Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: PLN

Modelización de temas de llamadas en tiempo real

Borrador

Autor: Manuel E. Gómez Montero

Tutora UOC: Ana Valdivia Garcia

Tutor TE: Antonio Fernández Gallardo

Profesor: Jordi Casas

Madrid, 23 de diciembre de 2019

Resumen

Un call-center es el área de una empresa el cuál se encarga de recibir y transmitir llamadas desde o hacia clientes, socios comerciales u otras compañías externas. Debido a la gran cantidad de información que se transfiere en estos centros, resulta una tarea esencial optimizar el tiempo de respuesta para así reaccionar en tiempo real a las peticiones de los clientes y mejorar la percepción que estos tienen sobre la compañía.

Una manera de mejorar el rendimiento es detectar el tema de las llamadas mediante técnicas de *machine learning* dando la posibilidad a la empresa de reaccionar en tiempo real, en función de la temática que se este tratando en cada momento.

El sistema que se presenta en el documento nos permite, a partir de la transcripción de las llamadas al *call-center* de Telefónica España, descubrir en tiempo real la temática de las mismas. Esta modelización de *topics* se ha realizado utilizando métodos de Procesamiento de Lenguaje Natural y aprendizaje profundo. El sistema realiza la clasificación de las nuevas llamadas en tiempo real, permitiendo a los usuarios visualizar la evolución en la temática de las mismas y generar alertas en base a anomalías.

TODO Es un borrador volver al resumen una vez acabado el proyecto.

Palabras clave: “natural language processing”, “sentiment analysis”, “real time”, “call center”, “topic modeling”, “deep learning”

Índice general

Índice de figuras

Parte I

Introdución: objetivos, estado del arte y arquitectura global

Capítulo 1

Introducción

Este primer capítulo del trabajo tiene como objetivo presentar, a grandes rasgos, la propuesta (sección 1.1), los objetivos que pretendemos lograr (sección 1.3), la motivación que nos ha llevado a abordar este proyecto (sección 1.2) y un repaso a las tareas que serán necesarias para la ejecución del mismo (sección 1.4).

Por último, dedicaremos una sección que describa brevemente los diferentes apartados de los que constará el documento y el objetivo de cada uno (sección 1.5).

1.1. Descripción general de la propuesta

En los últimos años, la explosión ingente en la generación de datos y el avance en las capacidades tecnológicas que nos permiten recolectar, almacenar y procesar los datos generados; han provocado que empecemos a abordar el estudio de otro tipo de datos no estructurados que antes no se podían analizar como imágenes, textos, audios, etc. Como resultado, diferentes áreas del conocimiento (Procesamiento del Lenguaje Natural, Análisis de Imágenes) han experimentado un creciente interés tanto en la comunidad científica como en el mundo de los negocios.

Dentro de los datos no estructurados, una de las fuentes de información con mayor potencial en todas las grandes empresas que prestan servicio al

público general, son las llamadas que los clientes realizan a su *call-center*, ya que nos permiten obtener una idea de la percepción que los clientes tienen de nuestra empresa y de sus preocupaciones en cada momento.

La propuesta que pretendemos abordar en este trabajo consiste en extraer la temática de estas llamadas en el momento en el que son capturadas. Aunque actualmente esta captura se hace periódicamente pretendemos construir una solución que nos permita el tratamiento de las mismas en tiempo real o streaming, y de esta manera mejorar el rendimiento de estos centros.

Esta extracción en tiempo real nos permitirá conocer cómo evolucionan los temas que tratan nuestros clientes cuando llaman a nuestro *call-center* para así poder reaccionar inmediatamente ante una preocupación concreta.

1.2. Motivación

La motivación que nos ha llevado a acometer un proyecto de esta naturaleza viene originada por diferentes factores que están ligados tanto al negocio como a las capacidades técnicas disponibles en la empresa.

Por un lado, la capacidad de obtener la temática de las llamadas en tiempo real se presenta como una oportunidad de mejorar la operatividad de un *call-center* y por ende la satisfacción de los clientes, permitiéndonos entenderlos mejor y así reaccionar de una manera ágil a sus necesidades reales.

Desde el punto de vista técnico, también es el momento ideal para emprender este proyecto debido tanto a la disponibilidad periódica de transcripciones de las llamadas, que nos permiten ahorrarnos el paso de realizar un *Speech 2 Text* para obtener nuestro conjunto de datos; como al aumento de capacidades técnicas en la empresa que nos permitirán tanto entrenar nuestros modelos, como poder tratar y explotar los datos en tiempo real.

1.3. Objetivos

En este apartado definiremos los objetivos que se pretenden conseguir con este proyecto. Estos objetivos deben ser *SMART*, es decir:

- ***S**pecific*: Deben plantearse de una forma detallada y concreta.
- ***M**easurable*: Deben poder medirse con facilidad.
- ***A**chievable*: Deben ser objetivos realistas.
- ***R**elevant*: Tienen que ser relevantes para la empresa y ofrecernos un beneficio claro.
- ***T**imely*: Estos objetivos tienen que tener un tiempo establecido.

El objetivo general es optimizar el proceso de atención de llamadas en el call-center mediante técnicas de Procesamiento del Lenguaje Natural y Aprendizaje Profundo. Concretamente, los objetivos específicos que se pretenden conseguir con este proyecto son:

- **Construir un modelo que nos permita extraer la temática de las llamadas** a partir de su transcripción a texto. Este objetivo debemos alcanzarlo en la fase de modelado y podremos medir su éxito atendiendo al porcentaje de llamadas que podamos clasificar correctamente en un proceso de test. Se trata del objetivo principal del proyecto.
- Desarrollar un mecanismo que nos permita **extraer esta temática para nuevas llamadas en tiempo real**. De este modo tendremos un sistema vigente cuando la frecuencia en la recepción de las llamadas aumente. Este objetivo se deberá alcanzar en la fase de productivización.
- Disponer de una **visualización en tiempo cuasi real** para que pueda visualizarse la evolución de las temáticas a lo largo del tiempo. Este objetivo se deberá alcanzar en la fase de productivización.
- Proporcionar un **sistema de alertado** que nos permita detectar anomalías en el número de llamadas que se reciben de un determinado tema. Este objetivo se deberá alcanzar en la fase de productivización.

En las conclusiones de este proyecto se evaluará el éxito o fracaso del mismo en función del grado de cumplimiento de estos objetivos.

