe evidencia de diversos estudios, aplicaciones y organizaciones dedicadas a solucionar el problema, ya se una vez está presente en la base de datos a través de métodos de calidad de información (data quality) o bien, desde antes del registro, a través de la implementación de reglas que no permiten ingresar la dirección en un formato libre [14]. Para

apps.png

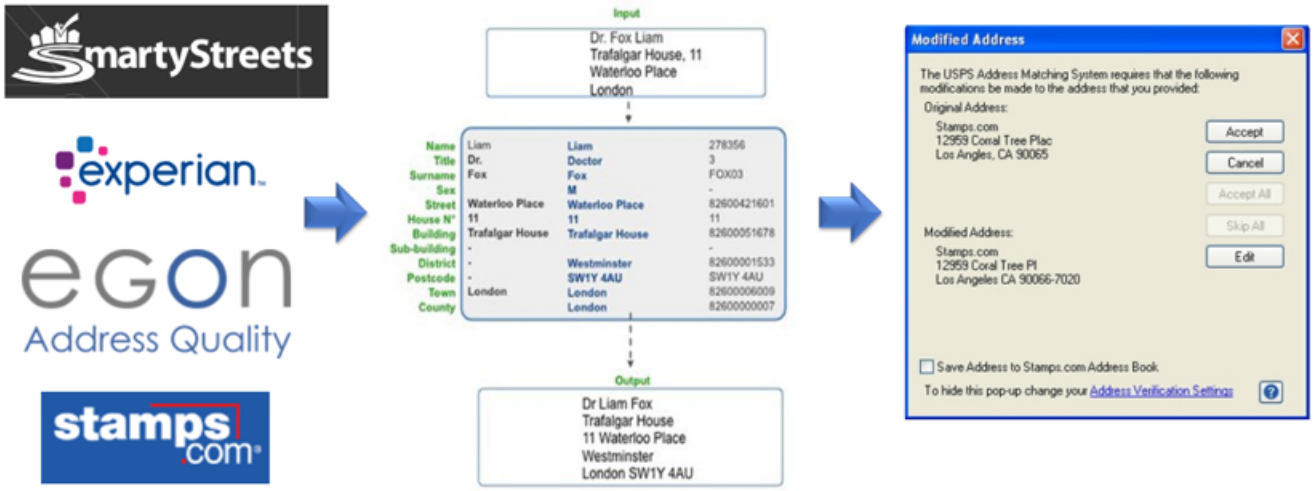

Figura 2-1.: Aplicaciones comerciales para limpieza de direcciones postales[25, 26, 27]

mayor detalle sobre los principales conceptos involucrados en el proceso de estandarización de direcciones y aplicaciones recientes, se puede consultar los informes presentes en [5, 13].

2.2.1. Aplicaciones comerciales para estandarización de direcciones.

A nivel global, existen varias organizaciones que ofrecen el servicio de limpieza, estandarización y, en algunos casos, geolocalización de direcciones, ya sea a partir de una base de datos o de registros individuales. A continuación, se mencionan aquellas con resultados más relevantes y con mayor alcance de acuerdo a la información disponible en páginas web especializadas en el tema. Dichas aplicaciones se muestran en la figura **2-1**.

Entre las aplicaciones comerciales se encuentra Smartystreets: Una compañía de servicios de software que provee verificación de direcciones y servicios de geolocalización a través de varias modalidades, entre ellas la limpieza de direcciones únicas, listado de direcciones y aplicaciones que se pueden integrar en diferentes páginas web u otros servicios [25]. Un servicio similar es ofrecido por Experian, empresa enfocada en la limpieza de direcciones para entrega de correos, utilizando las convenciones del servicio postal de los estados unidos [26]. La versión europea de este tipo de firmas se conoce como Egon address quality, empresa basada en Verona que busca constituirse como el líder en Europa en este tipo de aplicaciones [27].
Si bien no se tiene acceso directo al código utilizados por dichas organizaciones, es sabido que muchas de las aplicaciones utilizadas para limpieza de direcciones se basan en reglas estándar para cada uno de los componentes de la dirección que logran buenas generalizaciones de las direcciones, en contextos en los que el formato está bien definido como lo es el servicio postal

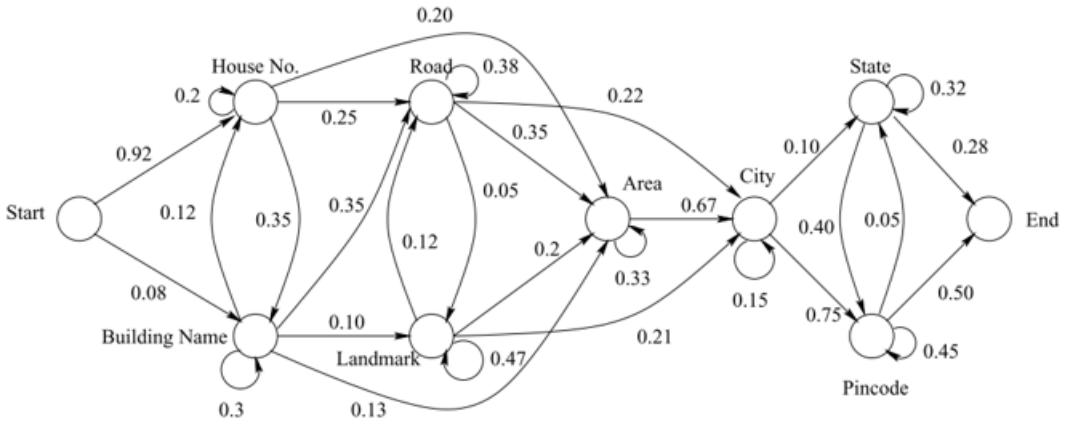

Figura 2-2.: Naive Hidden Markov Model. Estructura básica[1]

de los estados unidos. Sin embargo, en países en los que no existe un formato estándar para la dirección, crear reglar que cubran la gran variedad en los patrones posibles en la dirección requiere de mucho esfuerzo, para obtener precisiones aceptables [6]. Es por ello que son varios los estudios académicos que proponen soluciones para el problema de estandarización de las direcciones.

2.2.2. Modelos basados en cadenas de Markov ocultas (HMM)

Varios de los estudios que se han consolidado como modelos base para la estandarización de las direcciones están basados en cadenas de Markov ocultas, método que es aplicado a procesos en los que hay estados ocultos que generan evidencias desde las cuales se busca identificar el parámetro oculto [7]. En el caso específico de las direcciones, las cadenas de Markov ocultas (HMM) buscan identificar a qué elemento corresponde cada una de las palabras de la dirección y a partir de allí llevar dichas palabras o componentes a la forma estándar del elemento definido.
En [7] Borkar y compañía parten de un modelo de Markov simple como el que se muestra en la figura **2-2** y a partir de allí proponen un modelo anidado en el que cada uno de los nodos de la cadena de Markov es, a su vez, un segundo modelo de Markov que busca identificar cuantos elementos (palabras) tiene cada componente.
Con este modelo se obtuvieron resultados de 99.6 % de precisión en direcciones de estados unidos y 89 % en un set de datos de direcciones en India.

Una versión mejorada de este algoritmo fue propuesta por Bokar et. Al en [1], estudio en el que se agrega el concepto de taxonomía de los elementos que pretende diferenciar los elementos por parámetros como la longitud de la palabra o del número o la naturaleza

de las palabras, e incluye dicha taxonomía dentro del proceso de limpieza. Las cadenas de Markov ocultas se siguen utilizando en enfoques híbridos más recientes como el encontrado en [8], en el que se procesa la información con una serie de reglas y un diccionario creado específicamente para las direcciones de Tailandia antes de ingresar la dirección al modelo de Markov, Con este modelo logran una precisión del 97 % en el set de datos utilizado.

2.2.3. Otros modelos probabilísticos

Los HMM no son el único modelo utilizado para hacer limpieza de direcciones, recientemente Wang et al. [28] propusieron un método basando en campos condicionalmente aleatorios (CRF) y Gramática estocástica regular (SRG). Dichos métodos combinados demuestran ser mejor en la tarea de asignar a qué corresponde cada elemento de una determinada dirección frente a abreviaciones complejas como nombres de calles y carreras y obtienen una mejora promedio de 5 % vs los resultados obtenidos por modelos de HMM aplicados en los mismos sets de datos.
Además del CRF, existen aplicaciones recientes para solucionar el problema de estandarización de las direcciones basadas en procesamiento de lenguaje natural y redes neuronales: En el primer caso se utilizan algoritmos que toman cada elemento de la dirección y lo llevan a su forma estándar al identificar la palabra estándar que se asemeja más a la realmente utilizado en la dirección. Para ello utilizan algoritmos como la distancia de Levenshtein obteniendo un 99 % de direcciones geolocalizables versus un 64 % previo a la dirección sin proceso previo. [7]

2.2.4. Enfoques basados en redes neuronales

Todos los estudios basados en cadenas de Markov ocultas implican la segmentación manual de direcciones como set de entrenamiento de modo que se puedan establecer los estados y las probabilidades de transición de la cadena. Dado que esta es una tarea exhaustiva y que implica una gran cantidad de esfuerzo manual Kotari et. al proponen un método basado en redes neuronales para transferir conocimiento desde una fuente ya etiquetada a una no etiquetada utilizando un proceso jerárquico de Dirichlet demostrando una mejora del 13 % al alimentar un determinado modelo con información etiquetada de otra fuente. [12]
Por otro lado, las redes neuronales han demostrado poder dar solución a este problema de manera directa, como lo demuestra Sharma et. al en [3], estudio en el que proponen usar una red neuronal con una única capa oculta y una representación de tipo one-hot para clasificar cada uno de los elementos que componen la dirección, y con el cual obtienen un 97 % de precisión al aplicarlo sobre un set de datos de direcciones de Alemania y reino unido. Este resultado es 1.4 % mejor que el obtenido al aplicar el método de CRF, antes mencionado y 6 % mejor que el obtenido con modelos de Markov ocultos.

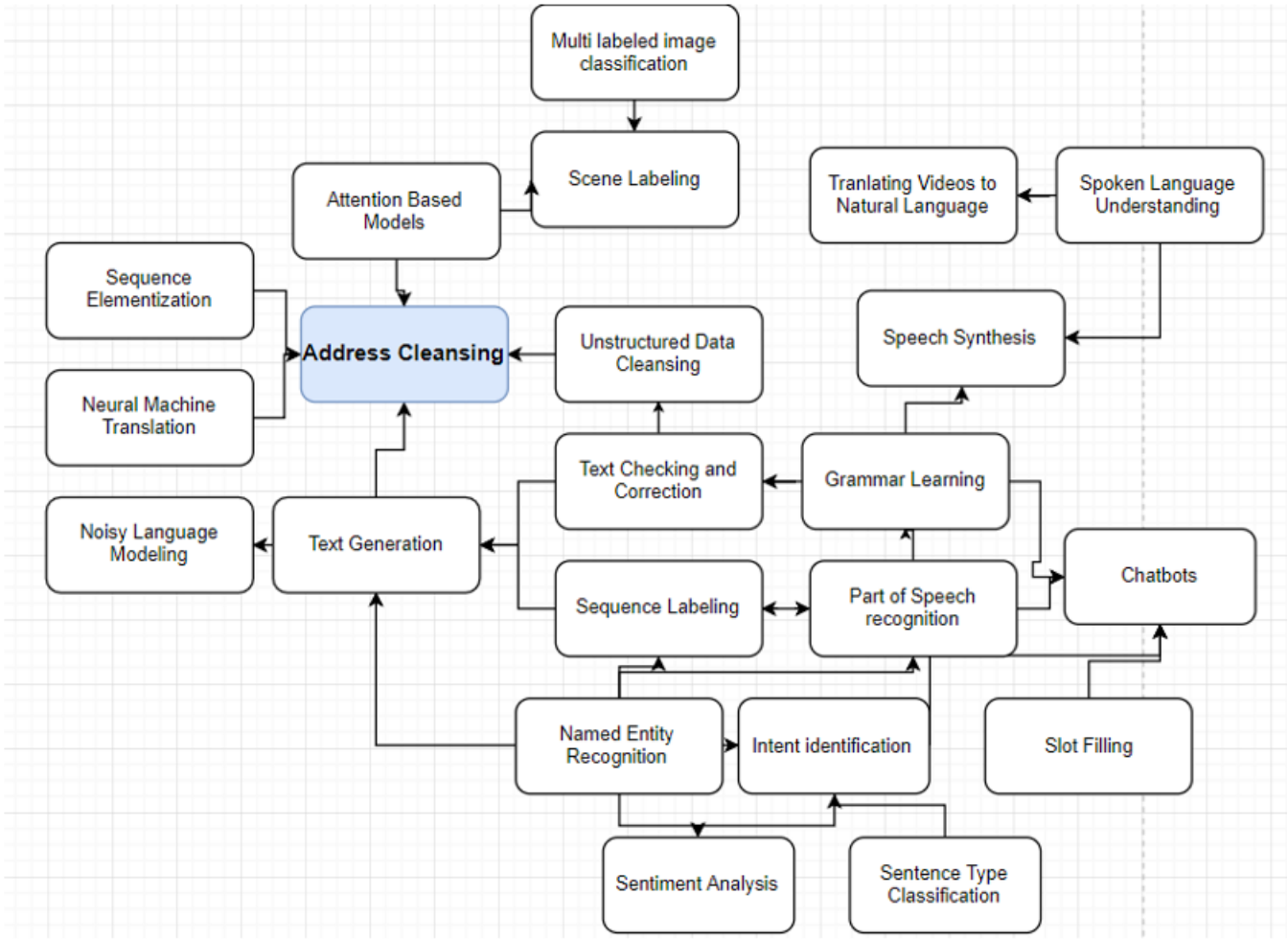

Figura 2-3.: Estado del arte. Aplicaciones relacionadas con limpieza y estandarización de direcciones. Elaboración propia a partir de [34],[35],[36],[37],[38],[39],[40]

2.2.5. Otras metodologías aplicadas a problemas similares

Si bien, a la fecha de construcción de la presente propuesta, no se encontró evidencia de aplicaciones adicionales de redes neuronales al problema de estandarización de dirección, este y otros tipos de redes si se han aplicado a problemas similares como lo son el part of speech tagging [29], sequence labeling y slot filing [30, 31, 32, 33], [31],[32], [33], entre otros [34, 35, 36, 37, 38, 39, 40]. La figura **2-3** muestra un esquema de las relaciones que se dan entre los diversos temas (y las metodologías aplicadas en cada uno de ellos) y que por lo tanto se presentan como estrategias que podrían ser aplicadas, de uno u otro modo en la elementización, limpieza y estandarización de direcciones postales.

De entre los presentes en la figura **2-3**, el análisis de sentimiento es uno de lo que más crecimiento ha presentado en los últimos años, debido al auge y disponibilidad de información en redes sociales e internet. Este campo, estudia y analiza las opiniones, sentimientos y

emociones de personas ante determinadas entidades a partir de oraciones expresadas de forma escrita [41]. Esta labor implica en muchos casos, clasificar cada uno de los elementos de la frase de acuerdo a si expresan algo positivo o negativo, así como identificar el elemento de la frase completa. Para ello se ha propuesto utilizar una variedad de redes neuronales tipo LSTM [37],[41],[42], en conjunto con redes neuronales convolucionales que procesan grupos de palabras antes de ingresar a la capa LSTM [41]. Este enfoque se utiliza también para la clasificacion de oraciones, la identificación de intención y el etiquetado de porciones de discursos, todos ellos relacionados con procesamiento de lenguaje natural [30],[40],[43]
Estas aplicaciones tienen una fuerte relación con aquellas que buscan interpretar parte de un texto, pero no con el fin de clasificarlo sino con el fin de corregirlo. Tal es el caso de slot filling en el que se busca identificar espacios vacíos y llenarlos con la palabra o grupo de caracteres que debería estar allí. [32],[33], lo cual tiene aplicación a la estandarización de direcciones en aquellos casos en los que en el registro de dicha dirección se omiten caracteres como identificadores de calles o avenidas o separadores entre los componentes. De este mismo modo los métodos que se concentran en la corrección de texto y el aprendizaje de elementos gramaticales podrían ser aplicados en este campo. [39], [43]
Por otro lado, las aplicaciones de redes neuronales utilizadas para la generación de cadenas de texto y el modelado de lenguaje ruidoso, proveen una base para la generación de modelos que permitan generar direcciones aleatoriamente y con ello aumentar los sets de entrenamiento utilizados. En este ámbito, las redes neuronales recurrentes y con memoria a corto y largo plazo también juegan un papel predominante como se puede observar en [44], [45]

3. Estrategia de Generación de Datos

Para el entrenamiento de este modelo es necesario contar con un set de datos significativo que tenga direcciones en la mayor cantidad de formatos posibles y su valor en el formato estándar definido. Algunos ejercicios similares al propuesto en este documento, como lo son los programas de traducción, utilizan set de datos disponibles online cuyo tamaño varía entre 2000 y 5000 líneas [46, 47, 48]. Este set de datos se puede considerar como pequeño para la tarea a realizar, por lo que generalmente los autores generan una gran cantidad de epochs para el entrenamiento del modelo (+1000) [47]. Puesto que para esta aplicación en particular no se cuenta con un set de datos ya construido, procedemos a generar uno sintético de 200.000 observaciones, de acuerdo con el siguiente proceso:

1. Identificar los principales elementos que componen una dirección y las diferentes secuencias (orden) de dichos elementos dentro de una dirección geográfica válida. Dichas secuencias también están regidas por probabilidades de transición entre un elemento y otro que es necesario identificar y utilizar a la hora de generar diferentes direcciones.

2. Identificar las principales causas de diferencias entre los diferentes formatos de codificación de una dirección. Esto es, determinar cuáles son aquellas variantes de palabras utilizadas en una misma ubicación de una dirección a partir de un set de datos real (ej. CL, CLL ó CLLE; #, -, No ó N.; CR, KR, CRR ó CARRERA).

3. Generar combinaciones aleatorias a partir de cada todas las variantes que hay en las diferentes posiciones de una dirección en particular.

3.1. identificación de variantes y frecuencia de aparición

En total se identifican 15 posibles términos que conforman una dirección. Dichos términos se listan en la tabla **3-1**. Nótese que los elementos listados no necesariamente están presentes en todas las direcciones geográficas y su oden puede variar significativamente entre una dirección y otra.

Término	Descripción
1	Primer identificador: Calle, Carrera, Diagonal o transversal
2	Número identificador de la Calle o carrera. RAND(1-200)
3	Identificar separador entre calles y carreras (#, No. NUM, -, Tab, etc).
4	Nombre de la avenida en cuestión en valor nominal (Ej. Avenida Pepe Sierra)
5	Complemento para el número de la avenida (A, AA, BIS, C, etc.)
6	Palabra clave para zonas en la ciudad (SUR, BIS, NORTE)
7	Segundo identificador: Calle, Carrera, Diagonal o transversal
8	segundo número identificador en la dirección. RAND(1-200)
9	separador segundo elemento de la dirección (#, No. NUM, -, Tab, etc).
10	Nombre de la avenida en cuestión en valor nominal (Ej. Av. Caracas)
11	Complemento para el número de la avenida (A, AA, BIS, C, etc.)
12	Palabra clave para zonas en la ciudad en el segundo elemento (SUR, BIS, NORTE)
13	segundo Complemento (barrio, manzana, urbanización, apartamento, etc.)
14	Separador de número (, -, tab)
15	Número correspondiente a la residencia específica. RAND (1-100)

Tabla 3-1.: Descripción de términos utilizados en la creación del set de datos

3.1.1. Cálculo de probabilidades de transición entre los diferentes elementos

Con el fin de calcular las posibles secuencias que se pueden formar con cada uno de los elementos mencionados en la tabla **3-1**, su orden y las probabilidades de transición de un estado a otro, se utilizó un total de 1200 direcciones correspondientes a establecimientos comerciales en Colombia. Dichas direcciones se descompusieron en unidades específicas de información (también llamados elementos o celdas), que luego fueron clasificados en las 15 categorías mencionadas.
Una vez realizado este proceso es posible calcular la probabilidad de que un elemento i de la dirección venga inmediatamente después de otro j aplicando la siguiente fórmula.

$$P_{ij} = \frac{\text{Conteo de transiciones del elemento i al elemento j}}{\text{Total de apariciones del elemento i}} \tag{3-1}$$

La figura **3-1** muestra gráficamente los estados y las probabilidades de ir de uno a otro en determinada secuencia de una dirección greográfica. Dichas probabilidades se utilizan para generar las direcciones.