1.4. Tareas y planificación

El proyecto se llevará a cabo desde el 16 de Septiembre hasta el 20 de Febrero. Para poder abordar la ejecución del mismo se han extraído las siguientes tareas principales:

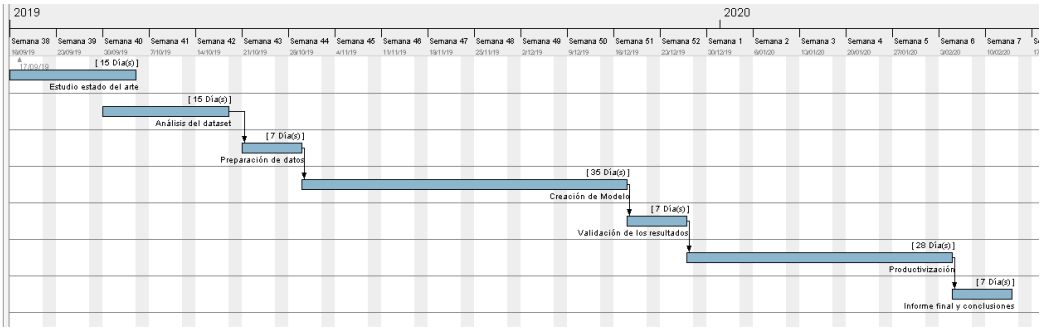

Figura 1.1: Diagrama de Gantt

- **Estudio estado del arte**: En esta fase se realizará una prospección para conocer el estado del arte en todos los puntos relacionados con el proyecto: Procesamiento del Lenguaje Natural, tecnologías de tratamiento de datos en tiempo real y *Big Data*.

- **Análisis del *dataset***: El propósito de esta tarea es entender el *dataset* y estudiar las posibilidades del mismo.

- **Preparación del *dataset***: Una vez realizado el estudio del *dataset* es necesario realizar labores de limpieza y transformación de los datos de modo que estos datos sean válidos para nuestro objetivo.

- **Creación del modelo**: En esta fase se procederá a la creación de un modelo capaz de obtener los temas de los que habla una determinada llamada. Este modelo será el *core* de nuestro proyecto.

- **Validación de los resultados**: Una vez entrenado el modelo será necesario validar los resultados obtenidos para poder evaluar la bondad de nuestro modelo.

- **Productivización**: El trabajo no acaba con la creación de un buen modelo que nos permita extraer los temas de nuestras llamadas. Este modelo tendrá que ser puesto en producción y permitir al usuario final extraer los temas de las llamadas en tiempo real y darle la opción de crear alarmas basadas en la variación del número de eventos (llamadas) de un determinado tema.

- **Informe final y conclusiones**: Por último, una vez llevado a a producción nuestro modelo, se realizará un informe final donde, entre otros puntos, se evaluarán los resultados obtenidos y se extraerán conclusiones y pasos futuros.

Figura 1.2: Fases del modelo CRISP-DM

Estas fases están basadas en el estándar ***CRISP-DM*** ([7]), añadiendo una última tarea para nuestro informe final, CRISP-DM nos proporciona una descripción del ciclo de vida de los proyectos de minería de datos de un modo bastante similar al que se aplica en los modelos de ciclo de vida de desarrollo *software*.

En la Figura 1.2 se observa el diseño de este modelo y cómo representa el ciclo de vida de un proyecto de minería de datos. En la imagen podemos ver en primer lugar un círculo exterior que refleja la naturaleza cíclica de los proyectos de minería de datos, además vemos cómo la secuencia de tareas no es rígida, pudiendo saltar hacia adelante o atrás entre tareas. En la gráfica se representan mediante flechas las dependencias más importantes y usuales entre tareas.

En nuestro desarrollo usaremos este modelo, aunque en el diagrama de la Figura 1.1 aparezca una secuencia de tareas más rígida, será usual, por ejemplo, el salto recíproco entre las fases de preparación de los datos y creación del modelo.

1.5. Estructura del documento

TODO Hacer repaso breve de los apartados del documento final.

Capítulo 2

Estado del Arte

El objetivo de este apartado es hacer un recorrido por el estado del arte relacionado con el proyecto, este recorrido lo enfocaremos desde tres puntos de vista diferentes:

- **Procesamiento del Lenguaje Natural**: En la sección 2.1 nos centraremos en el procesamiento del lenguaje natural y su evolución a lo largo del tiempo.

- ***Deep Learning* y aplicación al Procesamiento del Lenguaje Natural**: En la sección 2.2 pondremos foco en el *Deep Learning*, sus ventajas y cómo se están aplicando estos métodos al procesamiento del lenguaje natural.

- ***Big Data* y *Fast Data***: Por último, en la sección 2.3, haremos un repaso a la evolución del *Big Data* y cómo la tendencia actual es realizar el procesamiento en tiempo real mediante *Fast Data*.

Por último, una vez analizados los diferentes puntos de vista, en la sección 2.4 enumeraremos trabajos anteriores relacionados con nuestro proyecto. Estos trabajos nos serán de utilidad para justificar la realización de nuestro proyecto y su viabilidad.

2.1. Procesamiento de lenguaje natural

2.1.1. Historia

Para hablar de los orígenes del Procesamiento del Lenguaje Natural (a partir de ahora se usarán indistintamente las siglas PLN) tal y como lo conocemos, tendríamos que remontarnos a los años 50, concretamente al artículo "*Computing Machinery and Intelligence*" escrito por Alan Turing [31]. En este artículo aparece el PLN dentro del campo de la inteligencia artificial y se presenta por primera vez el conocido "Test de Turing". Este test convirtió la pregunta abstracta de "¿Son capaces de pensar las máquinas?" en un juego llamado: "*The Imitation Game*". El juego propuesto inicialmente, de forma muy resumida, consiste en ver si una persona (interrogador) interrogando a dos personas (un hombre y una mujer), era capaz de descubrir el sexo de cada una; la modificación del mismo sustituye las dos personas de distinto sexo por una persona y una máquina y el interrogador debe ser capaz de descubrir si las preguntas están siendo respondidas por un humano o una máquina. En el caso de que no sepa discernir, la computadora gana la partida. Podemos encontrar más información al respecto en el libro [32].

A partir de los avances de Turing y hasta los años 80 el crecimiento en el campo del PLN se produjo principalmente con la creación de complejos sistemas basados en reglas escritas a mano. Fue en esta década cuando empezamos a vivir la incorporación de algoritmos de *Machine Learning* enfocados al procesamiento del lenguaje natural. Este hecho se vio motivado principalmente por el increíble avance en la capacidad de cómputo, ya predicho por la ley de Moore, y por la aplicación de teorías ya existentes como los trabajos de Chomsky.

Desde el comienzo de la aplicación de modelos de *Machine Learning*, y de nuevo motivados por el crecimiento de la capacidad computacional de los sistemas actuales, se ha pasado de utilizar árboles de decisión, que creaban de manera automática reglas similares a las que se venían creando manualmente, a los modelos de *deep learning* que están en auge en la última década.

2.1.2. Aplicaciones

En el apartado anterior hicimos referencia a "*The Imitation Game*" como inicio de lo que hoy conocemos como procesamiento del lenguaje natural, sin embargo, las aplicaciones en este campo han crecido de forma vertiginosa en estos 70 años, principalmente en las últimas décadas. Hoy en día, si tuviéramos que contestar a la pregunta: "¿son capaces de pensar las máquinas?", implicaría algo más que superar el test de Turing. Mirando a nuestro alrededor nos encontraríamos con asistentes de voz como Alexa o Siri que, no solo contestan a nuestras preguntas, si no que realizan un trabajo de pasar nuestra voz a texto (*Speech to Text*) y de nuevo el texto resultante a voz (*Text to Speech*). Nos encontraríamos también con sistemas capaces de realizar traducciones simultáneas, otros capaces de autocompletar textos, de identificar preguntas y respuestas, de clasificar textos de acuerdo a temas o autores, incluso de analizar sentimientos positivos o negativos teniendo como entrada un texto u opinión.

Según [13] todos estos problemas tan diversos podríamos clasificarlos según en el punto del análisis que nos centremos:

- **Análisis de palabras**: En este tipo de problemas se pone foco en las palabras, como pueden ser "perro", "hablar", "piedra" y necesitamos decir algo sobre ellas. Por ejemplo: "¿estamos hablando de un ser vivo?", "¿a qué lenguaje pertenece?", "¿cuáles son sus sinónimos o antónimos?". Actualmente este tipo de problemas son menos frecuentes, ya que normalmente no pretendemos analizar palabras aisladas sino que es preferible basarse en un contexto.

- **Análisis de textos**: En este tipo de problemas no trabajamos solo con palabras aisladas, sino que disponemos de una pieza de texto que puede ser una frase, un párrafo o un documento completo y tenemos que decir algo sobre él. Por ejemplo: "¿se trata de spam?", "¿qué tipo de texto es?", "¿el tono es positivo o negativo?", "¿quién es su autor?". Este tipo de problemas son muy comunes y nos vamos a referir a ellos como **problemas de clasificación de documentos**.

- **Análisis de textos pareados**: En esta clase de análisis disponemos de dos textos (también podrían ser palabras aisladas) y tenemos que decir algo sobre ellos. Por ejemplo, "¿los textos son del mismo autor?", "¿son pregunta y respuesta?", "¿son sinónimos?" (para el caso de palabras aisladas).

- **Análisis de palabras en contexto**: En estos casos de uso, a diferencia del primer análisis que trataba unicamente con palabras aisladas, tenemos que clasificar una palabra en particular en función del contexto en el que se encuentra.

- **Análisis de relación entre palabras**: Este último tipo de análisis tiene como objetivo deducir la relación entre dos palabras existentes en un documento.

Dependiendo del problema que queramos abordar usaremos un tipo de características del lenguaje u otro, por ejemplo, es usual que si estamos analizando palabras aisladas nos centremos en las letras de una palabra, sus prefijos o sufijos, su longitud, la información léxica extraída de diccionarios como *WordNet* [11], etc. En cambio, si estamos trabajando con texto, lo normal es que nos fijemos en otros conceptos estadísticos como el histograma de las palabras dentro del texto, ratio de palabras cortas vs largas, número de veces que aparece una palabra en un texto comparado con el resto de textos, etc.

El proyecto que se presenta en este documento está centrado en el análisis de textos, concretamente en extraer los temas de un documento (o llamada). Este tipo de problemas se conoce como modelización de *topics*.

En el siguiente punto de este apartado nos centraremos en algunos modelos y avances en este área que puedan servirnos de apoyo para nuestro proyecto.

2.1.3. Modelización de temas

La modelización de topics hace referencia a un grupo de algoritmos de *Machine Learning* que infieren la estructura latente existente en un grupo de

documentos.

Aunque la mayoría de los algoritmos de modelización son no supervisados, al igual que los algoritmos tradicionales de *clustering*, existen también algunas variantes supervisadas que necesitan disponer de documentos etiquetados.

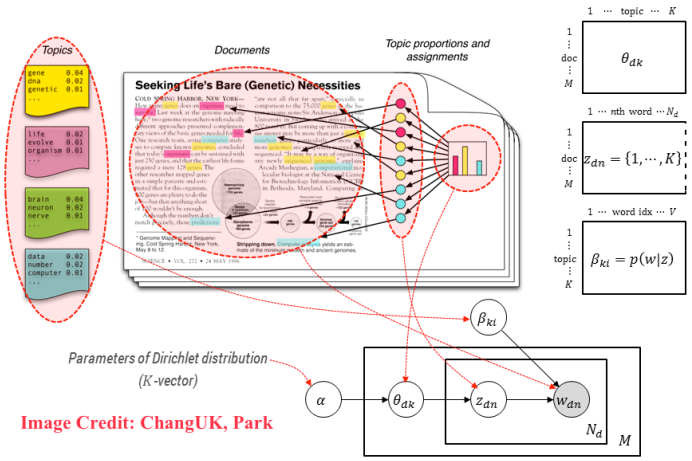

Figura 2.1: Representación gráfica de LDA. Fuente [22]

Quizás el algoritmo más conocido para la modelización de topics sea el *Latent Dirichlet Allocation* (normalmente conocido por su acrónimo, LDA). LDA fué presentado en 2003 en el artículo [6] y podemos ver una representación gŕafica del mismo en la figura 2.1. Este algoritmo no supervisado asume que cada documento es una distribución probabilística de *topics* y cada *topic*, a su vez, es una distribución de palabras del documento. LDA usa una aproximación llamada "*bag of words*", en la que cada documento es tratado como un vector con el conteo de las palabras que aparecen en el mismo. La principal característica de LDA es que la colección de documentos comparten los mismos topics, pero cada documento contiene esos topics en una proporción diferente.