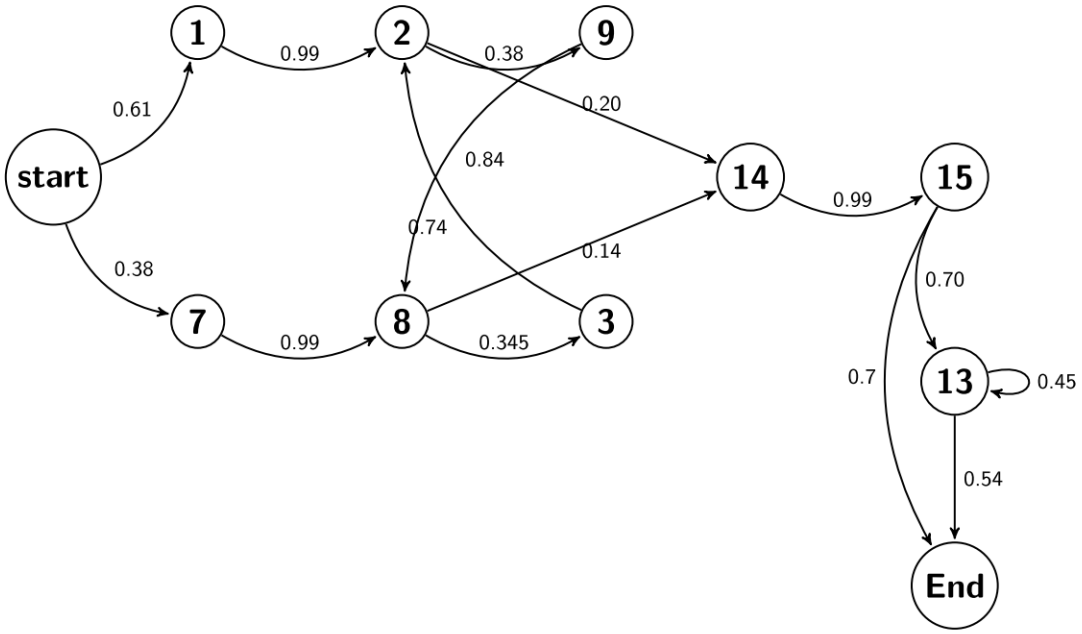

Figura 3-1.: Ilustración elementos y sus probabilidades de transición

3.1.2. Generación aleatoria de variaciones de un determinado elemento

Una vez determinada la secuencia o tipo de dirección que se va a generar es necesario considerar las posibles variaciones de un determinado elemento que se pueden presentar en las direcciones presentes en bases de datos reales que, en la gran mayoría de los casos, no están en una forma estandar si no sujetos a la subjetividad del usuario que ingresa dicha dirección.

Para reflejar estas variaciones dentro de nuestro generador aleatorio de direcciones, se utilizaron apróximadamente 200000 direcciones a partir de las cuales se identificaron las posibles palabras, o símbolos utilizados en cada uno de las diferentes categorías de elementos y el conteo de apariciones de dichos elementos a partir del cual se calcula la probabilidad. La figura **3-2** reflejan este caso para las variantes de la palabra "CARRERA", en su forma estandar. Con dichas probabilidades se calcula también la probabilidad acumulada que permite seleccionar una de las variantes generando un número aleatorio entre 0 y 1 e identificando el rango de la variante en el que éste se encuentra. Dichas probabilidades se muestran en la tabla **3-3**

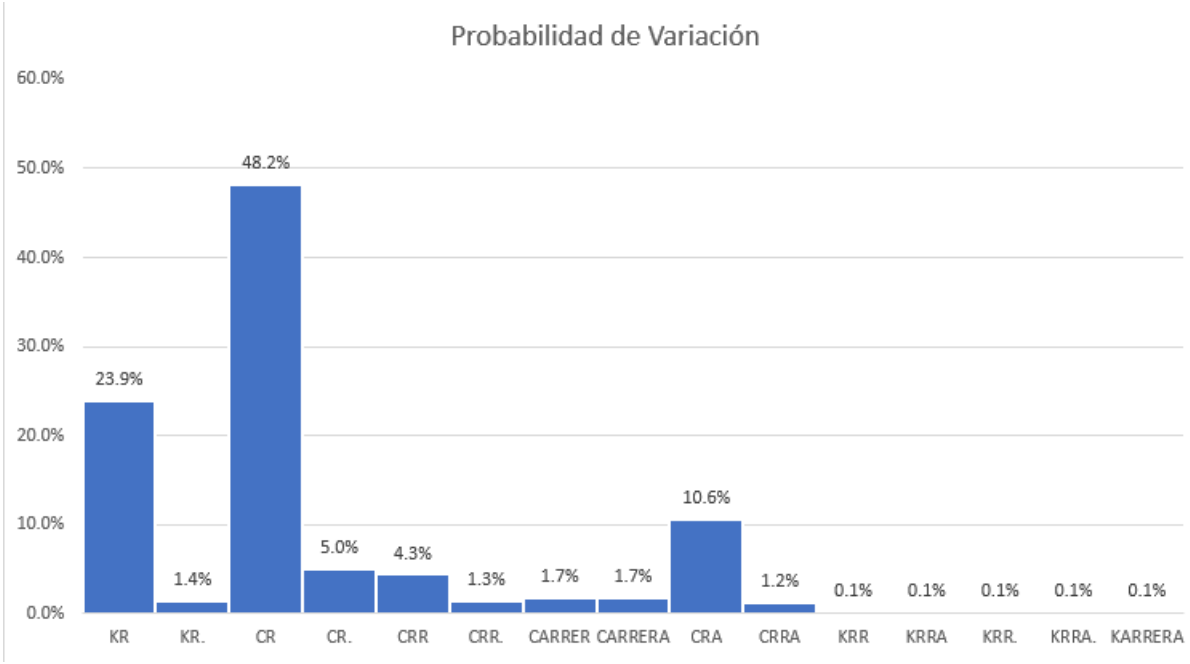

Figura 3-2.: Probabilidades de aparición para las variantes de "CARRERA"

Primer Término. Variaciones Carrera				
Correcto	**Variación**	**Conteo**	**Probabilidad**	**P. acumulada**
CARRERA	KR	19822	23.9%	23.857%
CARRERA	KR.	1139	1.4%	25.227%
CARRERA	CR	40031	48.2%	73.407%
CARRERA	CR.	4175	5.0%	78.431%
CARRERA	CRR	3612	4.3%	82.778%
CARRERA	CRR.	1111	1.3%	84.116%
CARRERA	CARRER	1425	1.7%	85.831%
CARRERA	CARRERA	1423	1.7%	87.543%
CARRERA	CRA	8836	10.6%	98.178%
CARRERA	CRRA	1000	1.2%	99.381%
CARRERA	KRR	109	0.1%	99.513%
CARRERA	KRRA	104	0.1%	99.638%
CARRERA	KRR.	100	0.1%	99.758%
CARRERA	KRRA.	100	0.1%	99.878%
CARRERA	KARRERA	101	0.1%	100.000%

Figura 3-3.: Probabilidades de aparición para las variantes de "CARRERA"

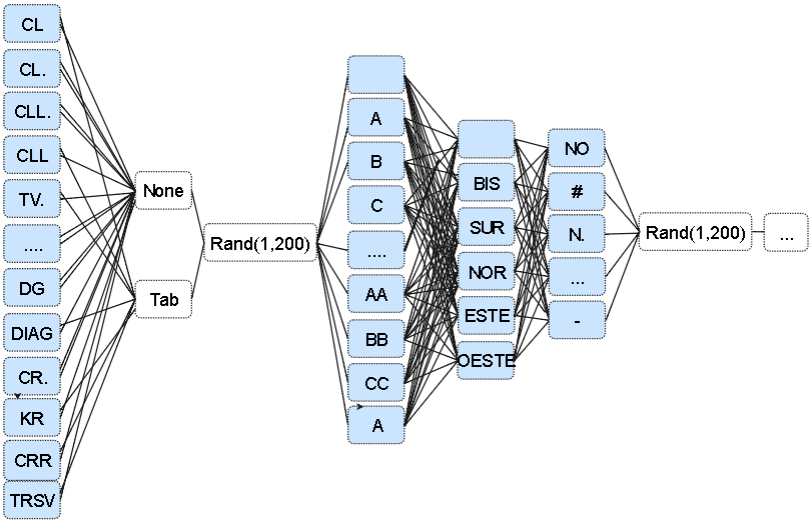

Figura 3-4.: Combinación aleatoria de elementos para general el set de datos

Dirección original	Dirección Limpia
CL80 112 30	CALLE 80 NO 112 - 30
DIAGONAL12 D BIS -30 70	DIAGONAL 12D BIS NO 30 - 70
CL58 AA KRRA. 192 BB 81	CALLE 58AA NO 192BB - 81
KR111 E # 154 89	CARRERA 111E NO 154 - 89
CRR.151 - 102E59	CARRERA 151 NO - 102E - 59
CLL 53D NUM. 156 B - 87	CALLE 53D NO 156B - 87
TV3B NORTE 50 C -73	TRANSVERSAL 3B NORTE NO 50C - 73
CLLE 128AA # 164 E81	CALLE 128AA NO 164E - 81
TV. 114 E BIS TRVL. 162 CC41	TRANSVERSAL 114E BIS NO 162CC - 41

Tabla 3-2.: Ejemplo de direcciones y su equivalente estandarizado

3.1.3. Generación aleatoria de direcciones

Una vez obtenida la secuencia de la dirección que se va a utilizar y las variantes a utilizar en cada uno de los elementos de dicha secuencia es posible generar una dirección geográfica en formatos similares a los presentes en las diferentes bases de datos. Puesto que las posiciones de los 15 elementos son fijas, es posible generar además el formato correcto para una determinada combinación asignando a cada uno de los dichos elementos la palabra estandar (Calle, Carrera, NO.,-, etc). . El proceso completo se ilustra en la figura figura **3-4**

El resultado de este proceso generara una tabla como la que se muestra en la tabla **3-2**, que servirá como base para entrenar cualquiera de los algoritmos utilizados.

4. Metodología Propuesta: RNN-LSTM

A diferencia de una red neuronal estándar, las redes neuronales recurrentes se presentan como redes con ciclos internos, también vistos como conexiones con estados anteriores de la misma red, que les permiten almacenar información de lo que ha sucedido en el pasado cercano [49]. Esta característica les ha permitido ser aplicadas con gran éxito en problemas que incluyan secuencias, listas, características asociadas al tiempo y modelos de lenguaje [44, 50, 51]. La figura **4-1** muestra la estructura básica de este tipo de redes.

Sin embargo, la estructura estándar de este tipo de redes sólo permite almacenar eventos por un corto periodo de tiempo. En determinados problemas, como lo puede ser el predecir la siguiente palabra dada una oración en un determinado contexto, es necesario considerar dependencias de largo plazo. En estos casos las redes neuronales recurrentes con estructura estándar han demostrado no ser tan efectivas por lo que se ha popularizado la aplicación de redes tipo LSTM (Long-Short Term Memory) cuya estructura varía ligeramente de la mostrada en la figura **4-1**, de tal manera que se permita almacenar información por un periodo más largo de tiempo. La estructura de estas redes se puede ver en la **4-2**, y su efectividad queda demostrada en aplicaciones como las mostradas en [42, 52, 40].

En el presente documento se detalla la implementación de una red neuronal con dos capas tipo LSTM para la codificación de una secuencia de caracteres asociada a una dirección en un formato no estandarizado, y su posterior decodificación en el formato estándar. A continuación, se describe en detalle la implementación de este tipo de red.

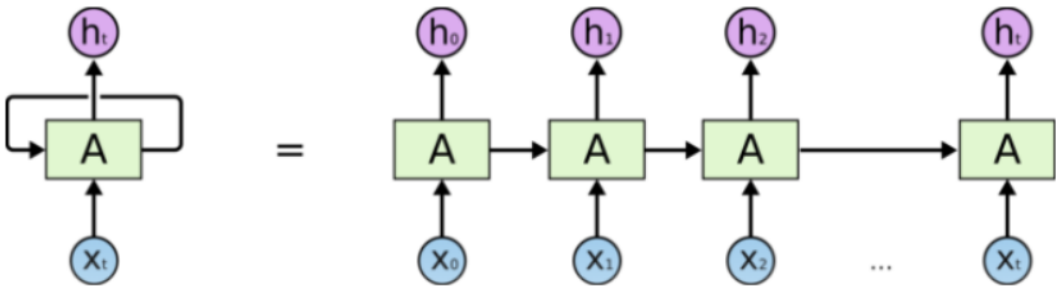

Figura 4-1.: Esquema Red Neuronal Recurrente

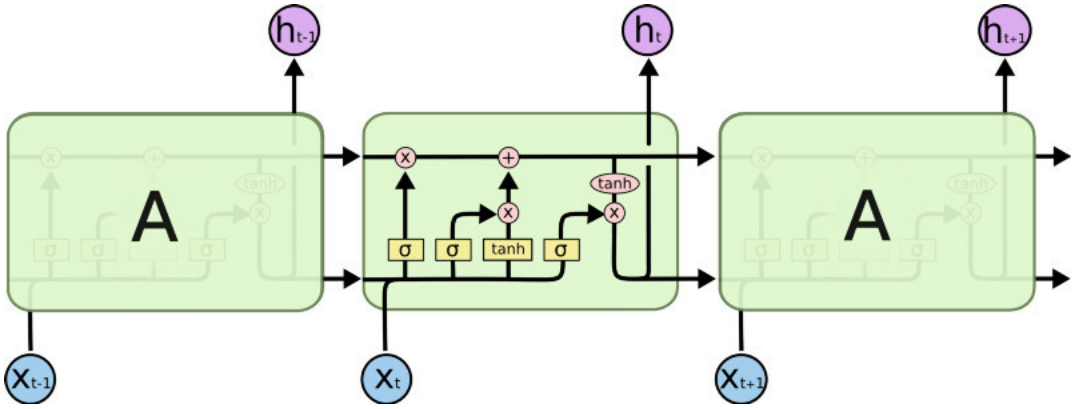

Figura 4-2.: Estructura interna LSTM

4.0.1. Modelo Propuesto

El modelo propuesto en el presente documento debe partir de una secuencia dada de caracteres $(x_1, x_2, x_{...}, x_n)$ (e.g."CL 63 No 130 - 15") y obtener una secuencia $(y_1, y_2, y_{...}, y_m)$ en el formato deseado ("CALLE 63 NO 130 - 15")
Nótese que n y m, que corresponden a los tamaños de las secuencias de entrada y de salida, pueden ser diferentes para un determinado texto.

Por tanto, lo que se quiere determinar la siguiente probabilidad.

$$P(y_1, y_2, \cdots, y_m | x_1, x_2, \cdots, x_n) \tag{4-1}$$

Para ello, se propone un modelo dividido en dos etapas o capas tipo LSTM. La primera sirve como codificador de los caracteres de entrada a un determinado formato s, mientras que la segunda sirve para generar un caracter de la secuencia de salida cada vez, a partir del formato codificado s, y los caracteres anteriores ya generados por esta capa. Matemáticamente se busca hallar la probabilidad condicional $P(y|x)$, que se puede ver como:

$$P(y|x) = \sum_{j=1}^{m} log P(y_j | y_{<j}, s) \tag{4-2}$$

En este caso, tal y como se propone en [40] utilizaremos la función softmax de tal modo que

$$P(y_j | y_{<j}, s) = softmax(g(h_j)) \tag{4-3}$$

Donde g es la función que decodifica el estado oculto h_j en un caracter determinado de la secuencia de salida. El modelo descrito anteriormente corresponde a un modelo de secuencia

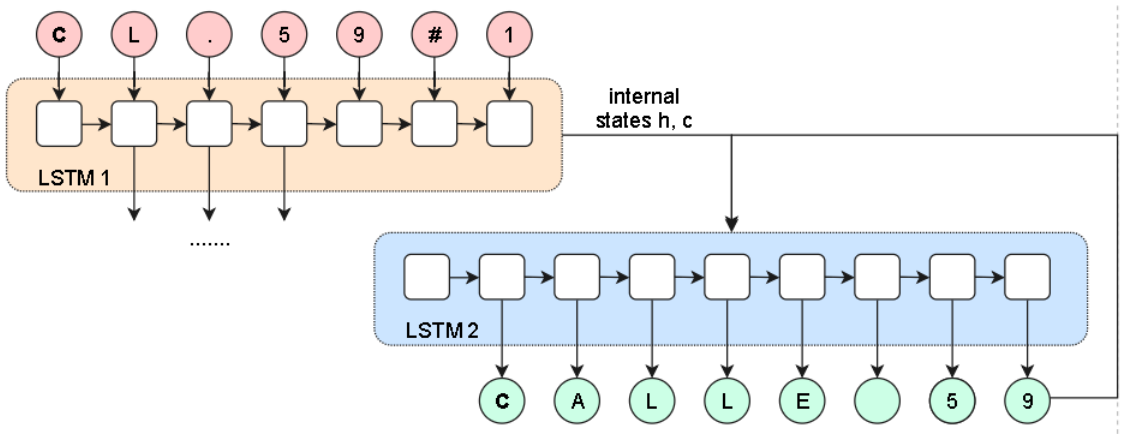

Figura 4-3.: Representación del modelo. Dos capaz LSTM para traducción caracter por caracter

a secuencia (varias entradas, varias salidas) en el que la predicción o transformación a la secuencia de salida se realiza caracter por caracter. La figura **4-3** muestra una representación de este modelo.
Si bien este modelo de dos capas genera más complejidad a la hora de realizar el entrenamiento del modelo y por lo tanto requiere mayor capacidad computacional, también hace que, en general, se obtengan mejores resultados para problemas de este tipo, como se especifica en [43]. Nótese adicionalmente que la segunda capa del modelo, encargada de la decodificación, únicamente considera el último estado oculto h, obtenido a partir de la primera capa y que brinda información sobre qué es lo que se supone que debe generar; así como también considera los caracteres que ya ha generado hasta ese momento del tiempo.

4.1. Definicion de las secuencias de entrada y salida

En esta sesión se describe el preprocesamiento y transformación realizado en las secuencias de entrada, correspondientes a direcciones postales, realizadas con el fin de facilitar el entrenamiento del modelo sin reducir la calidad del mismo. Dichas labores son necesarias para que el modelo sea escalable y reaplicable, independientemente de la base de datos de direcciones en la que se desee aplicar.

4.1.1. Preprocesamiento de las secuencias de entrada

Como se estableció en la definición del problema, del presente documento, la variabilidad intrínseca que le otorga el factor humano a la escritura de una dirección geográfica, hace que la cantidad de caracteres y combinaciones posibles en una secuencia de entrada sea muy alta [9, 28]. Con el fin de reducir dicha variabilidad y con ello, facilitar el entrenamiento, se

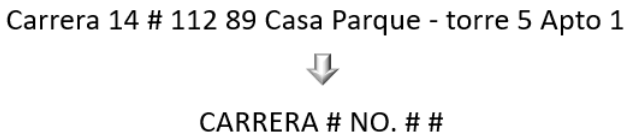

Figura 4-4.: resultado preprocesamiento en una secuencia de entrada "C"

propone realizar la siguiente secuencia de transformaciones, en orden, sobre el conjunto de direcciones.

- Remover acentos y caracteres especiales que no esten presentes en el formato estandar de direcciones para el lugar al que pertenezcan las direcciones. En el caso de Colombia esto incluye caracteres como: ("&" ,"°", "%", "~", "_" , "´", "^", etc.)
- Convertir toda la cadena de texto a letras mayúscalas
- Identificar y remover el complemento de la dirección que, en la mayoría de los casos, no aporta información adicional para los programas de geolocalización [2]. En el caso de Colombia, dicho complemento incluye normalmente información sobre apartamentos, locales, bloques, manzanas, nombre de barrio, comuna o lugar, interiores o pisos.
- Reemplazar el símbolo '' por las letras 'NO.'. Esto con el fin de garantizar que la dirección pueda ser procesada posteriormente en cualquiera de las APIs de geolocalización [53, 54]
- Identificar las posiciones de los números presentes en la dirección, en orden, y almacenar dichos números.
- Reemplazar todos los números presentes en la dirección por el símbolo "#". De esta manera, el modelo no necesitará codificar y decodificar todas las posibles combinaciones de dígitos presentes en una dirección, sino únicamente identificar la posición correcta de dichos números y reemplazarlos por los almacenados en el literal anterior, después del proceso de decodificación.

La figura **4-4** muestra un ejemplo de una secuencia original y su equivalente preprocesada y lista para entrar al modelo.

4.1.2. Representación de las secuencias de entrada y salida como vectores one-hot

Para la representación de las secuencias, tanto de entrada como de salida, se hará uso de una codificación de tipo One - hot en el vector formado por los diferentes caracteres que conforman los diferentes formatos de direcciones identificados en un set de datos inicial

0	#	-	.	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
G	I	K	L	M	N	O	R	S	T	U	V	I	m	n	u	o	\n	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figura 4-5.: One hot representacion Caracter "C"

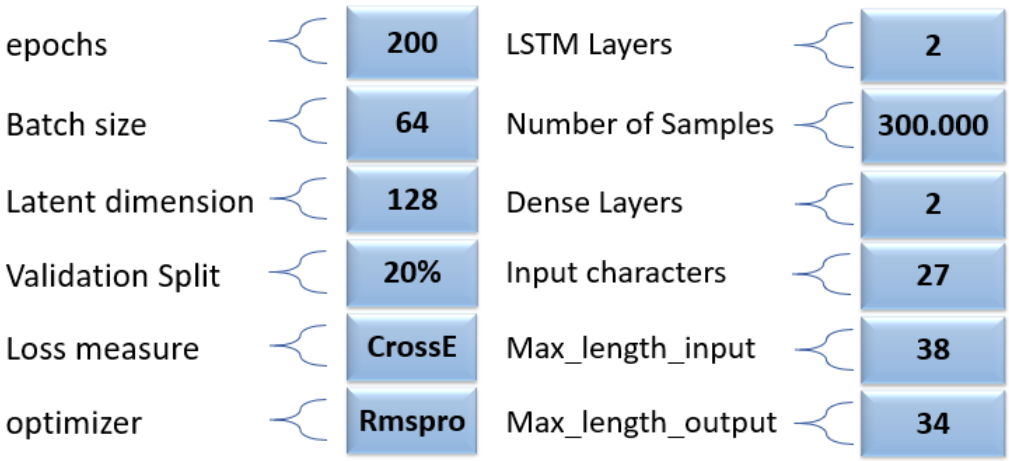

Figura 4-6.: Set de parámetros entrenamiento

de direcciones de tiendas. De esta manera cada caracter de la secuencia de entrada estará representado por un vector de longitud 27 (correspondiente al número de caracteres diferentes identificados) con un '1' en el espacio asignado al caracter que se está representando y "0.[en] el espacio asignado a los demás caracteres.
La figura **4-5** muestra dicha representación de 0s y 1s para el caracter 'C'.
Así, una secuencia correspondiente a una dirección en particular estará dada por un conjunto de vectores one - hot, uno correspondiente a cada caracter. La secuencia de Salida se representa de manera similar. La única diferencia radica en que, en este caso, el vector está dado únicamente por 26 posibles caracteres (puesto que la secuencia transformada puede excluir determinados elementos de la secuencia inicial antes de ser estandarizada).

4.2. Implementación y entrenamiento

4.2.1. 2 layer Encoder-Decoder RNN

La implementación y entrenamiento del modelo propuesto se realiza en Python utilizando las funciones predefinidas de la librería Keras. La figura **4-6**, se describe el set de parámetros utilizados para este entrenamiento.
Para el entrenamiento de ambas redes se utilizó La GPU virtual TESLA K80 (61GB RAM)

proveída por el servicio de Floydhub. En el caso de la primera red se utilizó un tamaño de lote de 64, según lo recomendado por Lewis et. al. en [43]. Para el entrenamiento se seleccionó un parámetro de Split del 10 % teniendo en cuenta que contamos con un set de datos mucho mayor a los usados por otros estudios que utilizan la mayoría de sus datos en el entrenamiento. De este modo el entrenamiento se realizará con 180000 muestras y se validará en 20000 aplicando el método de cross-entropy. Por otro lado la dimensión de la capa latente se fija en 128, número que se esperaría sea similar a la cantidad de caracteres de salida si se considerara un set de datos no sintético incluyendo acentos, letras mayúsculas, minúsculas y algunos otros signos, de acuerdo con la recomendación presente en [40]. Esto también habilita que el modelo sea escalable a aplicaciones más complejas[40]. El número epochs utilizadas en el entrenamiento final del modelo es de 200. Sin embargo, es de notar que tras haber entrenado la red completa, el valor de accuraccy y training-loss converge a partir de alrededor de 80 epochs. Todos los resultados mencionados en la sesión de resultados y evaluación se obtienen con los pesos obtenidos al finalizar dichas 200 epochs.
La figura **4-7** muestra el esquema final del modelo encoder-decoder propuesto para la limpieza de direcciones. Como pude verse, este modelo está compuesto por dos capas de entrada: la primera, que tiene como número de parámetros la cantidad de posibles caracteres que aparecen en las secuencias de entrada (27); y una segunda con 26 parámetros correspondientes a los caracteres que se espera obtener luego de la decodificación y que entra precisamente a dicha capa de decodificación. Adicionalmente, se agregan dos capas densas al final de la red como pasos de decodificación adicionales para obtener la dirección postal en formato estandar.

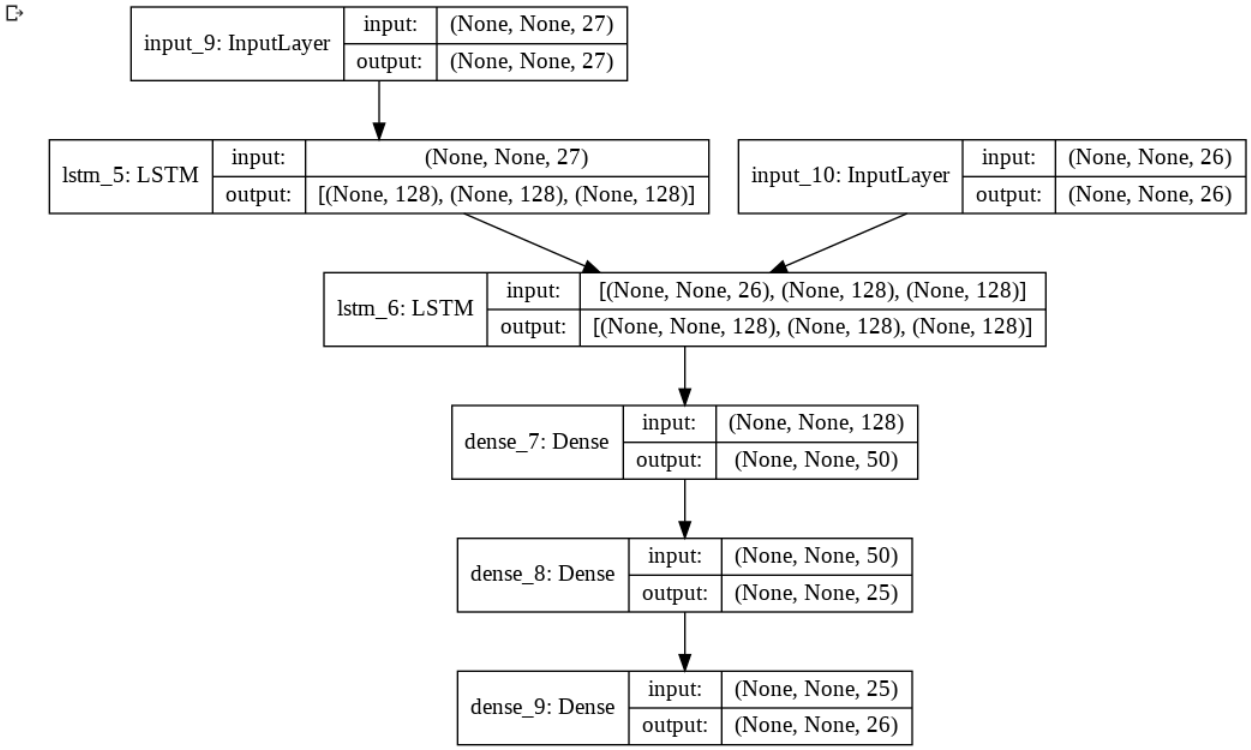

Figura 4-7.: Representación gráfica del modelo encoder-decoder

5. Evaluación Experimental

5.1. Medidas de desempeño

En el presente estudio se proponen dos tipos de medidas de desempeño, cada una enfocada en un aspecto diferente del proceso de geolocalización.

5.1.1. Medidas para la capacidad de estandarización del modelo

El primer set de medidas busca evaluar la capacidad del modelo para llevar la dirección a un formato estandar, tal y como lo establece el objetivo principal del presente trabajo. Para este fin se utilizarán varias de las medidas de similitud de texto, ampliamente utilizadas para la evaluación de sistemas de procesamiento de lenguaje natural [55], [56]: como lo son la distancia de Levehstein, La similitud de Jaro y La distancia de Jaccard, que se describen a continuación:

Algoritmo de distancia de Levenshtein

La distancia de Levenshtein, propuesta por el científico ruso Vladimir Levenshtein en [**?**], es una de las medidas más utilizadas para la comparación de cadenas de texto [57]. Su calculo se basa en determinar el mínimo número de transformaciones que se deben realizar para obtener una cadena de caracteres, a partir de otra. Se entiende como transformación una inserción de un determinado caracter, una eliminación, o una sustitución de un caracter por otro.
De esta manera, la distancia de Levenshtein está dada por la siguiente formula:

$$lev_{a,b}(i,j) = \text{mín} \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \tag{5-1}$$

donde $1_{(a_i \neq b_j)}$ es la funcion indicador, igual a 0 cuando $a_i = b_j$ e igual a 1 en caso contrario. $\text{lev}_{a,b}(i,j)$ es la distancia entre los primeros i caracteres de a y los primeros j caracteres de b.

En esta ecuación, el primer elemento corresponde a la eliminación de un caracter, el segundo elemento corresponde a la inserción y el tercero a la sustitución.

Medida de Similitud de Jaro

Como medida de la similitud entre las diferentes cadenas de texto, se propone utilizar la métrica de Jaro, ampliamente utilizada en el ámbito académico y que ha mostrado buenos resultados en el área de enlace de registros [55]. A diferencia de la distancia de Levenshtein, la métrica de Jaro no está basada en las distancias de edición sino en el número y orden de los caracteres comunes en ambas cadenas de texto. Así la medida de similutd de jaro entre dos cadenas $S1$ y $S2$ estará dada por:

$$Jaro(s1, s2) = \frac{1}{3}\left(\frac{m}{\mid s1 \mid} + \frac{m}{\mid s2 \mid} + \frac{m-t}{m}\right) \tag{5-2}$$

En donde:

- $\mid s1 \mid$ y $\mid s2 \mid$ son la longitud respectiva de cada una de las cadenas de texto.
- m es el número de caracteres equivalentes entre las dos cadenas de texto. LA equivalencia se da si son el mismo caracter están ubicadas en cada cadena de texto a una distancia no mayor de la siguiente expresión.

  $$H = \left\lfloor \frac{max(\mid s1 \mid, \mid s2 \mid)}{2} \right\rfloor - 1 \tag{5-3}$$

- t es el número de transposiciones entendidas como aquellas posiciones en los que los caracteres equivalentes, no estan en la misma posición dentro de las cadenas de texto formadas por todos los caracteres equivalentes de una palabra en la otra.

Coeficiente de Similitud de Jaccard

Como se estableció en la definición del problema, del presente documento, es posible entender una dirección geográfica como un conjunto de palabras o términos, cada una de ellas asociada a un componente específico de la dirección. Dichos componentes fueron listados en la tabla 3-1.
Es por esta razón que se propone como medida de similitud entre la dirección arrojada por un modelo y su correspondiente estandar conocida, el coeficiente de Jaccard [58], que mide qué tan similares son un par de conjuntos, independientemente del tipo de elementos que los conforman. En este caso, el conjunto corresponde a la secuencia completa de caracteres y los elementos corresponden a las palabras que conforman dicha secuencia. Puesto que el objetivo de los modelos es estandarizar la dirección se espera que ambas secuencias estén conformadas por las mismas palabras y en exactamente los mismos formatos.

Así, el índice de Jaccard se calcula al dividir la cardinalidad de la intersección de ambos conjuntos entre la cardinalidad de la unión, como se muestra a continuación:

$$J(A,B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} = \frac{\mid A \cap B \mid}{\mid A \mid + \mid B \mid - \mid A \cap B \mid} \tag{5-4}$$

Adicionalmente, para transformar el índice de Jaccard en una medida de distancia, tomamos su complemento, de manera que entre más cercano a cero (0), sea el resultado, menor será la distancia entre dos conjuntos (serán más parecidos entre sí. Así:

$$D(A,B) = 1 - J(A,B) \tag{5-5}$$

5.1.2. Precision de la geolocalización obtenida

La segunda medida busca identificar el efecto que tiene la estandarización, en la geolocalización de dichas direcciones a partir de APIs especializadas (ej. Googlemaps, open street maps, ArcGIS, etc.). Como lo establecen Dilek y Ugur en [10], la precisión de estos programas varía según el formato y la calidad de las direcciones de entrada, por lo que a través de este método aplicado, se busca medir la precisión de la ubicación (latitud, longitud) obtenida a partir de una dirección estandar versus la dirección original, para cada uno de los diferentes métodos de estandarización descritos en la siguiente sección.
En este caso, se tomarán como medidas de comparación, los estadísticos base tales como **la media, la mediana, la moda, los percentiles y las frecuencias en determinados rangos**, aplicados sobre la distancia entre el punto obtenido por la geolocalización y la ubicación real de dicho punto.

5.2. Modelos de línea base

5.2.1. Modelo basado en reglas de estandarización

Los modelos basados en reglas de elementización y estandarización son ampliamente utilizados en la industria, para labores relacionadas con limpieza de información estructurada como la que implica la limpieza de direcciones postales [12]. Dichos modelos se basan normalmente en dividir la dirección en cada uno de los diferentes elementos y luego transformar cada elemento en su forma canónica o estandar a través de diccionarios.

Este tipo de modelos ha demostrado tener muy buenos resultados en países y zonas geográficas con un esquema definido para la caracterización de direcciones como el sistema postal de Estados Unidos [9]. Sin embargo, cuando no hay un estandar, la cantidad de posibles variaciones y errores de codificación es muy alta para intentar cubrirla a través de diccionarios[9][12].

Con el fin de comprobar el desempeño de este tipo de modelos, se hace uso del modelo actualmente utilizado en una de las mayores empresas productoras de productos de consumo masivo en el país. Este modelo está actualmente implementado en KNIME Analytics Platform [59], a modo de ETL encargado de eliminar caracteres especiales, acentos y dobles espacios, llevar todo a letras mayúsculas y aplicar un total de 168 reglas, tipo diccionario, divididas así:

- 30 variaciones para el elemento “Calle”
- 30 variaciones del elemento “Carrera”
- 6 variaciones del elemento “Diagonal”
- 35 variaciones del elemento “Transversal”
- 18 variaciones del elemento “Avenida”
- 35 variaciones del elemento “Transversal”
- 27 variaciones para el separador de la segunda avenida “NO.”
- +20 Variaciones para los posibles complementos de la dirección (Barrio, Manzana, Local, Esquina, etc.).

Las 130 variaciones consideradas en el modelo base se listan en el Anexo A. Por su parte, la figura **5-1** muestra algunas de las reglas aplicadas sobre la palabra “CALLE”. Nótese que en este modelo base no se consideran transformaciones sobre los elementos numéricos de la dirección, pues estos se dejan en las mismas posiciones y formatos en los que se encuentren en la secuencia original.

5.2.2. Modelo oculto de Markov

Un modelo oculto de Markov (Hidden Markov Model - HMM) es un modelo estadlístico en el que se asume que el sistema que se modela corresponde a un proceso de markov con estados no observados direcatemente. El objetivo es estimar los parámetros desconocidos (de allí el nombre de ocultos) de una cadena de texto, a partir de las evidencias que se pueden observar directamente. Los parlámetros extralídos pueden ser entonces utilizados para anlálisis posteriores (ej. reconocimiento de patrones). El modelo oculto de markov puede ser considerado como la la manera más simple de una red dinámica Bayesiana. [60]
Como se muestra en la seccilón 2.2.2., los modelos ocultos de Markov han sido utilizados en diversos estudios para la estandarizacilón de direcciones postales [3, 7, 18]. Para el presente estudio, se utilizará el ”Naive markov model”base, descrito en [3]. Este Modelo oculto de markov es capaz de capturar las diferentes relaciones entre los elementos de la dirección como

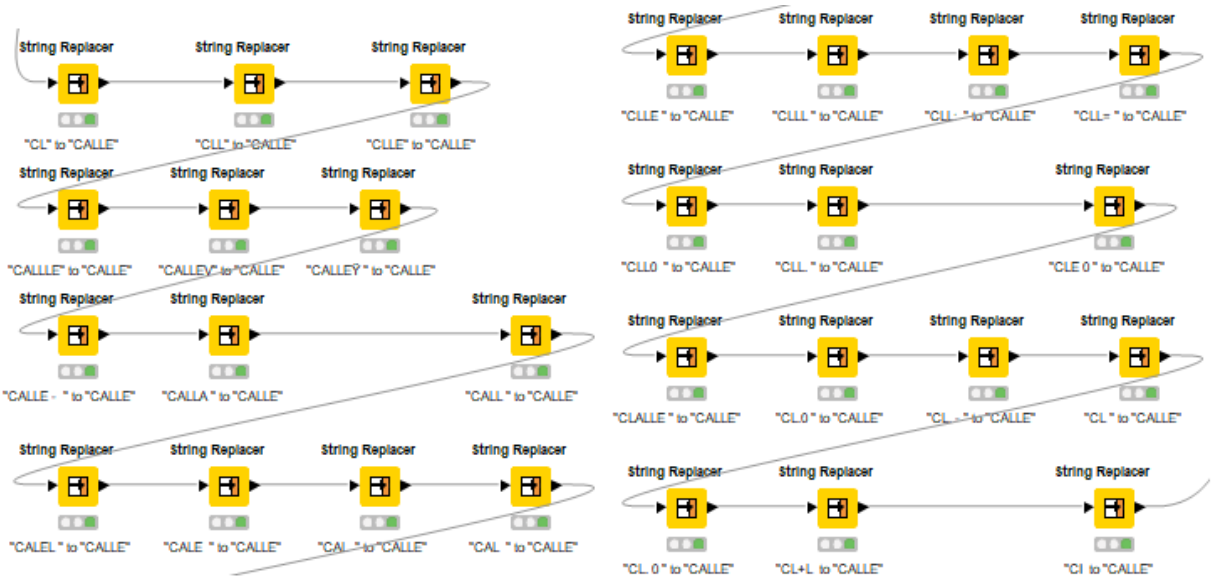

Figura 5-1.: Ejemplo de diccionario para "Calle"

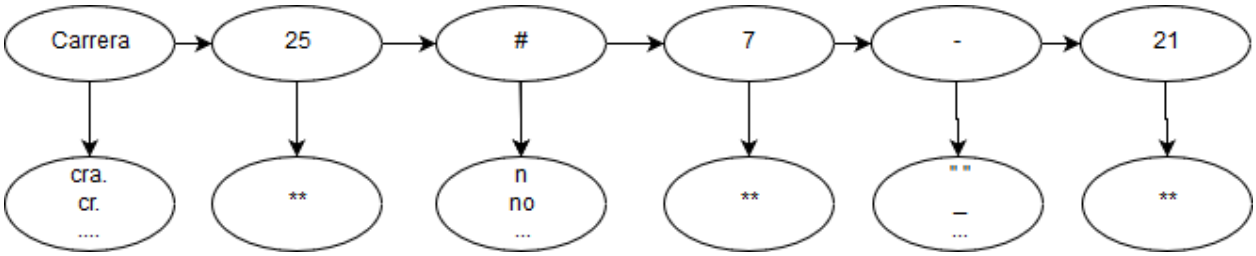

Figura 5-2.: Evidencias y Elementos en un modelo oculto de markov para una dirección geográfica

se muestra en la figura **5-2** que es una version simplificada del esquema de markov adaptado al proceso de segmentación de direcciones. En esta figura se puede notar la relacion entre la secuencia de terminos y posibles evidencias para cada uno de los campos. Dicha relacion es precisamente la que se busca identificar con este tipo de modelos.