A partir de LDA surgieron numerosas variantes que repasaremos de forma breve, por ejemplo, en el mismo año de la creación de LDA y también presentado por los mismos autores en [3], surgió una **variante jerárquica** que permitía representar los *topics* jerárquicamente. En 2006 en [5] se desa-

rrolla un modelo LDA dinámico denominado DTM (*Dinamic Topic Model*), en el que se introduce la variable temporal y los *topics* pueden ir cambiando a lo largo del tiempo. En el artículo [4] nos encontramos con otra variante de LDA llamada CTM (*Correlated topic model*) que nos permite encontrar correlaciones entre *topics*, ya que algunos temas es probable que sean más similares entre sí. Por último, nos encontramos con una variante de LDA denominada ATM (*Author-Topic Model*) propuesta por Michal Rosen-Zvi en su artículo [29] y desarrollada por el mismo en 2010, en la que los documentos son una distribución probabilística tanto de autores como de *topics*.

Podemos encontrar un resumen más completo del estado del arte en cuanto a la modelización de *topics* en el artículo [20].

2.2. Deep Learning y aplicación al PLN

El objetivo de esta sesión es entender el concepto de *Deep Learning* y analizar el estado del arte del *Deep Learning* aplicado al Procesamiento del Lenguaje Natural. Para poder entender el *Deep Learning* es conveniente entender los modelos de aprendizaje supervisados y saber qué provoca su aparición y popularidad de los últimos años. Posteriormente nos centraremos en los fundamentos del *Deep Learning* y cómo es utilizado en la representación de palabras. Por último, comentaremos algunas arquitecturas especializadas y su aplicación en el ámbito del Procesamiento del Lenguaje Natural.

2.2.1. Aprendizaje supervisado

El aprendizaje supervisado consiste en aprender una función a través de un conjunto de datos llamados de entrenamiento, mediante la cual podamos obtener una salida a partir de una determinada entrada. Se espera que esta función, una vez realizado el entrenamiento, sea capaz de producir una salida correcta incluso para datos nunca vistos. Es muy habitual el uso de estos tipos de algoritmos para casos de clasificación y/o predicción.

Buscar entre todas las posibles infinitas funciones para encontrar la que mejor se adapte a nuestro conjunto de datos es un trabajo inviable, es por

ello que normalmente se realiza la búsqueda entre un conjunto de funciones limitadas. En un primer lugar, y hasta hace aproximadamente una década, los modelos más populares de aprendizaje supervisado fueron los modelos lineales, provenientes del mundo de la estadística, estos modelos son fáciles de entrenar, fáciles de interpretar y muy efectivos en la práctica.

A partir de entonces, y motivado en parte por el aumento en las capacidades de cómputo, surgen otros modelos como las máquinas de vectores de soportes (*Support Vector Machines*, SVMs) o las redes neuronales, en las que nos centraremos en el siguiente apartado.

2.2.2. Deep Learning

Dentro del *Machine Learning* y usualmente relacionado con el aprendizaje supervisado, nos encontramos con un sub-campo denominado **Deep Learning** que utiliza las redes neuronales para la creación de modelos.

Como su nombre indica las redes neuronales consisten en unidades de cómputo llamadas neuronas que están interconectadas entre sí. Una neurona es una unidad de cómputo que posee múltiples entradas y una salida, esta neurona multiplica cada entrada por un peso para posteriormente realizar una suma y, por último, aplicar una función de salida no lineal. Si los pesos se establecen correctamente y tenemos un número suficiente de neuronas, una red neuronal puede aproximar a un conjunto muy amplio de funciones matemáticas.

En las redes neuronales, las neuronas suelen organizarse por capas que se encuentran conectadas entre sí. Mientras más capas tengamos, más características podremos extraer de nuestros datos de entrada y podremos aproximar un mayor número de funciones (sin perder de vista el sobrentrenamiento).

El primero y más simple de los tipos de redes neuronales es el denominado *Feed Forward Neural Network* (**FFNN**), este tipo de redes recibe este nombre porque no existen ciclos entre sus neuronas y las conexiones se realizan siempre desde las capas anteriores a las capas posteriores.

Una de las arquitectura más comunes de FFNN es el preceptrón multi-

capa (en inglés multilayer perceptron o **MLP**). Esta arquitectura contiene tres o más capas de neuronas totalmente conectadas, es decir, la salida de una neurona de una capa se encuentra conectada a la entrada de todas las neuronas de la siguiente capa. Las capas de una arquitectura MLP son:

- **Capa de entrada**: Se trata de la capa en la que introduciremos los datos en la red. Esta capa carece de procesamiento.
- **Capas ocultas**: Son las capas intermedias, cuyo número puede variar, tienen como entrada la salida de las neuronas de la capa anterior y su salida alimenta a las neuronas de la capa posterior.
- **Capa de salida**: Los valores de salida de las neuronas de esta capa se corresponden con la salida de la red.

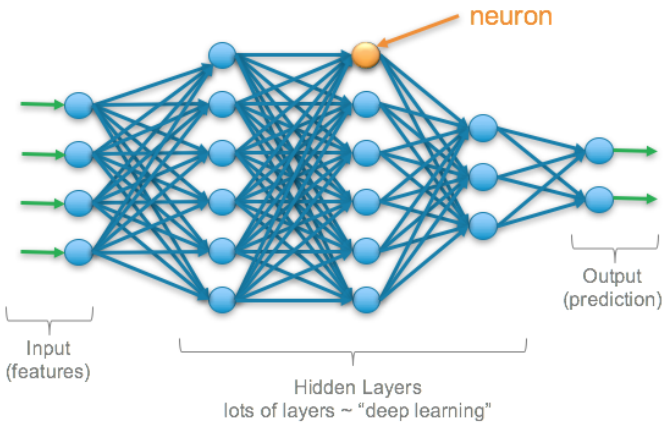

Figura 2.2: Ejemplo de arquitectura MLP. Fuente [28]

Hablamos que una red es profunda cuando contiene un gran número de capas, por ello el término de *Deep Learning*. En la Figura 2.2 observamos un ejemplo de arquitectura MLP y como la denominamos *Deep Learning* al crecer el número de capas ocultas.