En este documento, se utilizará como modelo base, el propuesto por Sharma et. al. en [3]. Para ello, el set de entrenamiento se segmenta manualmente en unidades atómicas, cada una de ellas correspondiente a un elemento de la dirección. A cada unidad se le asigna la clasificación correcta, correspondiente a alguno de los elmentos expresados en la tabla 2_1. Lo anterior también permite identificar las posibles variaciones que hay entre la notación de cada uno de los estados (evidencias), y con ello calcular la probabilidad de que se muestre la evidencia, dado un determinado estado, utilizando el método de suavización de Laplace, como se muestra a continuación.

```
['1', 'calle'],   ['11', '*a'],
['1', 'casa'],    ['11', 'a'],
['1', 'cl'],      ['11', 'b'],
['1', 'cl.'],     ['11', 'bb'],
['1', 'cll'],     ['11', 'centro'],
['1', 'clle'],    ['11', 'e'],
['1', 'kil'],     ['11', 'w'],
```

Figura 5-3.: Ejemplo de evidencia para los estados '1' y '11'

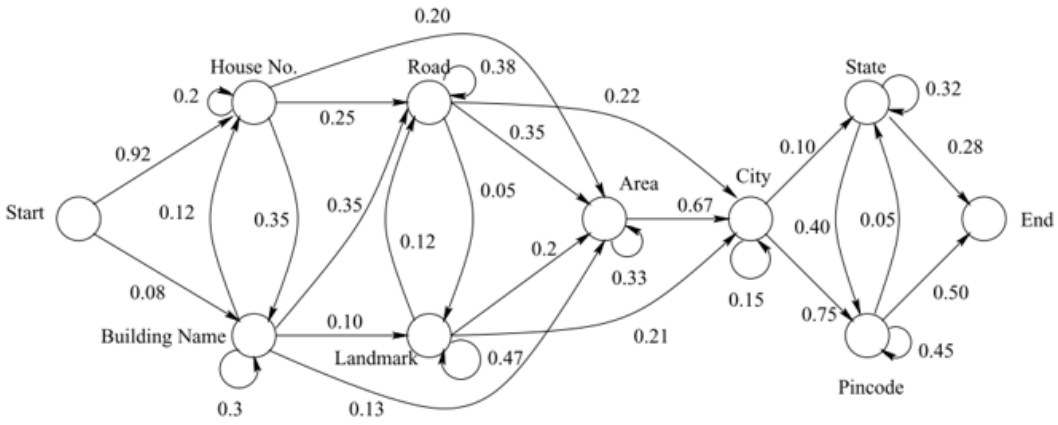

Figura 5-4.: Ejemplo cadena de Markov en segmentación de direcciones en Estados Unidos

$$P(evidence \mid state) = \frac{count(evidence, state) + k)}{count(state) + k \mid E \mid} \tag{5-6}$$

La figura **5-3** muestra algunos de las evidencias que se presentan para los estados '1' (Calle) y '11' (Complemento avenida), tomados directamente de la implementación del algoritmo.
La figura **5-4** muestra la estructura típica de un modelo oculto de markov, base el contexto de segmentación de direcciones utilizando direcciones de Estados Unidos. Para facilitar la interpretacián de la figura, sólo los estados más importantes que conforman una dirección, se muestran en la gráfica.
Así, el modelo escogido como modelo base y descrito está compuesto por los siguientes elementos:

- Un conjunto de n estados, cada uno correspondiente a un elemento de la dirección.

- Un diccionario de posibles salidas (evidencias) para cada uno de los estados.
- Una matriz de transición de probabilidad, de tamaño nxn, indicando la probabilidad de realizar una transición desde el estado i, al estado j.
- Un diccionario de n elementos de salida que corresponden a las probabilidades de que se presente una evidencia determinada desde cada uno de los n estados.

5.3. Set up experimental

A continuación, se describe la metodología utilizada para la comparación del modelo propuesto versus los modelos de línea base, y la correspondiente evaluación de resultados.

5.3.1. Selección del set de validación para medir la estandarización

De acuerdo con el objetivo principal del presente trabajo, se requiere medir la capacidad de estandarización. Para ello, es necesario comparar la salida de cada uno de los modelos al procesar una dirección en formato no estandar con la versión estandar de dicha dirección. Sin embargo, como se estableció en la definición del problema (Ver Capítulo 1), no es común encontrar direcciones ya estandarizadas en la industria y, en caso de que las haya, es poco probable que tengan un equivalente sucio que requiera ser estandarizado para esa misma dirección.

Por esta razón, se propone generar un conjunto de validación sintético, diferente al utilizado para el entrenamiento del modelo, que permita obtener variaciones desde una dirección ya estandarizada. Dicho set de entrenamiento se puede generar utilizando una versión modificada del modelo descrito en la sección 3, que obtiene versiones aleatorias de cada uno de los elementos de la dirección, incluyendo diferentes tipos de formatos. El archivo utilizado para generar dichas direcciones aleatorias se puede encontrar en el Anexo B adjunto a este documento. Con dicha metodología, se genero un total de 10000 direcciones aleatorias y su equivalente estandar contra el cual se compararán los resultados. El Anexo B del presente documento, muestra las primeras 50 direcciones autogeneradas, mientras que las primeras 15 se muestran en la tabla **5-1** En la documentación adjunta como soporte del trabajo está disponible también el total de direcciones generadas, que pueden ser utilizadas con fines de comparación en trabajos futuros.

En la tabla **5-1** la segunda columna corresponde a la dirección a limpiar y la tercera equivale a la cadena de texto contra la cual se va a medir la similitud del resultado de cada modelo, utilizando la distancia de Jaccard, la distancia de Levenshtein y la medida de similitud de Jaro descritas en el numeral 5-1-1.

	Original_Addresses	Target_Addresses
0	KR 115E SUR NO. 105DD23	CARRERA 115E SUR NO 105DD - 23
1	TRV161E DIAG 52 D -53	TRANSVERSAL 161E NO 52D - 53
2	CR 137C 31E10	CARRERA 137C NO 31E - 10
3	CALL 45 CC # 160CC17	CALLE 45CC NO 160CC - 17
4	CR169BB 150 24	CARRERA 169BB NO 150 - 24
5	TV. 40DD SUR TRVL. 97 75	TRANSVERSAL 40DD SUR NO 97 - 75
6	DIAG.186 BIS 119 B10	DIAGONAL 186 BIS NO 119B - 10
7	KR2AA NORTE #175D64	CARRERA 2AA NORTE NO 175D - 64
8	CALLE152 SUR KR.140 B42	CALLE 152 SUR NO 140B - 42
9	AVENIDA65 SUR N. 1C10	AVENIDA 65 SUR NO 1C - 10
10	CR 190 OESTE - 23 60	CARRERA 190 OESTE NO 23 - 60
11	CL 179D SUR NO 23 B35	CALLE 179D SUR NO 23B - 35
12	AV 25D - 171 A89	AVENIDA 25D NO 171A - 89
13	TRANS 154 D TRVL. 110 53	TRANSVERSAL 154D NO 110 - 53
14	CR155 E # 154 B13	CARRERA 155E NO 154B - 13

Tabla 5-1.: Primeras 15 Direcciones set de validación

5.3.2. Validación de resultados en base de datos real

Si bien la aproximación sugerida en el numeral anterior permite determinar la precisión de los modelos en un set de datos autogenerado, que replica varios de los errores comunes que se dan en la limpieza de direcciones, no es posible incluir en dicho modelo de autogeneración todas las posibles variaciones tipográficas, ortográficas o de formato que pueden haber en un set de datos real. Por lo que se considera relevante diseñar una metodología que permita medir el desempeño de los modelos base y el modelo propuesto sobre un conjunto de direcciones presentes en una base de datos real.

El desafío es, entonces, cómo medir la precisión de la dirección resultante si no se cuenta con la dirección limpia (pues este es justamente el problema que se busca solucionar).

Para lograrlo, se seleccionó un total de 5000 establecimientos comerciales en Bogotá, Medell´in y Cali que hacen parte de la base de datos de una importante empresa de consumo masivo en Colombia y para las que ya se contaba con una dirección sin estadarizar y la geolocalización (latitud, longitud). El hecho de tener las coordenadas geográficas de dichos establecimientos comerciales, permite tener un punto adicional de información que se puede utilizar para comparar contra el obtenido a partir de geolocalizar la dirección resultante de cada uno de los modelos cuando se aplican sobre la cadena de texto sin estandarizar.

Una vez definido el set de datos, se procede a geolocalizar dichas direcciones a través del API de Google Maps y se procede a aplicar las medidas de precisión descritas en el numeral 5-1-2 del presente documento.

	Original_Addresses	Target_Addresses	Rules_Addresses	Markov_Addresses	LSTM_Addresses
0	KR 115E SUR NO. 105DD23	CARRERA 115E SUR NO 105DD - 23	CARRERA 115E SUR NO 105DD23	CARRERA 115 E SUR NO. 105 DD 23	CARRERA 115E SUR NO 105DD - 23
1	TRV161E DIAG 52 D -53	TRANSVERSAL 161E NO 52D - 53	TRANSVERSAL 161E DIAGONAL 52 D - 53	TRANSVERSAL 161 E DIAGONAL 52 D - 53	TRANSVERSAL 161E NO 52D - 53
2	CR 137C 31E10	CARRERA 137C NO 31E - 10	CARRERA 137C 31E10	CARRERA 137 C 31 E 10	CARRERA 137C NO 31E - 10
3	CALL 45 CC # 160CC17	CALLE 45CC NO 160CC - 17	CALLE 45 CC NO 160CC17	CALLE 45 CC NO. 160 CC 17	CALLE 45CC NO 160CC - 17
4	CR169BB 150 24	CARRERA 169BB NO 150 - 24	CARRERA 169BB 150 24	CARRERA 169 BB 150 24	CARRERA 169BB NO 150 - 24
5	TV. 40DD SUR TRVL. 97 75	TRANSVERSAL 40DD SUR NO 97 - 75	TRANSVERSAL 40DD SUR TRVL 97 75	TRANSVERSAL 40 DD SUR TRANSVERSAL 97 75	TRANSVERSAL 40DD SUR NO 97 - 75
6	DIAG.186 BIS 119 B10	DIAGONAL 186 BIS NO 119B - 10	DIAGONAL 186 BIS 119 B10	DIAGONAL 186 BIS 119 B 10	DIAGONAL 186 BIS NO 119B - 10
7	KR2AA NORTE #175D64	CARRERA 2AA NORTE NO 175D - 64	CARRERA 2AA NORTE NO 175D64	CARRERA 2 AA NORTE NO. 175 D 64	CARRERA 2AA NORTE NO 175D - 64
8	CALLE152 SUR KR.140 B42	CALLE 152 SUR NO 140B - 42	CALLE 152 SUR CARRERA 140 B42	CALLE 152 SUR CARRERA 140 B 42	CALLE 152 SUR NO 140B - 42
9	AVENIDA65 SUR N. 1C10	AVENIDA 65 SUR NO 1C - 10	AVENIDA 65 SUR N 1C10	AVENIDA 65 SUR NO. 1 C 10	AVENIDA 65 SUR NO 1C - 10
10	CR 190 OESTE - 23 60	CARRERA 190 OESTE NO 23 - 60	CARRERA 190 OESTE - 23 60	CARRERA 190 OESTE NO. 23 60	CARRERA 190 OESTE NO 23 - 60
11	CL 179D SUR NO 23 B35	CALLE 179D SUR NO 23B - 35	CALLE 179D SUR NO 23 B35	CALLE 179 D SUR NO. 23 B 35	CALLE 179D SUR NO 23B - 35
12	AV 25D - 171 A89	AVENIDA 25D NO 171A - 89	AVENIDA 25D - 171 A89	AVENIDA 25 D NO. 171 A 89	AVENIDA 25D NO 171A - 89
13	TRANS 154 D TRVL. 110 53	TRANSVERSAL 154D NO 110 - 53	TRANSVERSAL 154 D TRVL 110 53	TRANSVERSAL 154 D TRANSVERSAL 110 53	TRANSVERSAL 154D NO 110 - 53
14	CR155 E # 154 B13	CARRERA 155E NO 154B - 13	CARRERA 155 E NO 154 B13	CARRERA 155 E NO. 154 B 13	CARRERA 155E NO 154B - 13

Figura 5-5.: Ejemplos resultados direcciones

Original_Addresses	Target_Addresses	Rules_Addresses	Markov_Addresses	LSTM_Addresses
KR 115E SUR NO. 105DD23	CARRERA 115E SUR NO 105DD - 23	CARRERA 115E SUR NO 105DD23	CARRERA 115 E SUR NO 105 DD 23	CARRERA 115E SUR NO 105DD - 23

Figura 5-6.: Primera línea resultados

5.4. Resultados y Discusión

5.4.1. Capacidad de estandarización del modelo

Tras aplicar la metodología propuesta en el numeral 5-3-1 para las direcciones del Anexo B la información se consolida en un conjunto de datos único, cuyas primeras 15 filas se muestran el la figura **5-5**. estas primeras líneas permiten realizar una validación visual de los resultados obtenidos de cada uno de los métodos y una comparación con la columna "Original Address", considerada como el formato estandar al que se desea llegar. Nótese, por ejemplo, que si se revisa únicamente la primera dirección de la tabla (Figura **5-6** se puede ver como el modelo de reglas no logra separar los elementos de la última parte de la dirección, mientras que el modelo de markov separa todos los elementos sin utilizar el "-çomo separador y sin identificar que "105DD"son un mismo elemento, que en este caso representa la calle. Por otro lado, el modelo LSTM da como resultado la secuencia exacta que se espera como Direccin Objetivo.

A continuación se evaluan los resultados cuantitativos tras calcular las medidas propuestas en la sección 5-1-1. El conjunto de datos utilizado para consolidar dichas medidas se encuentra en el ANEXO B.

En la tabla **5-2** se muestra el promedio y la desviación estandar para cada una de las medidas, obtenidas al resumir los resultados de comparar cada una de las 10000 direcciones con su correspondiente estandar. Nótese que se incluye además el radio de Levenshtein, obtenido al dividir la distancia de levenshtein sobre la longitud de la cadena más larga de texto de

Values	Rules	Markov	LSTM
Prom. Distancia de Levenshtein	6.14560	5.70720	0.00170
Prom. Radio de Levenshtein	0.86017	0.90025	0.99996
Prom. Similitud de Jaro	0.89244	0.87653	0.99998
Prom. índice de Jaccard	0.12709	0.10898	0.00005
StdDev. Distancia de Levenshtein	3.28954	3.46518	0.04793
StdDev. Radio de Levenshtein	0.07136	0.05037	0.00109
StdDev. Similitud de Jaro	0.05270	0.06992	0.00072
StdDev. índice de Jaccard	0.08270	0.07742	0.00186

Tabla 5-2.: Resultados estandarización

entre las que se están comparando. Dicho radio da un valor entre 0 y 1 y por lo tanto es comparable con el índice de similitud de Jaro.

De esta tabla puede verse que, en promedio, para pasar de la dirección procesada que genera el método base construido con reglas, es de 6.14 caracteres. Es decir que se necesitaría hacer en promedio 6.14 transformaciones, entre sustituciones, exclusiones e inclusiones para llegar del formato procesado por el modelo de reglas al formato estandar esperado. Mientras tanto, el modelo de markov mejora dicha medida tras necesitar, en promedio, 5.7 de estas transformaciones para llegar al formato estandar, lo anterior también se evidencia al observar el radio de Leveshtein que para el caso de Markov es de 0.9 versys 0.86 obtenido por el modelo basado en reglas. Dicho radio al ser superior a 0.85 indica una buena similitud de estandarización del modelo. Sin embargo, vale la pena resaltar que el modelo de markov también presenta una desviación estandar mayor a la del modelo base, resultado que se analizará más adelante a partir de las figuras **5-8** y **5-7**.

Nótese además que el modelo basado en redes neuronales tipo LSTM tiene una distancia de Levenshtein cercana a 0 y un radio de Levenshtein ≈ 1.

Esto indica que hay que hacer muy pocas transformaciones (ninguna en la mayoría de los casos) para llegar al formato estandar deseado. Lo anterior, da cuenta de la capacidad del modelo, diseñado especificamente para estandarizar, más allá de limpiar como es el caso de los modelos base.

La figura **5-7** muestra la comparación de los histogramas generados para la distancia de levenhstein en cada uno de los modelos, lo cual permite complementar el análisis realizado desde la tabla **5-2**.

En ella se puede evidenciar que la distribución de los resultados obtenidos por el modelo de markov es más sesgada a la izquierda que la distribución obtenida con el modelo basado en reglas que tiene un mayor número de incidencias en 3 y 4 transformaciones necesarias. Adicionalmente, puede verse que la máxima distancia de Levenshtein para el modelo de Markov es 16, mientras que la máxima con el modelo de reglas es de 20, lo que tambiém da cuenta de la eficacia del modelo de Markov sobre el modelo de reglas básico. Nótese además

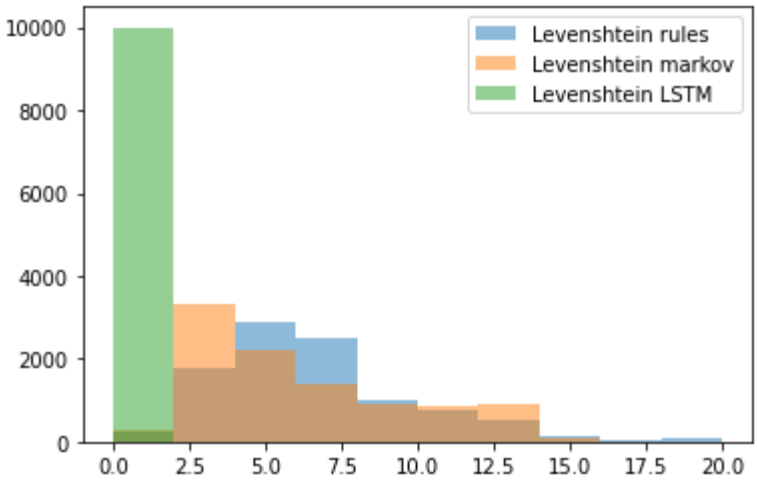

Figura 5-7.: Comparación de histogramas para la distancia de Levenshtein

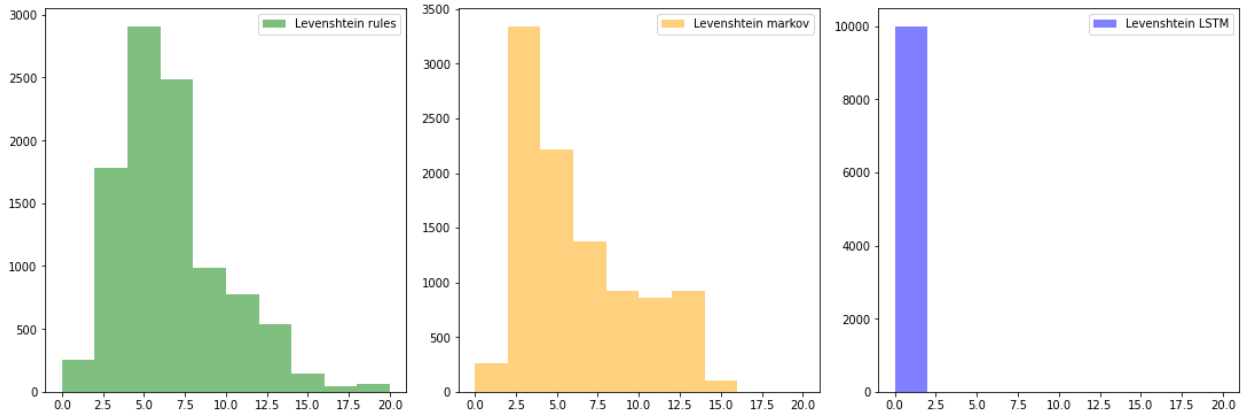

Figura 5-8.: Histogramas por modelos. Comparación para la distancia de Levenshtein

que la totalidad de las direcciones generadas por el modelo de LSTM se puede transformar en la dirección objetivo con menos de 2 cambios de caracteres, un valor muy inferior al de los otros dos modelos.

La figura **5-8** muestra el detalle de estos tres histogramas separado para una mejor visualización de los resultados. En esta puede verse más claramente el pico que hay en distancias de entre 2 y 3 transformaciones de diferencia. El análisis de las direcciones individuales que están a determinada distancia permite además entender mejor las falencias de cada modelo. Es así como, al filtrar las direcciones con distancia de levenshtein igual a 2 se obtienen casos como el que se muestra en la tabla **5-3**

En este caso, la diferencia corresponde a la inclusión de un "ƨ un espacio (" ") lo cual índica que el modelo de markov ve el separador entre el numero del segundo elemento de la dirección (Calle, carrera, diagonal o transversal) y el número de la casa como opcional y no

Target_Address	Markov_Addresses
CARRERA 115 ESTE NO 20 - 55	CARRERA 115 ESTE NO 20 55

Tabla 5-3.: table: Ejemplo distancia de Levenshtein = 2 en el modelo de Markov

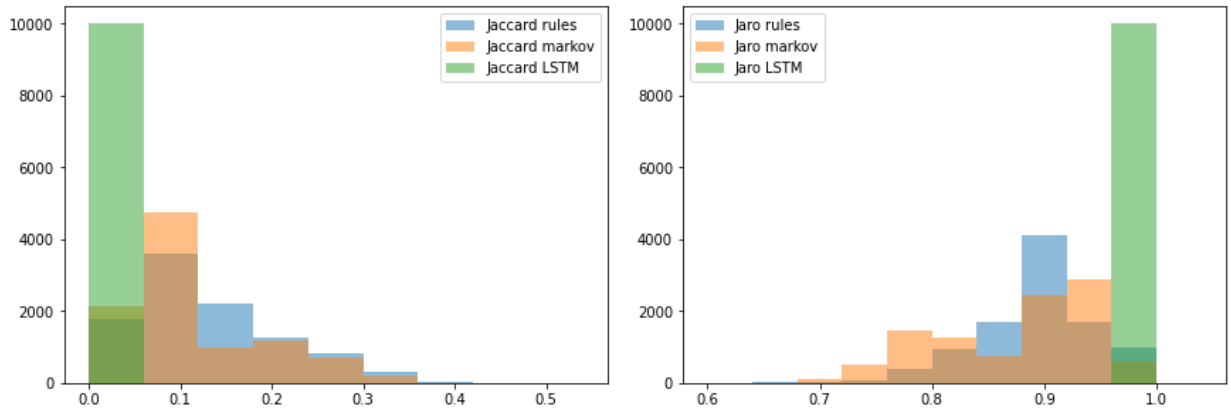

Figura 5-9.: Distribución del índice de Jaro y Jaccard para cada uno de los modelos

necesariamente lo incluye dentro de la secuencia de markov que le atribuye a la dirección a menos de que haya evidencia de él en la dirección sucia. Más adelante se muestra el análisis cualitativo comparando entre parejas de modelos y especificar más comportamientos de este tipo.

El mismo comportamiento obtenido para la distancia y radio de Levenshtein se ve en las medidas no basadas en transformaciones o ediciones necesarias, como lo son el coeficiente de similitud de Jaro y el índice de Jaccard.

La parte izquierda de la figura **5-9**, muestra el histograma para los resultados de la distancia de Jaccard. En este caso, al ser una medida de distancia (entendida como el complemento del índice de similitud de Jaccard) se busca que el resultado sea lo más cercano a 0, y por lo tanto esté más cerca. Es importante recordar que esta medida de similitud está basada en el número de elementos comunes, en este caso entre direcciones. Por otro lado, en la métrica de distancia de Jaro, un valor cercano a 1 implica un grado mayor de similitud entre las cadenas de texto comparadas.

Ambos comparativos de histogramas, muestran nuevamente la superioridad del modelo basado en redes neuronales por sobre los modelos base y un mejor comportamiento del modelo de markov versus el modelo basado en reglas. Sin embargo, se puede notar que en el caso de la distancia de Jaro, que está basado en la cantidad y orden de los caracteres comunes, se puede ver que la variabilidad es mayor en el modelo de markov lo cual indica que hay un determinado tipo de direcciones en las que el modelo no logra generar direcciones con caracteres en un orden y posiciones cercanas a las del modelo estandar. Con el fin de entender mejor este y otros casos similares, se realizan los siguientes análisis cualitiativos complemen-

Rules_Addresses	Markov_Addresses
CARRERA 115E SUR NO 105DD23	CARRERA 115 E SUR NO 105 DD 23
TRANSVERSAL 161E DIAGONAL 52 D - 53	TRANSVERSAL 161 E DIAGONAL 52 D - 53
CARRERA 169BB 150 24	CARRERA 169 BB 150 24
TRANSVERSAL 40DD SUR TRVL 97 75	TRANSVERSAL 40 DD SUR TRANSVERSAL 97
TRANSVERSAL 154 D TRVL 110 53	TRANSVERSAL 154 D TRANSVERSAL 110 53
CALLE 134 BIS CRR26 6	CALLE 134 BIS CARRERA 26 6
TRANSVERSAL 124 DIA123BB48	TRANSVERSAL 124 DIAGONAL 123 BB 48

Tabla 5-4.: Markov versus Rules

tarios, comparando las direcciones generadas por parejas de modelo como se muestra en las tablas **5-4** y **5-5**.

Las primeras tres filas de la tabla **5-4** muestran direcciones para los cuales las medidas de distancia obtenidas, versus la dirección estandar son muy cercanas entre ambos modelos. Por otro lado, las siguientes cuatro direcciones son ejemplos en los que la diferencia en la distancia es mayor.

En este caso, se puede ver cómo, cuando la diferencia es menor, las cadenas de texto obtenidas son prácticamente iguales, sin embargo la principal diferencia se da en que el modelo de markov separa todos los elementos de la dirección, y el modelo de reglas no siempre lo logra, generando diferencias como “105DD23” vs. “105 DD 23”, “161E” vs. “161 E” o “169BB” vs. “169 BB”.

Por otro lado, en los casos en los que mayor diferencia se presenta la razón es que el modelo basado en reglas no siempre logra diferenciar el segundo elemento de la dirección para ser separado y posteriormente limpiado. Es por esto que se ven elementos aun no estandarizados en la dirección basada en reglas como lo son “CRR26 6”, “TVRL” y “DIA123BB48”. El modelo de Markov si logra identificar estos elementos y los limpia con la palabra completa (“CARRERA”, “TRANSVERSAL” y “DIAGONAL”). Sin embargo, esto hace que el modelo de Markov genere direcciones más largas versus el modelo estandar, lo que afecta medidas como la distancia de Jaro, basadas en que los caracteres comunes esten en posiciones similares dentro de la cadena. Este comportamiento ya se había evidenciado en el extremo derecho de la figura **5-9**

Para la comparación del modelo de Markov versus el modelo LSTM, se realiza el mismo ejercicio. En este caso las primeras 4 filas de la tabla **5-5** muestran direcciones para las que hay una diferencia significativa en las medidas de distancia versus la dirección real, mientras que las siguientes 5 líneas muestran casos en los que la distancia es baja y por lo tanto muestra cadenas de texto muy similares entre sí. Se peude ver como en las primeras direcciones, la principal diferencia se da en el hecho de que las cadenas de markov identifican el segundo elemento de la dirección con las palabras que le corresponden, mientras que en la versión LSTM, al igual que en la estandar, este segundo elemento se cambia por la palabra “NO”.

Markov_Addresses	LSTM_Addresses
DIAGONAL 199 A SUR TRANSVERSAL 156 CC 63	DIAGONAL 199A SUR NO 156CC - 63
DIAGONAL 147 A TRANSVERSAL 180 AA 1	DIAGONAL 147A NO 180AA - 1
DIAGONAL 38 BB TRANSVERSAL 58 BB 29	DIAGONAL 38BB NO 58BB - 29
CARRERA 44 B SUR CALLE 76 - 86	CARRERA 44B SUR NO 76 - 86
CALLE 198 86 20	CALLE 198 NO 86 - 20
CALLE 82 C SUR 114 B 31	CALLE 82C SUR NO 114B - 31
CALLE 46 OESTE 124 DD 43	CALLE 46 OESTE NO 124DD - 43
TRANSVERSAL 164 23 53	TRANSVERSAL 164 NO 23 - 53
TRANSVERSAL 102 A 49 63	TRANSVERSAL 102A NO 49 - 63

Tabla 5-5.: Markov versus LSTM

Target_Addresses	LSTM_Addresses
TRANSVERSAL 160C OESTE NO 19G - 51	TRANSVERSAL 160C OESTE NO 19 - 51
CALLE 57G NO 198AA - 42	CALLE 57D NO 198A - 42
TRANSVERSAL 167AA NORTE NO 46AA - 28	TRANSVERSAL 167AA NORTE NO 46A - 28
TRANSVERSAL 91A ESTE NO 85A - 66	TRANSVERSAL 91A ESTE NO 85CC - 66
TRANSVERSAL 136A SUR NO 163DD - 13	TRANSVERSAL 136A SUR NO 163D - 13
CARRERA 181AA NORTE NO 107D - 39	CARRERA 181AA NORTE NO 107E - 39
TRANSVERSAL 106BB NO 158D - 84	TRANSVERSAL 106BB NO 158E - 84
CARRERA 127C ESTE NO 22AA - 1	CARRERA 127C ESTE NO 22A - 1
CARRERA 177 ESTE NO 21CC - 13	CARRERA 177 ESTE NO 21C - 13

Tabla 5-6.: Errores comunes modelo LSTM

En los casos, más similares, la principal diferencia se da en que justamente la cadena de markov no asigna ningún atributo para identificar dicho segundo elemento y tampoco utiliza un separador entre los dos últimos números de la dirección. Dicho elemento, en general está presente como un guión en la dirección generada por la red.
Luego de verificar la eficacia del modelo de estandarización, basado en redes neuronales recurrentes de tipo LSTM, se busca analizar también, para aquellos casos en los que presenta alguna diferencia contra la dirección estandar base, el porqué de dicha diferencia. La tabla **5-6**, muestra precisamente algunas cadenas de texto cuya distancia de Levenshtein fue de 1 o 2 transformaciones. Al buscar los elementos comunes entre ellas se puede observar que, en todos los casos, la dirección objetivo tiene una letra que acompaña al segundo número presente en la dirección y que en muchos de los casos este complemento del número está conformado por más de una letra. A continuación , se listan varios de los posibles errores evidenciados en la direccón estandar obtenida con este modelo.

- Casos en los que la dirección obtenida, no contiene el elemento que acompaña al núme-

	Base	Markov	Clean
error	27.0	23.0	4.0
count	4999.0	4999.0	4999.0
mean	26880.0	16562.8	13408.5
std	489257.5	329940.3	294788.0
min	0.3	0.3	0.0
25 %	10.6	10.4	10.7
50 %	23.7	22.5	23.0
75 %	492.7	238.6	266.9
max	17225348.8	9081889.1	9081889.1

Tabla 5-7.: Resultados para la distancia a la ubicación real

ro. Caso “19**G** - 51” versus “19 - 51”

- Casos en los que el complemento original tiene dos letras y en la dirección obtenida solo se muestra una letra (“198AA” versus “198A”, “163DD” versus “163D”
- Casos en los que cambia la letra que corresponde al complemento (“107D” versus “107E”)
- Casos en los que cambia el complemento en cantidad y letras correspondientes (“85A” versus “85CC”)

5.4.2. Precisión de la geolocalización obtenida como resultado del modelo

Uno de los mayores usos de la estandarización de direcciones está en la posibilidad de geolocalizarlas. En esta sección se busca medir y comparar la precisión que se obtiene al geolocalizar la dirección estandarizada con cada uno de los métodos contra la ubicación real que corresponde a dicha dirección. La metodología utilizada es la descrita en la sección 5-3-2, similar a la propuesta en estudios como el de Matci en [2].

La tabla **5-7**, resume los principales resultados obtenidos para la distancia entre el punto obtenido con la dirección y la ubicación real de las 5000 direcciones con las que se cuenta. Se muestran la media, desviación estandar y los percentiles. Con el fin de analizarlos visualmente, se muestran también el gráfico de caja para cada uno de los modelos en la figura **5-10**.

El primero elemento a resaltar de la tabla es que el número de errores, obtenido al geolocalizar la cadena de texto resultante de la dirección, disminuye a medida que se limpian las direcciones y es de 0 para el modelo LSTM. Esto se explica porque la red neuronal está

entrenada para devolver siempre una dirección en formato geolocalizable, incluso cuando se trata de direcciones como "CALLE DEL COMERCIO PARQUE PRINCIPAL", mientras que el modelo basado en reglas deja esta dirección tal y como está, al igual que el modelo de markov al no encontrar una secuencia que cumpla con las evidencias presentadas. Sin embargo, esto no quiere decir que la dirección obtenida por el modelo LSTM vaya a estar cerca a la ubicación real del establecimiento, especialmente en estos casos en los que la secuencia generada no necesariamente describe bien la dirección original.

Por lo anterior, es necesario evaluar, en conjunto, todas las distancias encontradas para comparar los modelos entre sí. Al hacerlo, se puede observar como la distancia promedio en metros, obtenida a cada uno de los puntos, disminuye con cada uno de los modelos, pasando de 34km a 24km con el modelo de Markov y a 21km con el modelo tipo LSTM, al igual que la desviación estandar. Si bien esto representa una mejora, si se tiene en cuenta que se está tratando con geolocalizaciones, una distancia en kilometros no es un rango aceptable para la ayoría de aplicaciones. Sin embargo, esta distancia se explica por valores atípicos, causados por direcciones que contienen palabras clave como "AMALFí", las cuales a pesar de ser ubicaciones en Colombia, son geolocalizacdas por el API de Google Maps en ciudades de otros países (en este caso particular "San Francisco" en Estados Unidos). Aun así, esto indica que con el modelo tipo LSTM se obtienen menos direcciones de este tipo.

Esta presencia de valores atípicos, puede verse claramente en la parte izquierda de la figura **5-10**. Nótese cómo los valores extremos disminuyen significativamente entre el modelo basado en Reglas y los otros dos modelos. Puesto que dichos valores, no corresponden a la mayoría de ubicaciones, se procede a analizar la mediana y los cuantiles obtenidos, que dan una mejor perspectiva de lo que pasa con la mayoría de los datos. Para ello, se muestran los gráficos de caja, excluyendo valores atípicos de cada uno de los sets de datos de distancia, en la parte derecha de la figura.

La figura **5-10** muestra que, para los tres modelos, el 50 % de las ubicaciones está a aproximadamente 23 metros de la ubicación registrada. Si se tiene en cuenta que son establecimientos comerciales, es probable que a esta distancia el establecimiento se encuentra a la vista de quien lo visite, habilitando de este modo muchas de las aplicaciones de geolocalización en la industria (Ver Capitulo 6). Las diferencias empiezan a ser más notables en el 50 % restante, rango en el que el modelo de markov demuestra un mejor desempeño, con un rango intercuartil un poco más estrecho e inferior al del modelo LSTM.

Con el fin de analizar detalladamente los resultados de los modelos entre sí, se muestra, en la figura **5-11** la distribución de las distancias obtenidas en cada uno de ellos. La parte izquierda muestra la comparación entre el modelo de Markov y el modelo base construido con reglas, mientras que la parte derecha muestra la comparacion entre el modelo LSTM y

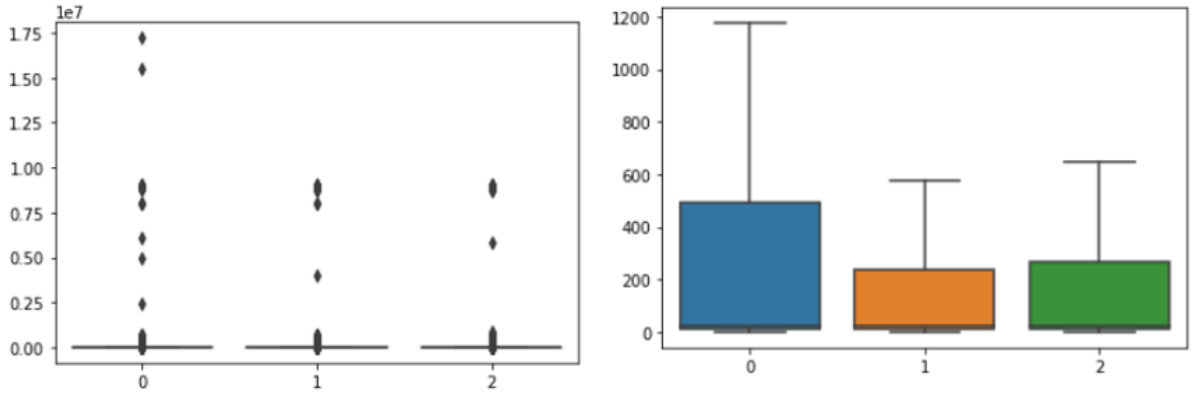

Figura 5-10.: Boxplots para cada uno de los modelos revisados 0. Reglas 1. Markov 2. LSTM

el modelo de Markov.

Cuando se comparan todos los rangos de distancias obtenidos, es fácil observar cómo, el modelo de Markov tiene un conteo mayor de direcciones en todos los rangos entre 0 y 500 metros, mientras que el modelo en basado en reglas obtiene más direcciones en todos los rangos a partir de un kilómetro de distancia contra la ubicación real. Estas diferencias son menos claras al comparar los resultados del modelo LSTM versus el modelo de Markov, pues, si bien el conteo de direcciones obtenidas con el modelo LSTM supera al modelo de Markov en los rangos inferiores a 30 metros, muestra valores inferiores en los rangos inferiores a 50 y 100 metros. Sin embargo, el hecho de que presente un mayor conteo de direcciones ubicadas entre 2 y 20km, es probablemente lo que causa las diferencias entre los rangos intercuartiles reflejada en la figura **5-10**. Aun así, en el extremo superior (rangos superiores a 50km), el modelo LSTM presenta un menor número de direcciones y por lo tanto un mejor comportamiento.