2.2.3. Representación de palabras en PLN

Es usual, en el ámbito del reconocimiento de imágenes, utilizar información acerca de la dimensionalidad de las mismas. Este tipo de información nos

permite extraer características teniendo en cuenta los píxeles vecinos. Tradicionalmente, en el ámbito del Procesamiento del Lenguaje Natural, esto no se ha llevado a cabo debido a que cada palabra (o n-grama) se trataba como una entidad aislada utilizando una codificación de las palabras denominada **one-hot encoding**.

En cambio, existe otro método de representar las palabras en el lenguaje natural que sí es capaz de captar la "dimensionalidad" de una forma similar a como lo realizamos en las imágenes. Este modo, conocido como ***word embedding***, deja de tratar la palabra como un ente aislado y es capaz de captar el significado de la misma, esta representación se denomina distribuida y consiste en convertir las palabras en vectores en los que cada dimensión capte características diferentes de las palabras. Este tipo de representaciones dará lugar a vectores similares para palabras semánticamente parecidas.

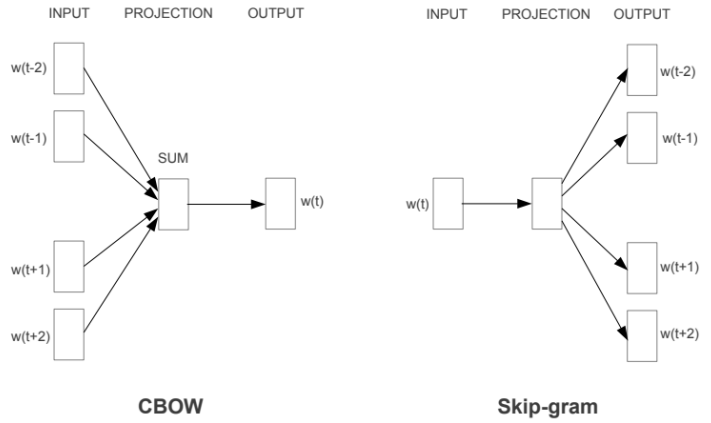

Figura 2.3: Arquitecturas CBOW y Skip-gram. Fuente [23]

Una de las soluciones más populares que nos permiten convertir una palabra a un vector (*word2vec*) que contenga información de la palabra en función del contexto se detallan en el artículo [23] . Aquí se presentaron dos modelos llamados **Skip-Gram** y **CBOW** cuya arquitectura podemos ver en la Figura 2.3. Estos modelos utilizan redes neuronales para predecir una palabra en función de su contexto o el contexto en función de una palabra, el vector que se utiliza para representar la palabra es el vector de pesos de la capa oculta.

2.2.4. Arquitecturas especializadas

Después de introducir las redes neuronales y el modo en el que podemos representar las palabras, frases o documentos para ser usados como entrada en nuestro modelo; vamos a centrarnos en comentar dos tipos de redes neuronales que se usan de manera tradicional en tareas de Procesamiento de Lenguaje Natural.

Los dos tipos de redes neuronales que comentaremos son las redes neuronales convolucionales y las redes neuronales recurrentes. Haremos una introducción a cada una de ellas, comentaremos sus aplicaciones al PLN y sus ventajas e inconvenientes con respecto a otro tipo de métodos.

Redes neuronales convolucionales

Las redes neuronales convolucionales (llamadas usualmente CNN, por su nombre en inglés ***C**onvolutional **N**eural **N**etworks*) son un tipo de redes neuronales que deben su nombre a la operación matemática de convolución que realizan. Esta operación consiste en aplicar a una matriz de entrada multidimensional, un filtro o kernel también multidimensional y obtener una salida, también denominada mapa de características. En la Figura 2.4 podemos ver una representación gráfica de esta operación para un ejemplo de 2 dimensiones.

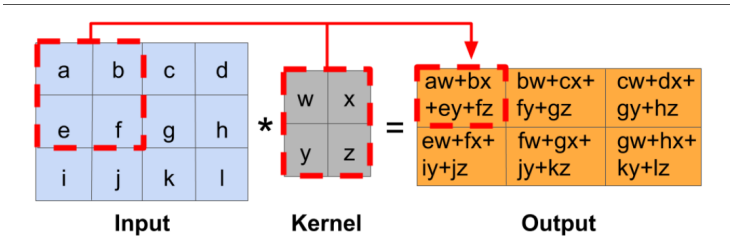

Figura 2.4: Ejemplo de una convolución de dos dimensiones. Fuente [2]

Es usual en las redes convolucionales utilizar diferentes kernels sobre una misma entrada, obteniendo diferentes salidas que permitan reconocer distintos patrones. Los pesos del kernel junto con el sesgo, son los parámetros

que serán necesarios calcular en el entrenamiento y en el caso de las redes convolucionales, se denominan mapas de características.

Aunque la convolución simple que comentamos es la operación básica en las redes convolucionales, es usual añadirle algunas variantes (o configurarla con algunos parámetros) que nos permitan variar la dimensión de salida una vez hemos aplicado la convolución; algunas de estas variaciones más comunes son el ***zero padding*** y la **convolución por pasos**.

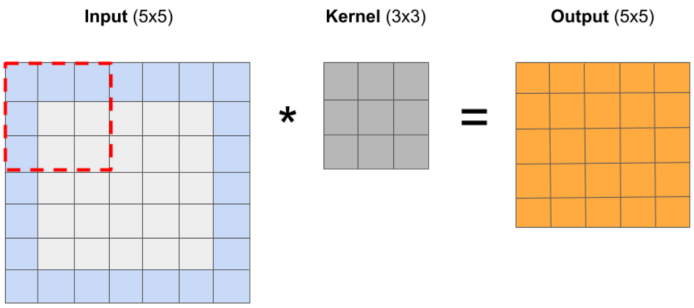

Figura 2.5: Ejemplo de aplicación de *zero padding* para mantener la dimensionalidad. Fuente [2]

Como observamos en la Figura 2.4, al aplicar el kernel a los datos de entrada estamos reduciendo la dimensionalidad de la salida. A menudo es posible que esto no nos interese y queramos mantener la dimensionalidad en la salida; para ello recurrimos a un método denominado *zero padding* que consiste en añadir '0s' en los bordes de nuestra entrada con el objetivo de preservar la dimensionalidad en la salida. En la Figura 2.5 podemos ver un ejemplo de aplicación de *zero padding*.