El API de google, provee determinadas clasificaciones para las direcciones una vez geolocalizadas, que permiten tener una medida de la precisión de cada uno de los métodos. Estas clasificaciones se definen a continuación.

- ROOFTOP: Indica que el resultado es una geolocalización precisa, para la cual Google cuenta con información precisa a nivel de Calle.
- RANGE_INTERPOLATED:Indica que la ubicación obtenida es el resultado de una interpolación entre dos puntos precisos (generalmente cruces entre calles). Esto ocurre generalmente cuando google no cuenta con geolocalizaciones precisas para una calle en

Location type	Base	Markov	LSTM
Errores	27	23	0
ROOFTOP	1676	1683	1663
RANGE_INTERPOLATED	2694	2815	2911
GEOMETRIC_CENTER	357	283	228
APPROXIMATE	245	195	186

Tabla 5-8.: Conteo de ubicaciones por clasificación según el API de geolocalización de google

particular.

- GEOMETRIC_CENTRIC: Indica que el resultado es el centro geométrico de una determinada calle o zona geográfica que google identifica con la dirección.
- Indica que el resultado obtenido es una aproximación, menos precisa que en los tres casos anteriores, generalmente a nivel de ciudad.

La tabla **5-8** provee el conteo de direcciones obtenidas para cada una de estos tipos de clasificaciones, así como el conteo de errores, divididos por método de estandarización. Es importante tener en cuenta en el análisis, que el hecho de que una dirección esté clasificada como rooftop, no necesariamente indica que la ubicación obtenida va a coincidir exactamente con la presente en la base de datos, sino únicamente que en google tiene una ubicación específica para la direccón de entrada (que al ser registrada manualmente podría estar sujeta a errores, por ejemplo, en los indicadores de calle o carrera).

Una vez más se observa un aumento en el número de direcciones ubicadas en los clasificaciones de precisión superiores, en especial en el rango "RANGE_INTERPOLATED" con los modelos de Markov y el modelo LSTM. Esto a su vez explica que el número de direcciones, sobretodo en el rango "APPROXIMATE", cuyo margen de error está a nivel de municipio (y por lo tanto puede ser de varios kilómetros) también disminuye con el modelo de markov y el modelo basado en redes neuronales propuesto.

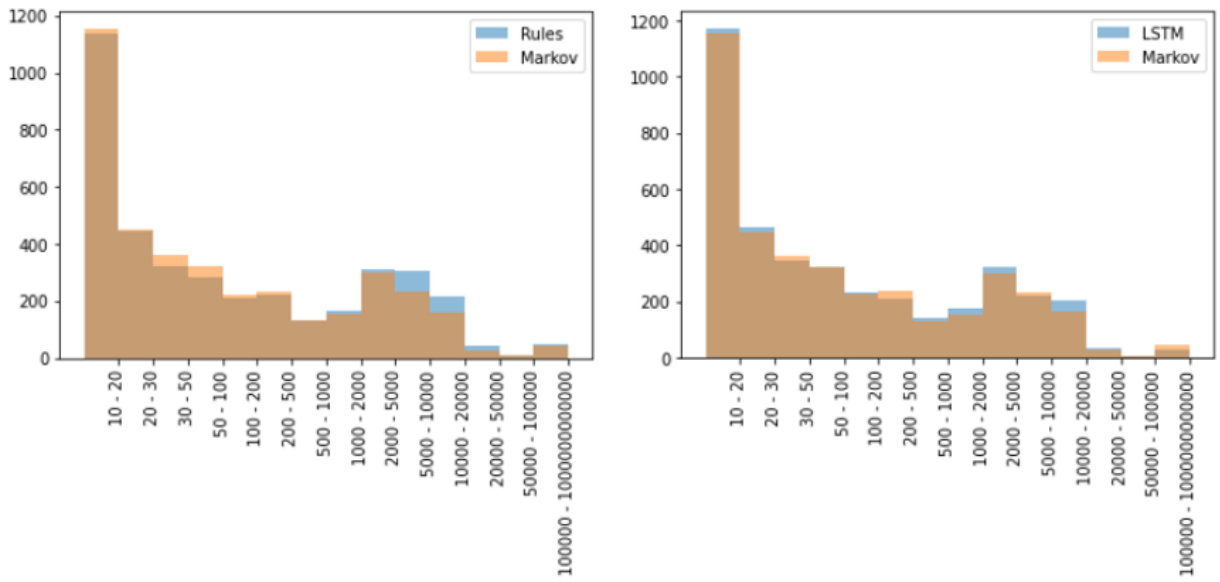

Figura 5-11.: Distribución de distancias. 1. Base vs Markov 2. MArkov vs LSTM

6. Aplicaciones en la industria

En el presente capítulo se evidencian los resultados del algoritmo de limpieza de direcciones postales + geolocalización, aplicado en diversas áreas dentro de una de las compañías de productos de consumo masivo en el país, y el mundo. Dichas aplicaciones, buscan demostrar el potencial que tiene la estandarización y posterior limpieza de direcciones en sus bases de datos propias, en la industria en general, para la toma de decisiones en atención al cliente, cadena de suplemento, búsqueda de eficiencias en el proceso de venta, evaluación del desempeño y diseñ y ejecución de pautas publicitarias.

6.1. Limpieza de bases de datos - Caso aplicado en Atención al cliente

La dirección geográfica de un determinado cliente, y por consiguiente su ubicación, es el elemento principal en la atención, para servicios y productos que se ofrecen a domicilio, pues de acuerdo a ella se debe establecer la operación que garantice la entrega apropiada y en el menor tiempo posible de dicho servicio o producto. Esta es una de las razones por las que, en los últimos años, se ha visto un incremento importante en startups y compañías dedicadas a operaciones de logística de última milla y entrega de Productos [17].
Sin embargo, gran parte de la operación tradicional de domicilios en el país, aún se realiza vía telefónica o a través de formularios en páginas Web que le permiten a quien registra un determinado pedido, escribir una determinada dirección según lo crea conveniente. Lo anterior hace que un mismo cliente pueda tener varias direcciones registradas en el sistema o incluso la misma escrita de diferentes formas. Si a ello se le suma que un determinado cliente puede pedir desde diferentes ubicaciones a una misma central de domicilios (Por ejemplo, desde la oficina, la casa de un amigo o el propio hogar), entonces la decisión de, desde qué tienda atender cada uno de los pedidos, cómo programar la entrega y el lugar exacto al que se debe enviar, es una tarea compleja desde el punto de vista analítico.
Tal es el caso de una importante cadena de Farmacias en Cali, que cuenta con una base de datos de aproximadamente +440000 direcciones diferentes para aproximadamente 150.000 clientes en la ciudad, todas ellas dígitadas manualmente por los operarios que reciben los pedidos. Dicha cadena de farmacias no cuenta con las coordenadas exactas de la ubicación de sus tiendas y la base de datos de direcciones de tiendas, no está integrada con su base de datos de clientes. En este caso, las preguntas de negocio que se quieren solucionar son:

- ¿Cuál es la dirección más probable desde la que un cliente me realiza un pedido
- ¿Desde qué tienda debería atender el pedido de un determinado cliente

Para resolver la primera pregunta de negocio se identificaron inicialmente todas las variaciones de las direcciónes asociadas a un determinado cliente (por código), Dichas variaciones se estandarizaron, unificándolas en un solo formato y removiendo los valores duplicados. Gracias a esto se obtienen todas las direcciones reales y diferentes, desde la cual un determinado cliente ha hecho un pedido históricamente. Por último, para seleccionar la dirección más probable desde la que un determinado cliente solicita un determinado producto se tomaron en cuenta aquellos pedidos realizados durante la etapa de confinamiento por COVID-19 en Colombia, bajo la presunción de que esta es precisamente su dirección de estadía, para aquellos clientes con pocos o ningún pedido durante esta etapa, se toma la moda. Vale la pena aclarar también que del total de clientes, solo un 41 % habían realizado pedido desde más de una ubicación, una vez estandarizadas dichas ubicaciones.
La aplicación del algoritmo permitió, entre otras cosas:

- Asignar coordenadas a +70.000 pedidos sin ubicación en la base de datos .Esto representa un 80 % del total no geolocalizado. El restante 20 %, se encuentra en direcciones no geolocalizables como es el caso de "BANCO DE SANGRE 1 PISO CLINICA VALLE DEL LILI TORRE1.º "La BUITRERA KM3".
- Identificar coordinadas asignadas erróneamente, o que no coincidan con la ubicación registrada
- Proponer una ubicación más probable para cada cliente, con la cual resolver la segunda pregunta de negocio.

Una vez geolocalizados la gran mayoría de pedidos y clientes. Se realizó el proceso de estandarización y geolocalización, para un total de 65 tiendas en la ciudad y se asignó a cada uno de los clientes la tienda más cercana, como aquella desde la que se debería atender un determinado pedido.
La figura **6-1**, muestra el resultado de la ejecución del algoritmo completo, resolviendo ambas preguntas de negocio. En este mapa, las farmacias son dibujados con marcadores más grandes y los clientes asignados ellas con marcadores más pequeños, representados con un mismo color.

6.2. Optimización de fuerzas de venta

El mercado de productos de consumo masivo no comestible en Colombia, está dividido principalmente en 4 Canales principales: Grandes superficios, supermercados regionales, farmacias y canal tradicional que incluye las que conocemos como "tiendas de barrio". Este último,

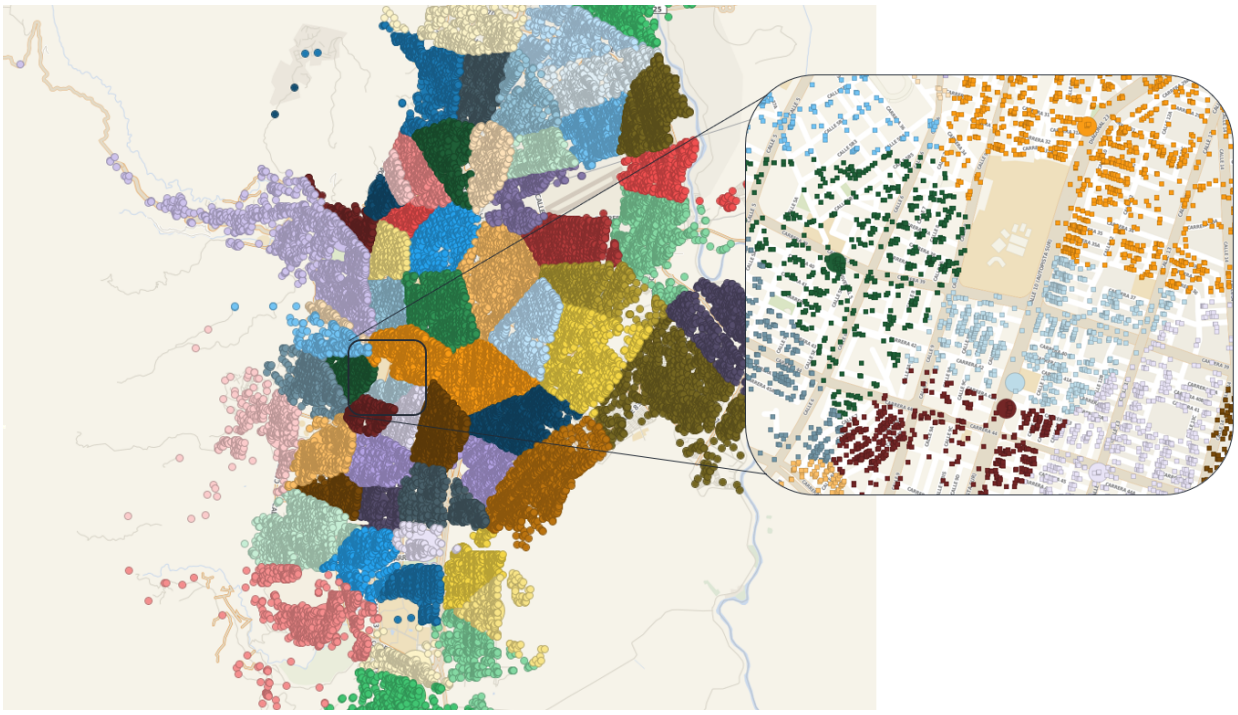

Figura 6-1.: Asignación final de clientes a farmacias después del proceso de estandarizacón

representa un gran porcentaje de las ventas para las compañías dedicadas a la venta de este tipo de productos. Sin embargo, es también uno de los más difíciles de atender de manera efectiva, debido a la gran cantidad de establecimientos (En Colombia se estiman en más de 400000 [[61]) y al alto grado de informalidad [61]

Lo anterior resalta la importancia que tiene para este tipo de empresas, poder contar con una base de datos de direcciones limpias y en un formato único, que permita además la geolocalización precisa de las mismas. A continuación se muestran varias aplicaciones realizadas a partir de la estandarización de dichas direcciones, enfocadas en la atención del canal tradicional.

6.2.1. Evaluación del desempeño por clusters

En el canal tradicional, la atención se da normalmente por vendedores asignados a determinadas zonas. Evaluar el desempeño de dichos fuerzas de venta, sus oportunidades y fortalezas a lo largo del mes, es una necesidad para garantizar hacer todo lo posible para garantizar la distribución requerida de los productos, y las metas establecidas. Sin embargo, realizar intervenciones con alto nivel de granularidad es un reto desde el punto de vista técnico.

Para hacerlo posible, en esta aplicación se tomó la información de 13200 tiendas ubicadas en Bogotá, Medellín y Cali, de las cuales se contaba con la secuencia asociada a la dirección, tal y como se encontraban en la base de datos del distribuidor a cargo. Se estandarizó y se

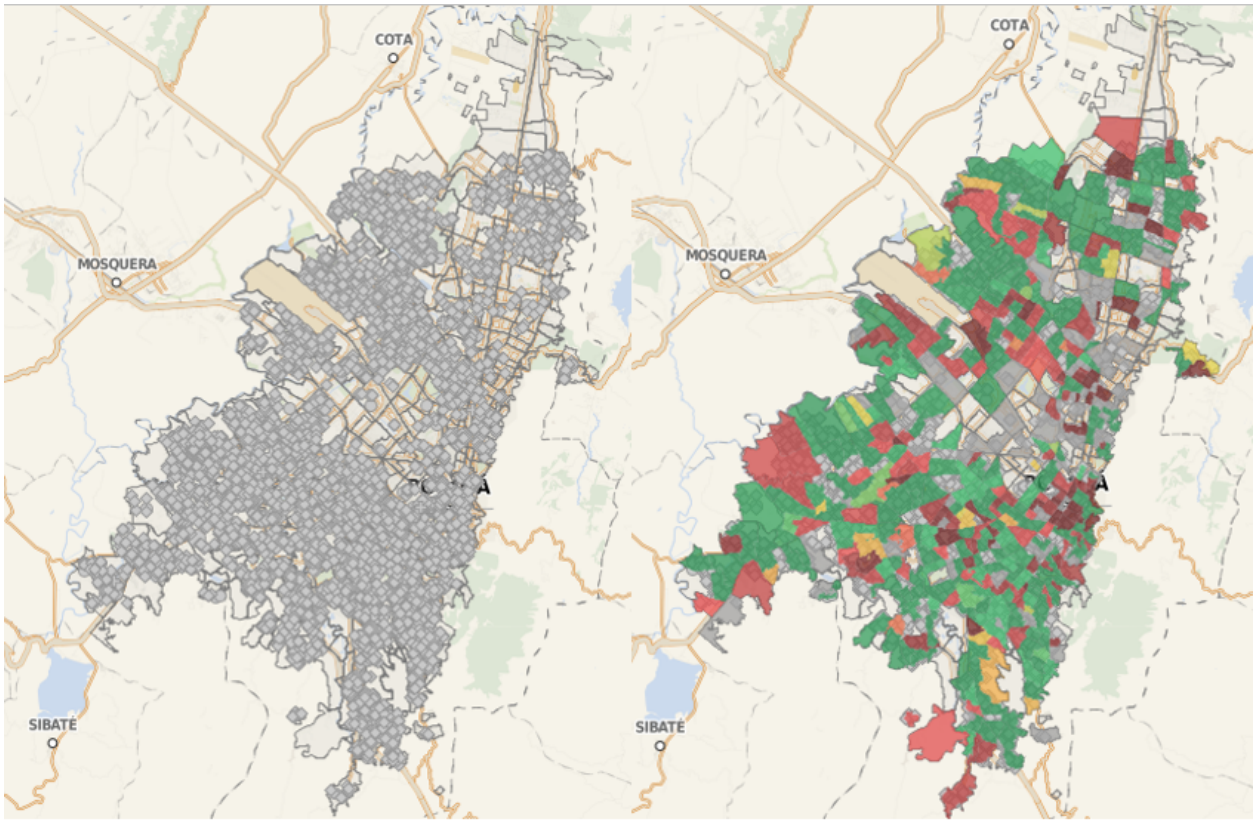

Figura 6-2.: Medidas de desempeño por clusters de tiendas.

ubicó cada una de ellas en el mapa de cada una de las ciudades utilizando GOOGLE API. Este corresponde al mismo proceso y set de datos, utilizado para el calculo de la medida de desempeño de los diferentes modelos, en el capítulo 5 de este trabajo.
Las tiendas gelocalizadas, para el caso de Bogotá, se pueden ver en la sección izquierda de la figura **6-2**. Una vez geolocalizadas, se cruzo esta información con el mapa de sectores de la ciudad (Un intermedio entre Barrio y Localidad), proveído por el DANE [62], y se utilizó un codígo de color que funciona de para diferentes medidas que un determinado usuario puede seleccionar y que incluyen ventas, distribución, cantidad de productos comprados en promedio por tienda o por vendedor, cantidad de transacciones, entre otras útiles para la toma de decisiones en el canal. Esto se habilitó a modo de tablero digital y se actualiza diariamente todos los días. La visualización final para una de las medidas específicas, se muestra en la parte derecha de la **6-2**.
Este tablero, permite tomar decisiones. A continuación se listan algunas de ellas.

- Identificar zonas vacantes o con restricciones de capacidad para habilitar más rutas en dichas zonas.

- Evaluar posibles aperturas de subdistribuidores para atender determinadas zonas.

- Evidenciar posibles restricciones al flujo dentro de determinadas zonas que afecten las medidas de desempeño
- Diseño de combos, promociones y actividades con la fuerza de ventas para promover la distribución de determinados productos o categorías.
- Diseño de lanzamientos en determinadas zonas, que eviten la canibalización de productos de la misma compañía.

6.2.2. Algoritmos de recomendación de productos

El rápido crecimiento del e-commerce y las plataformas digitales para la gestión del proceso de ventas, ha generado que un determinado cliente se vea sobrecargado por muchas opciones a considerar en muy poco tiempo para tomar una decisión de compra. Es por esto que los sistemas de recomendación han tomado cada vez más relevancia como soporte al proceso de compra y venta de productos para ofrecer lo que realmente puede interesarle al cliente [63]. En el contexto de las tiendas de canal tradicional, en cuestión, se evidencia exactamente esta situación, teniendo en cuenta que las tiendas manejan más de productos en su portafolio [64] y generalmente no cuentan con mucho tiempo para atender a los vendedores de las diferentes empresas por lo que las recomendaciones del vendedor son cada vez más relevantes.
Si bien, existen muchos tipos de algoritmos de recomendación, como se puede ver en [63][65], la gran mayoría de ellos se basa en identificar clientes similares y sus comportamientos de compra para determinar lo que le puede interesar a cada uno de ellos. Esta medida de similitud entre clientes se basa generalmente en sus transacciones, pero la zona en la que están ubicados juega un papel determinante en el contexto de consumo masivo [64]. Del mismo modo, identificar las características de una determinada zona, permite también buscar zonas con características geográficas o demográficas similares aún si están distantes unas de otras, y de este modo aumentar el universo de clientes similares.
Para este fin, se aplicó el proceso de estandarización y geolocalización de direcciones, en este caso en perú, justamente para identificar los vecindarios que se debían considerar en el diseño de recomendaciones de cada uno de los clientes. La figura **6-3** ilustra los vecinos greográficos de una determinada tienda el Lima.

6.2.3. Ruteo de fuerzas de ventas

Una de las grandes áreas de estudio en lo que tiene que ver con gestion de fuerzas de venta, es el aumento de la eficiencia con la que las personas que componen estos grupos de vendedores, atienden a cada uno de sus clientes. Lo anterior incluye, entre otros, los problemas de asignación de clientes a cada uno de los vendedores de un determinado grupo; y la definición del orden en el que cada vendedor debería atender a cada uno de los clientes, que da lugar a un problema de ruteo [**?**].

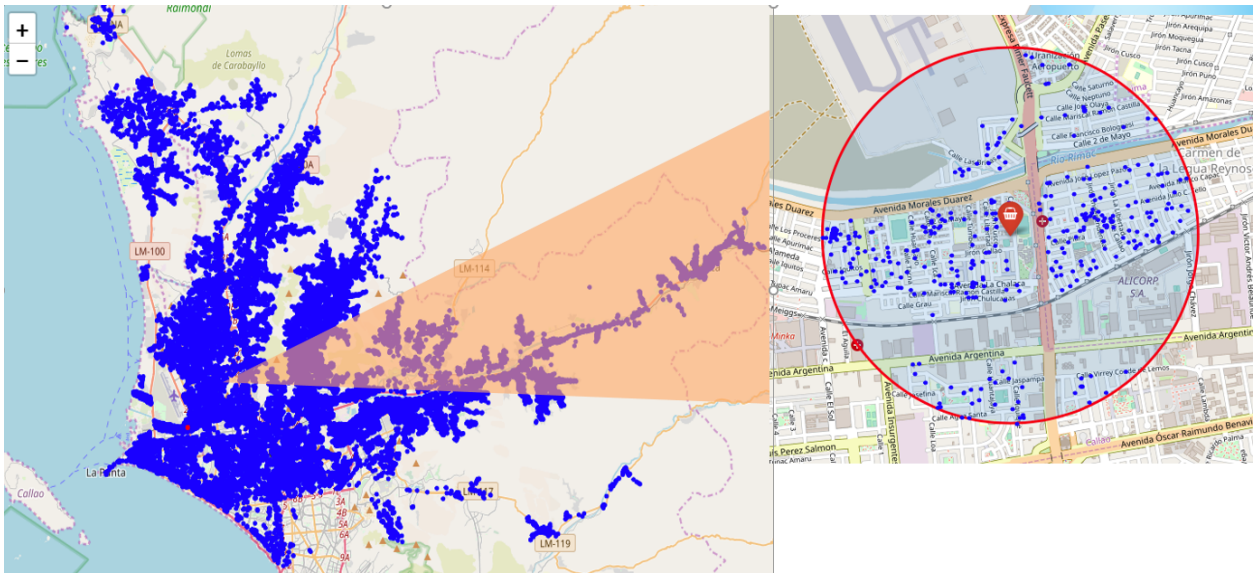

Figura 6-3.: Muestra Vecindario Tiendas Lima

Aunque este es un problema analítico, ampliamente estudiado en la literatura [**?**, 16], su implementación básica requiere tener la ubicación de los clientes que se desea atender y, al igual que en el caso anterior, dicha información no está disponible y los vendedores únicamente cuentan con la dirección, por lo que los atienden según su conocimiento de la zona y experiencia, en el orden en el que lo consideran eficiente.
Con el modelo descrito, se procedió a correr el ejercicio para saber cual debería ser el número de personas a contratar para realizar ciertas labores en tiendas de retail, teniendo en cuenta que estas se debían visitar un determinado número de veces y por un determinado periodo de tiempo, a la semana.
Tras correr el proceso de limpieza de direcciones, geolocalización de tiendas, y correr varios escenarios basados en metaheurísticas para el problema del vendedor viajero se obtuvieron algunos escenarios como el mostrado en la figura **6-4**, en el que se pueden diferencias las diferentes rutas por color, garantizando que cumplieran con algunas restricciones adicionales del modelo (como lo es, por ejemplo, la exclusividad de rutas entre clientes). Uno de estos escenarios demostró tener un 3 % de eficiencia, por sobre el modelo actual, trayendo beneficios cuantificables a la organización en cuestión.

6.3. Aplicaciones en Mercadeo y Mercadeo digital

La cada vez mayor conectividad y utilización de elementos digitales como los telefónos inteligentes, junto con múltiples tecnologías emergentes en el área del retail, en los últimos años tales como la baliza electrónica (También conocida como "beacon" por su nombre en ingle) han resaltado el rol de la ubicación de un potencial cliente y sus patrones de consumo en el

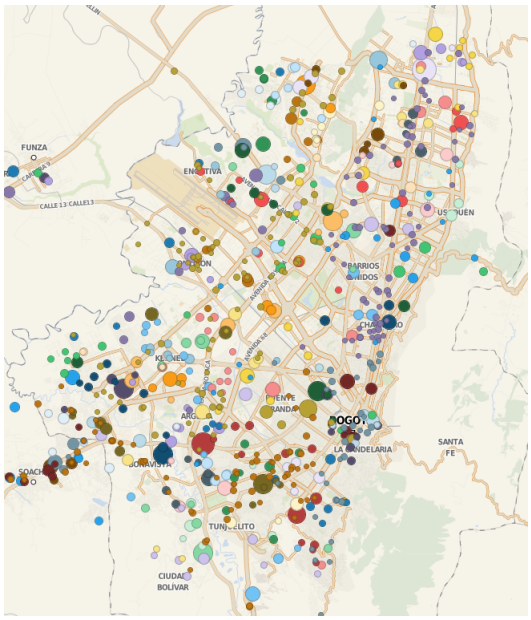

Figura 6-4.: Asignación geolocalizada de clientes a grupos de vendedores

diseño de pautas por parte de las empresas, de manera que se ofrezcan productos y servicios a la medida de dichos clientes. [66] .
Es así como la toma de decisiones basadas en la geolocalización ha permeado el mundo del marketing y por lo tanto, todas las fuentes de información que permitan realizar pautas más eficientes, son cada vez más valoradas. Las direcciones, nuevamente, son una de estas posibles fuentes de información que le permiten a las empresas saber la ubicación aproximada de un determinado consumidor y tomar decisiones con base en ella, siempre y cuando esta se logre geolocalizar. A continuación se muestran dos ejemplos de aplicaciones del algoritmo en el contexto del Marketing, para productos de consumo masivo.

6.3.1. Entendimiento de consumidor de acuerdo a información censal

Una importante fuente de información para el entendimiento del consumidor colombiano, es la recolectada por el Departamento Nacional de Estadística, DANE, una vez realizado el censo poblacional [62]. Esta información es pública, y puede ser descargada en forma de mapas que permiten ver los resultados de las medidas demográficas tomadas en el cuestionario, hasta el nivel de sector dentro de una determinada ciudad o población. Los demográficos disponibles incluyen, entre otros, número de personas por hogar, la edad de los integrantes de cada hogar, el nivel de estudios, promedio de ingresos, estrato socioeconómico, ocupación, entre otros que pueden ser relevantes para la caracterización del consumidor de una determinada zona. La información, agregada en el sector permite identificar características comunes de los consumidores reales o potenciales de dicha zona geográfica y por lo tanto, proveen información valiosa para la activación de posibles campañas publicitarias o activaciones en

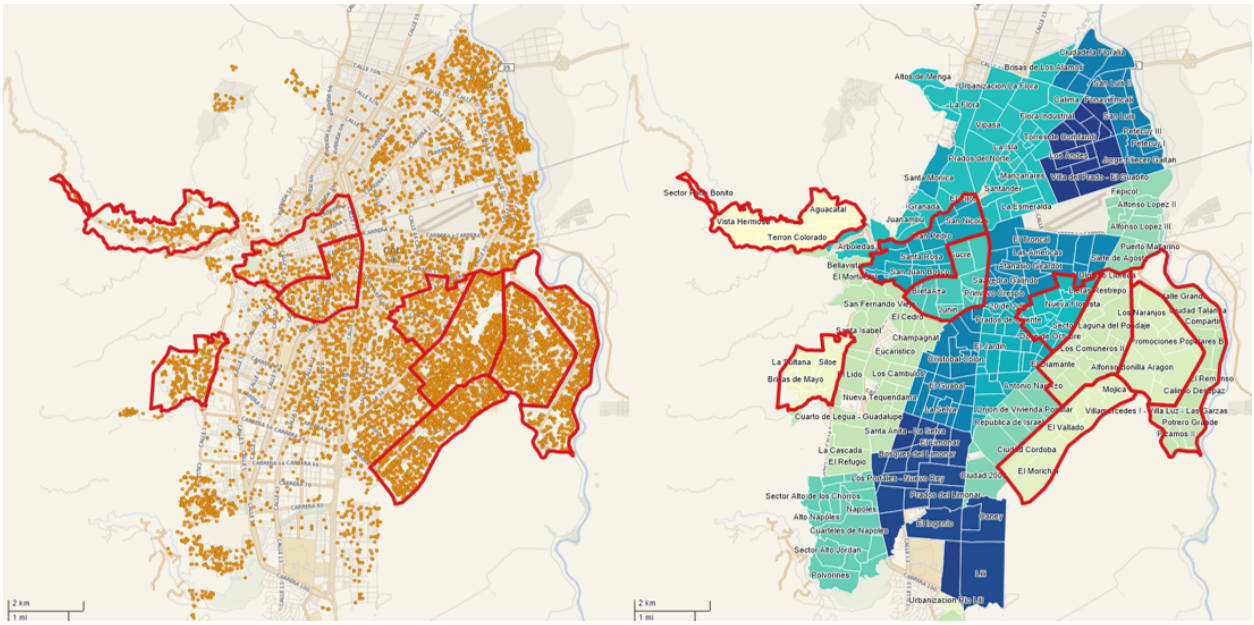

Figura 6-5.: Cruce de ubicación con información censal para la toma de decisiones

los establecimientos comerciales cercanos a dichos consumidores.

La figura **6-5** muestra un ejemplo de activación de este tipo en la ciudad de Cali. La parte izquierda, muestra la ubicación de las tiendas de barrio de Cali, registradas en determinada base de datos; mientras que la derecha, se muestra con código de color, en escala de azul, el nivel de ingresos promedio, de los hogares presentes en cada una de las zonas y con un borde rojo las zonas de enfoque para la activación de un determinado plan de medios.

En este caso, dicho plan consistió en material adicional en tienda, pauta en televisión y radio, alianzas con los equipos de futbol de máyor influencia en esta zona y el lanzamiento de un producto diseñado para las caracterísiticias de ese consumidor, lo que llevo a un crecimiento del 17 % en ventas, luego de la ejecución del plan.

6.3.2. Mercadeo Geolocalizado

Adicionalmente, el cruce de información demográfica con información de patrones de consumo, permite no solo identificar qué comunicar y dónde comunicarlo, sino también cuándo comunicarlo. Este es el caso ilustrado en la figura **6-6**, en el que se ilustra cómo, para un determinado producto, el patrón de compra varía, dependiendo de si el consumo se da en sectores con un promedio de ingreso bajo, o alto, medido según el estrato socioeconómico del sector.

Tras realizar el análisis de los patrones de consumo en las tiendas de esos sectores, se puede identificar que, el producto en cuestión, se compra mucho más durante los días de semana en estas zonas, mientras que en sectores con estrato más alto, el producto se suele comprar

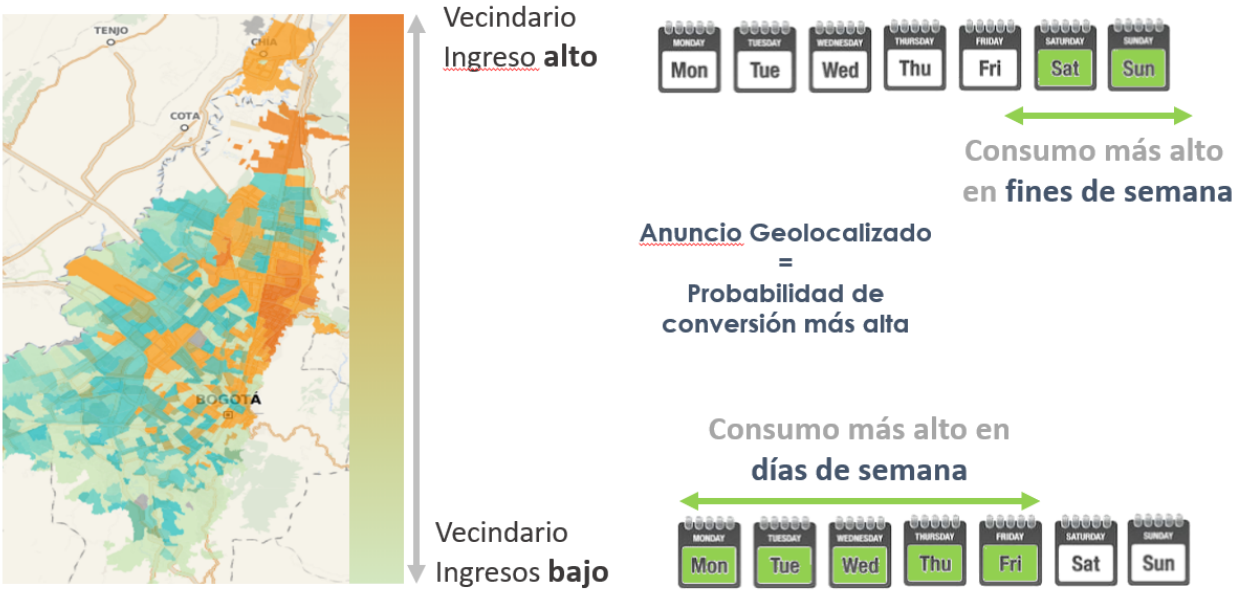

Figura 6-6.: Patrón de consumo para un determinado producto según el nivel de ingresos

durante los fines de semana. Esta conclusión, permitió definir planes de comunicación a los dispositivos móviles presentes en estos sectores para que los usuarios vean determinados anuncios en los días en los que es más probable que salgan a comprar, alcanzando incrementos de hasta 200 % en las ventas de dicho producto en determinados sectores.
Para hacerlo posible, se utilizó la información de direcciones de 850 tiendas de retail, que fueron estandarizadas y geolocalizadas con la red neuronal tipo LSTM, en conjunto con la información histórica de ventas de cada tienda de este determinado producto, lo que a su vez lo hace un modelo escalable para otros posibles productos en el portafolio de la compañia.

7. Conclusiones y recomendaciones

7.1. Conclusiones

En el presente trabajo se ha demostrado que la metodología propuesta para la limpieza y estandarización de direcciones geográficas, basada en una red neuronal recurrente de tipo LSTM y entrenada sobre conjuntos de datos autogenerados, permite obtener una cadena de texto en un formato estructurado y estandar, incluso a partir de múltiples variaciones de una misma dirección. Estas variaciones se dan generalmente, en aplicaciones reales, por la libertad que tiene el usuario a la hora de registrar dicha información en una base de datos, lo que a su vez implica que el método de limpieza y estandarización debe ser capaz de adaptarse a tantas variaciones como pueda un humano registrar cada elemento de la dirección geográfica.

Es precisamente en este último aspecto, en el que el modelo LSTM demuestra un mejor desempeño que el de los dos modelos base seleccionados (por su recurrente aplicación en la industria y en la academia): el modelo basado en reglas de limpieza y el modelo basado en cadenas de Markov ocultas. Dicho comportamiento, se evidenció tanto en las medidas de similitud entre cadenas de texto utilizadas, como en las diferencias entre la geolocalización obtenida desde la direccón estandar y la ubicación real, para determinados puntos de referencia. Esta metodología, permitió a su vez comparar los diferentes modelos utilizando datos simulados y datos reales para los que no se tenía una dirección estandar.

Por otro lado, el uso de la distancia de Levenshtein, La distancia de Jaccard y la similitud de Jaro entre cadenas de texto, aplicadas en conjunto con una comparación cualitativa entre las direcciones obtenidas por cada uno de los modelos, permite identificar comportamientos recurrentes en las direcciones obtenidas. De esta forma, se identificó que el modelo basado en reglas no es capaz de diferenciar varios de los elementos de la dirección, cuando las cadenas de texto asociadas a ellos están unidas entre sí (ej. “CL14BB62”). El modelo de markov, por su parte, sí es capaz de separar dichos elementos, pero suele agregar un identificador completo a la segunda mitad de la dirección (ej. “ Calle ”, “Transveral”) cuando se espera un separador único (“NO”). Este comportamiento, hace que para este tipo de direcciones, la similitud de Jaro del modelo de markov, basada en la ubicación y orden de los caracteres, tenga un desempeño inferior al del equivalente basado en reglas.

A pesar de su buen desempeño, los modelos de machine learning, como el LSTM propuesto,

necesitan un conjunto de datos de entrenamiento que, para el caso de limpieza de información no estructurada, implica tener el correspondiente estandar de lo que se desea limpiar. Esto representa una desventaja cuando se compara con los modelos de línea base, que parten de reglas y evidencias, ya presentes en el formato de la dirección. Por lo anterior, se hace necesario contar con una estrategia de generación de datos de entrenamiento y validación. En este sentido, la estrategia propuesta en este trabajo, basada en cadenas de markov ocultas, selección aleatoria de variaciones e introducción de error aleatorio, ha demostrado ser efectiva para el entrenamiento de la red neuronal, lo que se evidencia al medir el desempeño en set de datos reales, medido como la distancia a puntos de referencia previamente ubicados.

La metodología propuesta en este estudio debe ser aplicada de manera integral, lo cual incluye la estrategia de generación de data de validación sintética, la construccion y entrenamiento del modelo LSTM y la ejecución en limpieza de direcciones, ya sean sintéticas o reales. Sin embargo, el valor de la dirección geográfica no está únicamente en el hecho de tenerla limpia, sino también en las aplicaciones que su limpieza y posterior geolocalización habilita. Este es el caso de la optimización de fuerzas de venta o los mecanismos de atención a clientes; áreas que se ven beneficiadas con la posibilidad de realizar seguimiento a nivel geográfico, comparar clientes y consumidores cercanos entre sí, diseñar rutas de venta optimas, entender las características de un consumidor y llegar a dicho consumidor con publicidad personalizada, entre muchas otras necesidades presentes en las organizaciones actuales.

7.2. Recomendaciones

Durante la validación de la metodología propuesta y el entrenamiento del modelo LSTM, se evidenció una mejora significativa tras simplificar la dirección de entrada, lo cual incluye la sustitución de los caracteres numéricos de la dirección por un símbolo (“#”), la sustitución de caracteres especiales y la eliminación del complemento de la dirección que generalmente no brinda información adicional sobre la ubicación (Tal es el caso de palabras como “APTO”, “BARRIO”, “LOCAL”, “CONJUNTO”, etc). Como trabajo futuro, se recomienda incluir estos elementos adicionales de la dirección, en el proceso de limpieza y estandarización, siempre que estos sean importantes para el caso de uso en cuestión. Para ello, se debe tener en cuenta que estos elementos pueden tener un mayor número de variaciones que las ya consideradas, e incluirlas dentro del conjunto de entrenamiento.

Adicionalmente, se recomienda evaluar y adaptar el modelo, de manera que funcione para zonas geográficas específicas, más allá de las consideradas en este estudio; en las que el identificador de una determinada vía puede tener un nombre en lugar de un número (Ej, “Calle del comercio”, “Parque principal” o “Calle Colombia”) diferentes a los ya considerados en el conjunto de datos de entrenamiento disponible en el ANEXO B.