Por otro lado, es también posible que queramos reducir aún más la dimensión de salida, principalmente por un tema de eficiencia y reducción de los tiempos de ejecución, a costa de perder información de algunas características en la salida. Para ello podemos utilizar la convolución por pasos (o *strided* por su nombre en inglés). Este método consiste en aplicar el kernel realizando saltos en lugar de hacerlo sobre celdas consecutivas, tal y como podemos ver en la Figura 2.6.

El proceso explicado anteriormente correspondería con una capa convolucional, que son el corazón de las redes neuronales convolucionales, sin em-

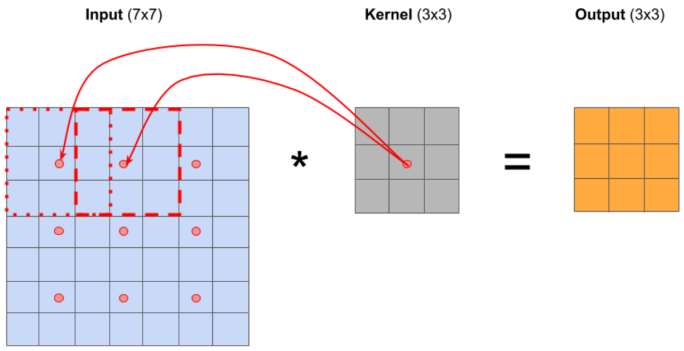

Figura 2.6: Ejemplo de aplicación de convolución por pasos para reducir la dimensionalidad. Fuente [2]

bargo, en una red neuronal convolucional estas capas coexisten con otro tipo de capas que nos ayudaran a mejorar nuestros modelos. Las más usuales son:

- Capa de agrupamiento (*polling* en inglés): El objetivo de esta capa es agrupar un conjunto de salidas para obtener un único valor. Al conjunto de valores de entrada (seleccionado de nuevo con un filtro) se le aplica una función para obtener un único valor. Aunque se pueden utilizar diferentes funciones, como puede ser la media, lo más usual es aplicar la función de máximo (*max-polling*). Es habitual, intuir que usando esta función nos estamos quedando con las características más relevantes de cada cuadrante (del tamaño del filtro) de la entrada. En la Figura 2.7 podemos ver un ejemplo de agrupamiento utilizando la función de máximo.

- Capa totalmente conectada: Hemos visto ejemplos de capas totalmente conectadas al introducir las redes neuronales, esta capas usualmente se usan al final de nuestra red para tareas de clasificación, teniendo la última capa un número de neuronas igual al número de clases que pretendemos clasificar.

- Capa RELU: Si observamos la descripción de la operación de convolución nos damos cuenta de que se trata de una operación totalmente lineal, es por ello que después de cada capa de convolución es usual

agregar una capa no lineal (también llamada capa de activación). Aunque se pueden utilizar otras funciones como la tangente o la función sigmoide, lo más usual es utilizar la función RELU.

- Capa de Dropout: Esta capa tiene como funcionalidad prevenir el sobreentrenamiento en las redes neuronales, desactivando un número aleatorio de entradas de la capa, forzando a la red a ser redundante y permitiendo dar una clasificación correcta sin tener todas las entradas activas.

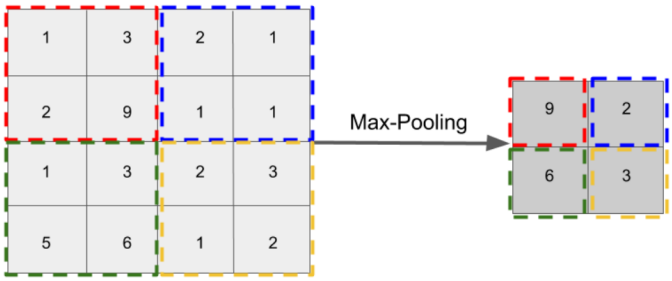

Figura 2.7: Ejemplo de aplicación de *max-pooling*. Fuente [2]

Tras esta visión general sobre las redes neuronales generales, podemos enumerar las ventajas que conllevan:

- Por un lado, aunque no hemos entrado en detalles sobre el proceso de entrenamiento de las redes convolucionales, se puede intuir que el aplicar un mismo kernel sobre toda la entrada provoca que el número de parámetros a aprender (los valores del kernel) con respecto a una red totalmente conectada será mucho menor. Esto provoca una **reducción del tiempo de entrenamiento necesario**.

- Por otro lado, el hecho de compartir el kernel provoca que podamos **capturar una misma característica en la entrada a pesar de su traslación**. Por ejemplo, si estamos detectando un objeto en una imagen un modelo convolucional podrá detectar ese objeto a pesar de su movimiento por la imagen.

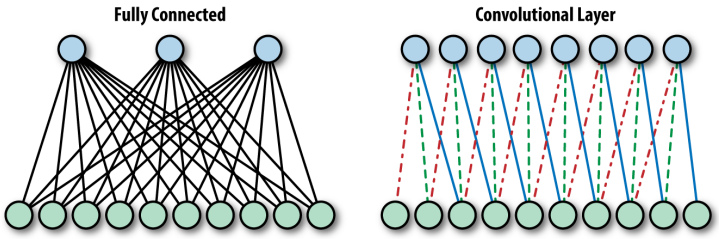

Figura 2.8: Comparación capa tradicional totalmente conectada con capa convolucional. Fuente [15]

Sin embargo, las redes neuronales convolucionales deben usarse en datos que contengan coherencia local ya que esa es su fortaleza. **En datos sin coherencia local las redes neuronales convolucionales no lograrían obtener un buen rendimiento** como las redes neuronales tradicionales vistas anteriormente. Si observamos la Figura 2.8 podemos ver la diferencia entre las capas de ambos tipos de redes y cómo la capa convolucional se centra más en las estructuras locales.

Encontramos una explicación más profunda sobre las redes convolucionales y su uso general en [2]. Aunque, como podemos imaginar, el uso más extendido de este tipo de redes es para el tratamiento de imágenes, nosotros nos centraremos en tener una breve visión de su aplicación al Procesamiento del Lenguaje Natural, que también puede ampliarse en [13].

Al aplicar las redes tradicionales al PLN, solemos ignorar el orden en el que las palabras aparecen en las frases, o las frases en el documento, siguiendo una aproximación CBOW; esto suele ser problemático a la hora de realizar, por ejemplo, un análisis de sentimientos ya que no es lo mismo encontrar la palabra “malo” aislada que el bigrama “no malo”. Aunque el uso de bigramas y N-gramas de mayor orden puede mejorar esta situación el coste puede volverse inasumible.