A pesar de los buenos resultados obtenidos con el modelo LSTM, Se debe tener en cuenta que la red utilizada tiene una arquitectura y número de neuronas fijo. Es posible que existan variaciones de los parámetros de la red que permitan obtener resultados similares o mejores, con menos entrenamiento. Por lo tanto, se recomienda explorar cambios en la arquitectura de la red, el numero de neuronas, el tipo de funciones de activación utilizada y la codificación de las secuencias de entrada y salida, para llegar a modelos que pueden ser más eficientes, teniendo el aquí propuesto como nueva base de comparación.

Por último, a la hora de evaluar el desempeño utilizando distancias versus puntos de referencia, es importante considerar (y de ser posible, separar), el error causado por la geolocalización con el API seleccionada, puesto que estas no garantizan un 100 % de precisión para todas las direcciones. Si se tiene en cuenta que estas aplicaciones permiten obtener una medida del error promedio generado en la geolocalización, una posible área de estudio relacionada, es la selección de la mejor API a utilizar, dependiendo de la zona geográfica, el caso de estudio y el formato de direcciones con los que se esté tratando.

A. Anexo: Diccionario de transformaciones para método basado en reglas

Tabla A-1.: Serie de transformaciones realizadas en el método basado en reglas

Original	**Limpio**	**Original**	**Limpio**
CL	CALLE	CR	CARRERA
CLL	CALLE	CRR	CARRERA
CLLE	CALLE	CRA	CARRERA
CALLLE	CALLE	K	CARRERA
CALLEV	CALLE	KR	CARRERA
CALLEY	CALLE	KRA	CARRERA
CALLE-	CALLE	CAR	CARRERA
CALLA	CALLE	CARRERA-	CARRERA
CALL	CALLE	CAREEA	CARRERA
CALEL	CALLE	CARREERA	CARRERA
CALE	CALLE	CARR	CARRERA
CAL	CALLE	CARRA	CARRERA
CAL.	CALLE	CARRE	CARRERA
CLLE.	CALLE	CARR	CARRERA
CLLL	CALLE	CARERRA	CARRERA
CLL.	CALLE	CARERA	CARRERA
CLL=	CALLE	CARA	CARRERA
CLL0	CALLE	CAREERA	CARRERA
CLL'	CALLE	CAR	CARRERA
CLE 0	CALLE	CCARRERA	CARRERA
CLALLE	CALLE	KRA.	CARRERA
CL. -	CALLE	KR.	CARRERA
CL.0	CALLE	KR,	CARRERA
CL .	CALLE	KRR	CARRERA
CL. 0	CALLE	CARRRERA	CARRERA
CL+L	CALLE	CRRA	CARRERA

Cl	CALLE	CTCARRERA	CARRERA
ACVALLE	CALLE	CRA:	CARRERA
AC	CALLE	CRA.	CARRERA
”	”	CR.	CARRERA
No.	NO	AVCARRERA	CARRERA
”_”	”	AK	CARRERA
DG	DIAGONAL	AV.	AVENIDA
DIA	DIAGONAL	AVNDA	AVENIDA
DIAG	DIAGONAL	AVNDA	AVENIDA
DIAG.	DIAGONAL	AVENIDA-	AVENIDA
DIG	DIAGONAL	AVL	AVENIDA
DIANGONAL	DIAGONAL	AVN	AVENIDA
TRANS	TRANSVERSAL	AVENIDA-	AVENIDA
TRANSV	TRANSVERSAL	AVEND	AVENIDA
TRAV	TRANSVERSAL	AVENIDA-	AVENIDA
TRV	TRANSVERSAL	AVE	AVENIDA
TRB	TRANSVERSAL	AVD	AVENIDA
TV 0	TRANSVERSAL	AVDA	AVENIDA
TRAN	TRANSVERSAL	AV-	AVENIDA
TRAN.	TRANSVERSAL	AV	AVENIDA
TRANS.	TRANSVERSAL	N=	NO
TRANSNVERSAL	TRANSVERSAL	N0	NO
TRANSP	TRANSVERSAL	NON.	NO
TRASSVERSAL	TRANSVERSAL	N	NO
TRASV	TRANSVERSAL	N-	NO
TRASNVERSAL	TRANSVERSAL	N:	NO
TRASNSVERSAL	TRANSVERSAL	NRO	NO
TRANSSVERSAL	TRANSVERSAL	NRO.	NO
TRANSV–	TRANSVERSAL	NÂ¶Ã˜	NO
TRANSV	TRANSVERSAL	NUMERO	NO
TRANSVE	TRANSVERSAL	NUMERO	NO
TRASSVERSAL.	TRANSVERSAL	N.	NO
TRASV.	TRANSVERSAL	N?	NO
TRASV	TRANSVERSAL	NUMERO	NO
TRASVERSAL	TRANSVERSAL	NÃ˜	NO
TV	TRANSVERSAL	#	NO
TANSVERASAL	TRANSVERSAL	NUMERO	NO
TRANSVER	TRANSVERSAL	NUM.	NO

B. Anexo: Set de datos de validación para la capacidad de estandarización

Tabla B-1.: Primeras 50 direcciones. Set de entrenamiento

Original_Address	Clean_Address
KR 148 SUR #72 72	CARRERA 148 SUR NO 72 - 72
CARRER116 AA 162AA 60	CARRERA 116AA NO 162AA - 60
KR142B NORTE NO. 144 2	CARRERA 142B NORTE NO 144 - 2
KR181 G NUM 22 18	CARRERA 181G NO 22 - 18
CR 46 D - 84D - 17	CARRERA 46D NO 84D - 17
CRA 199DD OESTE NO 27 B7	CARRERA 199DD OESTE NO 27B - 7
KR144 NORTE # 109E4	CARRERA 144 NORTE NO 109E - 4
KR 153E OESTE NO. 153E4	CARRERA 153E OESTE NO 153E - 4
CRA 168 NUM. 17A78	CARRERA 168 NO 17A - 78
CRR200 NORTE CALLE 5A40	CARRERA 200 NORTE NO 5A - 40
CR 59A BIS - 116 85	CARRERA 59A BIS NO 116 - 85
CR111 BIS # 94 AA45	CARRERA 111 BIS NO 94AA - 45
KR 54 OESTE KL 166 - 37	CARRERA 54 OESTE NO 166 - 37
KR 89 BB SUR NO 9 16	CARRERA 89BB SUR NO 9 - 16
KR124 - 179 AA17	CARRERA 124 NO 179AA - 17
KR 131 B OESTE - 65DD65	CARRERA 131B OESTE NO 65DD - 65
KR15A NO. 130 DD60	CARRERA 15A NO 130DD - 60
CRA 156AA ESTE NO 85 9	CARRERA 156AA ESTE NO 85 - 9
CR 178 # 121 32	CARRERA 178 NO 121 - 32
CR.93C ESTE NO 104 - 27	CARRERA 93C ESTE NO 104 - 27
KR169 - 50D24	CARRERA 169 NO 50D - 24
CR195 A # 149C 79	CARRERA 195A NO 149C - 79
CR170A OESTE N 180 21	CARRERA 170A OESTE NO 180 - 21
CR93 #78 E15	CARRERA 93 NO 78E - 15
KR79 E # 54 41	CARRERA 79E NO 54 - 41
CR 120 OESTE - 99C29	CARRERA 120 OESTE NO 99C - 29
CR 29A OESTE CLE 144B23	CARRERA 29A OESTE NO 144B - 23

CR. 78 BB 5 54	CARRERA 78BB NO 5 - 54
CR 158 SUR - 104 D64	CARRERA 158 SUR NO 104D - 64
CR87 ESTE 95 -66	CARRERA 87 ESTE NO 95 - 66
KR97 D BIS N147D39	CARRERA 97D BIS NO 147D - 39
CR 55 BB # 106 D71	CARRERA 55BB NO 106D - 71
CR 144 OESTE # 4B-46	CARRERA 144 OESTE NO 4B - 46
KR68A - 46D32	CARRERA 68A NO 46D - 32
CR 166 - 195F46	CARRERA 166 NO 195F - 46
CR 42E CL. 32A85	CARRERA 42E NO 32A - 85
CRA43 NORTE KLL 123 B55	CARRERA 43 NORTE NO 123B - 55
CR121 NUM131 B73	CARRERA 121 NO 131B - 73
CR173 E CLL 92 25	CARRERA 173E NO 92 - 25
CR 14 - 182E 45	CARRERA 14 NO 182E - 45
CR 96 E ESTE 161 BB24	CARRERA 96E ESTE NO 161BB - 24
CARRER134 9 C26	CARRERA 134 NO 9C - 26
KR48D SUR # 78 B53	CARRERA 48D SUR NO 78B - 53
CR191 SUR 138 B59	CARRERA 191 SUR NO 138B - 59
CRA192 NORTE N.187C48	CARRERA 192 NORTE NO 187C - 48
CR8 E - 97 45	CARRERA 8E NO 97 - 45
CR 4 SUR # 113 14	CARRERA 4 SUR NO 113 - 14
KR51 AA ESTE # 1 3	CARRERA 51AA ESTE NO 1 - 3
CR123 # 128 26	CARRERA 123 NO 128 - 26

Bibliografía

[1] V. Borkar, K. Deshmukh, and S. Sarawagi, "Automatic segmentation of text into structured records," *ACM SIGMOD Record*, vol. 30, no. 2, pp. 175–186, 2001.

[2] D.üçük Matci@ and U. Avdan, "Address standardization using the natural language process for improving geocoding results," *Computers, Environment and Urban Systems*, vol. 70, no. February, pp. 1–8, 2018.

[3] G. Sharma, Shikhar; Ratti, Ritesh; arora, Ishaan; Solanki, Anshul,; Bhatt, "Automated Parsing of Geographical Addresses : A Multilayer Feedforward Neural Network based approach," in *IEEE international Conference on Semantic Computing*, pp. 123–130, 2018.

[4] O. F. I. Pachón Quevedo and S. I. Tellez, "Propuesta de Estándar de las Direcciones Urbanas para los Equipamientos del Ministerio de Educación," p. 42, 2009.

[5] D. W. Goldberg, J. N. Swift, and J. P. Wilson, "Address Standardization," Tech. Rep. 12, 2017.

[6] K. Malik, Muhammad Noman; Abdul, "Address Standardization using Supervised Machine Learning," in *2011 International Conference on Computer Communication and Management*, no. November, 2015.

[7] V. Borkar, K. Deshmukh, and S. Sarawagi, "Automatically extracting structure from free text addresses," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 27–32, 2000.

[8] I. Mulasastra and A. Taplaksint, "Elementization of Thai Postal Addresses : A Hybrid Approach," in *2015 IEEE International Conference on Electrical and Computer Engineering (WIECON-ECE)*, 2015.

[9] G. Kothari, T. A. Faruquie, L. V. Subramaniam, K. H. Prasad, and M. K. Mohania, "Transfer of supervision for improved address standardization," *Proceedings - International Conference on Pattern Recognition*, pp. 2178–2181, 2010.

[10] D.üçük Matci@ and U. Avdan, "Address standardization using the natural language process for improving geocoding results," *Computers, Environment and Urban Systems*, vol. 70, no. January 2017, pp. 1–8, 2018.

[11] M. N. Masrek and Z. A. Razak, "Malaysian address semantic: The process of standardization," *2nd International Conference on Computer Research and Development, ICCRD 2010*, pp. 77–80, 2010.

[12] G. K. Tanveer, A. F. L. Venkata, S. K. Hima, and P. Mukesh, "Transfer of supervision for improved address standardization," in *2010 International Conference on Pattern Recognition*, pp. 2182–2185, 2010.

[13] Informatica, "Address Validation Best Practices for Interpreting and AnalizingAddress Data Quality Results," 2013.

[14] Runner enterprise Data Quality, "ADDRESS DATA CLEANSING: A BETTER APPROACH," 2017.

[15] R. A. Abbasi, "Information Extraction Techniques for Postal Address Standardization," *Faculty of Computing - Riphap International University*, 2005.

[16] C. Lin, K. Choy, G. Ho, S. Chung, and H. Lam, "Survey of Green Vehicle Routing Problem: Past and future trends," *Expert Systems with Applications*, vol. 41, pp. 1118–1138, mar 2014.

[17] H. Jafari, "e-Commerce Logistics â" Contemporary Literature," *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 1196–1200, 2018.

[18] P. Christen, T. Churches, and A. Willmore, "A probabilistic geocoding system based on a national address file," *Proceedings of the 3rd Australasian Data Mining Conference, Cairns*, 2004.

[19] P. Rogerson, D. Han, J. L. Freudenheim, J. E. Vena, M. R. Bonner, and J. Nie, "Positional Accuracy of Geocoded Addresses in Epidemiologic Research," *Epidemiology*, vol. 14, no. 4, pp. 408–412, 2004.

[20] S. A. Collier, L. J. Stockman, L. A. Hicks, L. E. Garrison, F. J. Zhou, and M. J. Beach, "Direct healthcare costs of selected diseases primarily or partially transmitted by water.," *Epidemiology and infection*, vol. 140, pp. 2003–13, nov 2012.

[21] M. R. Cayo and T. O. Talbot, "Positional error in automated geocoding of residential addresses," *International Journal of Health Geographics*, vol. 2, pp. 1–12, 2003.

[22] C. A. Davis and F. T. Fonseca, "Assessing the certainty of locations produced by an address geocoding system," *GeoInformatica*, vol. 11, no. 1, pp. 103–129, 2007.

[23] J. H. Ratcliffe, "Geocoding crime and a first estimate of a minimum acceptable hit rate," *International Journal of Geographical Information Science*, vol. 18, pp. 61–72, jan 2004.

[24] D. P. Johnson, A. Stanforth, V. Lulla, and G. Luber, "Developing an applied extreme heat vulnerability index utilizing socioeconomic and environmental data," *Applied Geography*, vol. 35, pp. 23–31, nov 2012.

[25] SmartyStreets, "USPS & International Address Verification - SmartyStreets."

[26] egon: Address Quality, "EGON - Company informations," 2019.

[27] EXPERIAN, "Address validation from Experian QAS," 2018.

[28] M. Wang, V. Haberland, A. Yeo, A. Martin, J. Howroyd, and J. M. Bishop, "A Probabilistic Address Parser Using Conditional Random Fields and Stochastic Regular Grammar," *IEEE International Conference on Data Mining Workshops, ICDMW*, pp. 225–232, 2017.

[29] R. G. Crowder, *Principles of Learning and Memory: Classic Edition*, vol. 2014. 2014.

[30] N. Reimers and I. Gurevych, "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging Nils," in *Ubiquitous Knowledge Processing Lab (UKP-DIPF)*, 2017.

[31] E. Ma, Xuezhe; Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," in *Language Tecnologies Institute*, 2016.

[32] G. Xiang, Bing; Kurata, "Leveraging Sentence-level Information with Encoder LSTM for Semantic Slot Filling," in *IBM Research*, 2016.

[33] B. Liu and I. Lane, "Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling."

[34] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," in *University of British Columbia; Honda Research Institute Japan CO*, no. 2003, 2014.

[35] F. Xu, G. Yi, W. Qi, and F. Zhen, "Research on Automatic Summary of Chinese Short Text Based on LSTM and Keywords Correction *," in *Tenth International Conference on Advanced Computational Intelligence (ICACI)*, no. 17, pp. 467–472, 2018.

[36] S. Pascual and A. Bonafonte, "Multi-output RNN-LSTM for multiple speaker speech synthesis and adaptation," in *European Signal Processing Conference (EUSIPCO)*, pp. 2325–2329, 2016.

[37] D. Wei, B. Wang, G. Lin, D. Liu, Z. Dong, H. Liu, and Y. Liu, "Research on Unstructured Text Data Mining and Fault Classification Based on RNN-LSTM with Malfunction Inspection Report," *Energies*, vol. 10, no. 406, 2017.

[38] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "SPOKEN LANGUAGE UNDERSTANDING USING LONG SHORT-TERM MEMORY NEURAL NETWORKS," in *Microsoft*, pp. 189–194, 2014.

[39] O. Morillot, L. Likforman-Sulem, and E. Grosicki, "New baseline correction algorithm for text-line recognition with bidirectional recurrent neural networks," *Journal of Electronic Imaging*, vol. 22, no. 2, p. 023028, 2013.

[40] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," 2015.

[41] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Systems With Applications*, vol. 72, pp. 221–230, 2017.

[42] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems (NIPS)*, pp. 3104–3112, 2014.

[43] G. Lewis, "Sentence Correction using Recurrent Neural Networks," pp. 1–7, 2015.

[44] J. Martens, "Generating Text with Recurrent Neural Networks," *Neural Networks*, vol. 131, no. 1, pp. 1017–1024, 2011.

[45] J. Li, K. Ouazzane, H. B. Kazemian, and M. S. Afzal, "Neural network approaches for noisy language modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 11, pp. 1773–1784, 2013.

[46] S. Zhu and K. Yu, "ENCODER-DECODER WITH FOCUS-MECHANISM FOR SEQUENCE LABELLING BASED SPOKEN LANGUAGE UNDERSTANDING," *Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering SpeechLab , Department of Computer Scie*, pp. 5675–5679, 2017.

[47] F. Liu, T. M. Hospedales, W. Yang, and C. Sun, "Semantic Regularisation for Recurrent Image Annotation," in *Computer Vision Foundation*, 2016.

[48] L. Liu, J. Shang, X. Ren, F. F. Xu, H. Gui, J. Peng, and J. Han, "Empower Sequence Labeling with Task-Aware Neural Language Model," 2017.

[49] E. Alpayding, *Introduction to Machine Learning Second Edition*, vol. 1107. 2010.

[50] D. P. Mandic and J. A. Chambers, *Recurrent Neural Networks for Prediction*. John Wiley & Sons, Ltd, aug 2001.

[51] B. V. Merri, "Learning Phrase Representations using RNN Encoderâ"Decoder for Statistical Machine Translation," 2013.

[52] J. Hochreiter, Sepp; Schmidhuber, "LONG SHORT-TERM MEMORY," *Neural Computation*, vol. 9, no. 8, pp. 1–32, 1997.

[53] Google Inc, "Google Maps Platform."

[54] OpenStreetMap, "Researcher Information OpenStreetMap," 2017.

[55] W. Cohen, P. Ravikumar, and S. Fienberg, "A Comparison of String Distance Metrics for Name-Matching Tasks William," *Software: Practice and Experience*, vol. 12, no. 1, pp. 57–66, 2003.

[56] P. Achananuparp, X. Hu, and X. Shen, "The evaluation of sentence similarity measures," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5182 LNCS, pp. 305–316, 2008.

[57] T. Kohonen and P. Somervuo, "Self-organizing maps of symbol strings," *Neurocomputing*, vol. 21, no. 1-3, pp. 19–30, 1998.

[58] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," *Lecture Notes in Engineering and Computer Science*, vol. 2202, no. May 2017, pp. 380–384, 2013.

[59] Knime.org — Open for innovation, "KNIME Analytics Platform," 2015.

[60] R. Hughey and A. Krogh, "Hidden markov models for sequence analysis: Extension and analysis of the basic method," *Bioinformatics*, vol. 12, no. 2, pp. 95–107, 1996.

[61] Superintendencia de Industria y Comercio, "Estudio económico del sector Retail en Colombia (2010-2012)," 2012.

[62] Departamento Adminisitrativo Nacional de Estadística (DANE), "Censo Nacional de Población y Vivienda 2018," 2018.

[63] F. Ricci, L. Rokach, B. Shapira, and P. Kantor, *Recommender Systems Handbook*. 2011.

[64] Tienda Registrada— Sabemos de Tiendas, "Noticias de la Tienda. Para la industria del consumo masivo," Tech. Rep. 48, Medellín, 2019.

[65] P. Jariha and S. K. Jain, "A state-of-the-art Recommender Systems: An overview on Concepts, Methodology and Challenges," *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, no. Icicct, pp. 1769–1774, 2018.

[66] S. van de Sanden, K. Willems, and M. Brengman, "In-store location-based marketing with beacons: from inflated expectations to smart use in retailing," *Journal of Marketing Management*, vol. 35, no. 15-16, pp. 1514–1541, 2019.