Es en este ámbito dónde las redes neuronales convolucionales pueden ser de gran ayuda, ya que gracias a la capacidad comentada para detectar estructuras locales serían capaces de identificar estos N-gramas de forma automática para ser usados posteriormente en tareas predictivas.

Redes neuronales recurrentes

Otro de los modelos usualmente usados en tareas de Procesamiento del Lenguaje Natural son las redes neuronales recurrentes, (llamadas usualmente RNN, por su nombre en inglés ***R**ecurrent **N**eural **N**etworks*). Hasta ahora todos los modelos de redes neuronales que hemos citado funcionaban siempre en una dirección, las neuronas de una capa anterior producían una salida que era la encargada de activar las neuronas de la capa posterior. En las redes recurrentes veremos que las salidas de una neurona en una capa posterior pueden tener una conexión con una neurona de una capa anterior. Esto crea una especie de **memoria** que nos permite modificar la respuesta de la red en función de los datos que se hayan procesado anteriormente (incluso reaccionar con datos que lleguen posteriormente).

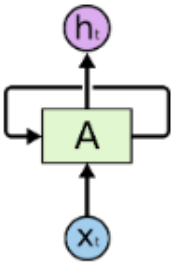

Figura 2.9: Ejemplo RNN. Fuente [27]

Las conexiones en una red neuronal recurrente pueden tener muchas variaciones por lo que es usual hablar del concepto de **celda**. Una celda suele tener como entrada los valores de la secuencia y el estado de la red neuronal en el paso anterior; y como salida la respuesta de la red neuronal a dicha entrada y el estado de la red neuronal en el paso actual.

En la Figura 2.9 observamos un ejemplo de red neuronal recurrente en el que tenemos como entrada el valor de la secuencia x_t y el estado de la red en el paso anterior (conexión en bucle). Producimos una respuesta h_t y un estado (conexión en bucle). Sin embargo, posiblemente esta representación sea algo más confusa para comprender su funcionamiento que si procedemos a "desenrollar" la red neuronal haciéndola más similar a los modelos vistos hasta ahora. En la Figura 2.10 podemos ver el resultado de "desenrollar"

la red; hay que tener en cuenta con esta representación que los parámetros usados por cada celda son exactamente los mismos.

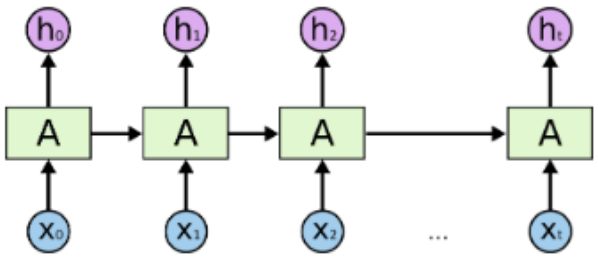

Figura 2.10: Ejemplo RNN "desenrollada". Fuente [27]

Aunque no entraremos en detalles sobre el entrenamiento de redes neuronales, es importante saber que existen dos problemas diferentes provocados ambos por usar los mismos parámetros en todas las celdas que provocan la inestabilidad durante el proceso de entrenamiento. Estos problemas son la **desaparición del gradiente**, que ocurre al multiplicar el gradiente consigo mismo múltiples veces cuando este es menor que 1, y la **explosión del gradiente**, que ocurre por el mismo motivo cuando este es mayor que 1.

Para mitigar estos problemas es importante el diseño de las celdas, a continuación veremos de manera resumida los dos tipos de celdas más usados en las redes neuronales recurrentes. Las celdas que comentaremos están compuestas por diferentes mecanismos internos, denominados puertas, que gestionan el flujo de información a lo largo de la misma.

El primer tipo de celdas son las celdas *Long Short Term Memory* (**LSTM**). Este tipo de celdas se comportan bien en situaciones que queremos encontrar patrones entre registros que se encuentran separados en la secuencia, esto es algo muy usual, por ejemplo, en el caso de PLN cuando en una misma frase una palabra hace referencia a otra que apareció a una distancia de varias palabras.

Para conseguir este objetivo, parece evidente que es importante controlar la memoria en cada una de las celdas. Una celda LSTM realiza esta tarea con las siguientes puertas:

- Puerta de entrada: Controla que información se añade a la memoria de

la red.

- Puerta de olvido: Controla, a partir de la memoria del paso anterior y de la entrada, qué información debe conservarse en la memoria.
- Puerta de salida: Es la encargada de calcular la salida de la red, h_t, en el paso actual.

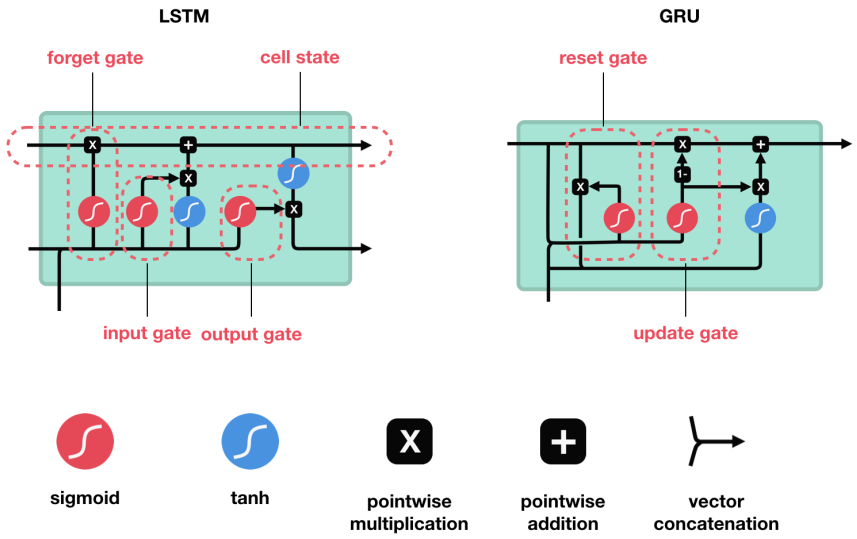

Figura 2.11: Arquitectura celdas LSTM y GRU. Fuente [26]

Debido al éxito de las celdas LSTM, y al constante esfuerzo de optimización de las mismas, han surgido diferentes variantes. La más conocida de todas son las *Gated Recurrent Unit* (**GRU**) introducidas en 2014. La celda GRU es una simplificación de la celda LSTM y produce unos rendimientos bastante similares con un menor coste. De manera muy resumida una celda GRU se compone de:

- Puerta de reset: Permite seleccionar que información de la memoria va a ser utilizada en un paso concreto.
- Puerta de actualización: Realiza la función de las puertas de olvido y de entrada que hemos visto en las celdas LSTM.

Aunque no hemos entrado en el detalle del funcionamiento de cada una de las puertas, en la Figura 2.11 podemos ver la arquitectura completa de ambos tipos, y observar la mayor simplicidad de las celdas GRU frente a las LSTM.

Como podemos imaginar, las redes neuronales recurrentes son ampliamente usadas en el mundo del Procesamiento del Lenguaje Natural, debido a que una palabra no es otra cosa que una secuencia de letras, una frase a su vez se trata de una secuencia de palabras y un documento una secuencia de frases. Alguno de los usos de las RNN en este ámbito son:

- **Análisis de sentimientos**: Detectar el sentimiento positivo o negativo de un texto utilizando únicamente la última salida de la red. En [19] tenemos un ejemplo de análisis de sentimientos utilizando redes recurrentes con LSTM sobre los datos de una red social China.

- **Generador de texto**: Predecir la siguiente palabra de una secuencia, utilizando la salida de cada una de las celdas. Podemos ver una comparación de métodos para generar textos de diferentes temáticas para poder entrenar posteriormente modelos *Deep Learning* en [1]. Entre los métodos de la comparación se encuentran las redes neuronales recurrentes con celdas GRU y con celdas LSTM.

- **Traductores**: Traducción automática de textos entre idiomas, llamado *Neural Machine Translation* (NMT) cuando se utilizan redes neuronales. Quizás el mayor caso de éxito en este punto es el *Google's Neural Machine Translation System* [33] cuya base es una red neuronal profunda construida con celdas LSTM.

Una vez analizadas, de manera global, las redes neuronales recurrentes podemos ver que nos ofrecen **numerosas ventajas al trabajar con secuencias** como hacemos en el ámbito del PLN. Entre ellas podemos ver, que vamos incluso más allá que con las CNN analizando la relación entre las diferentes palabras, fundamentalmente con las celdas LSTM y GRU, pudiendo detectarlas entre palabras que estén alejadas entre sí y no ceñirnos al tamaño del kernel. Sin embargo, como hemos visto, y aunque sean mitigados

por el uso de celdas LSTM y GRU, las redes neuronales recurrentes presentan **problemas durante el entrenamiento** tanto de desvanecimiento como de explosión del gradiente.

2.3. *BigData* y *Fast Data*

A lo largo de esta sección intentaremos tener una visión general del *Big Data* y su evolución a lo largo del tiempo hasta llegar al *Fast Data*. Posteriormente veremos las arquitecturas más usadas en el mundo del *Big Data*.

2.3.1. Evolución: del *Big Data* al *Fast Data*

El primer uso del término *Big Data* se da en un artículo de Michael Cox y David Ellsworth de la NASA publicado en 1997 [8], donde hacen referencia a la dificultad de procesar grandes volúmenes de datos con los métodos de la época. Sin embargo, fue en 2001 cuando encontramos la definición más conocida y aceptada de *Big Data* hecha por el analista Laney Douglas en su artículo "*3D Data Management: Controlling Data Volume, Velocity y Variety*" [17] en el que se hacía referencia a las ya "famosas" tres *V*s:

- **V**olumen: Cada vez los volúmenes de datos son mayores.
- **V**elocidad: Es cada vez mayor la velocidad con la que se generan los datos.
- **V**ariedad: Dejamos de tener únicamente datos completamente estructurados para trabajar con datos no estructurados y/o semi-estructurados.

Google, como es obvio, también se enfrentó a un importante problema a la hora de procesar la ingente cantidad de datos que generaba día a día y que no podían ser procesados de manera eficiente con el *software* existente, es por ello que en el año 2003 presenta en [12] su *"Google File System"* (GFS) y un año después *Map Reduce* [9], estas dos capas de almacenamiento y procesamiento distribuido dieron lugar al nacimiento de lo que hoy conocemos como ***Big Data***.

Sin embargo, estas aportaciones no empezaron a tomar una repercusión relevante fuera de Google hasta el nacimiento del *framework* Hadoop en 2006, un ecosistema con una gran cantidad de servicios pero cuya base fue Map Reduce y HDFS (basado en GFS). La complejidad del ecosistema *Hadoop* hizo que éste no empezara a aparecer en la mayoría de las empresas hasta la creación de la compañía *Cloudera* en 2009, que empezó a empaquetar los diferentes componentes del ecosistema *Hadoop*, ofreciendo distribuciones estables y soporte para sus clientes.

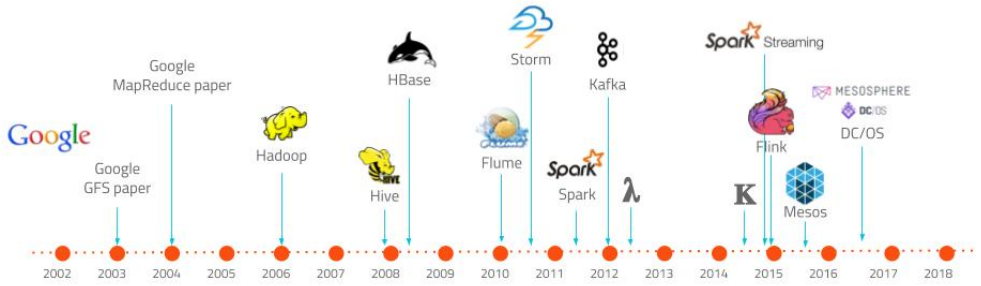

Figura 2.12: Evolución del *Big Data*. Fuente [10]

Durante estos 10 años la popularidad de *Hadoop* ha crecido exponencialmente y junto con las BBDD NoSQL, nacidas también a partir de Google con su BigTable, forman lo que hoy conocemos como Big Data. En la Figura 2.12 observamos esta evolución en el mundo del *Big Data* con los hitos de aparición de algunas tecnologías representativas.

El auge del **Big Data** ha llevado a algunas empresas a tener verdadera obsesión por el almacenamiento de todos los datos de sus clientes y las operaciones realizadas, creando inmensos *datalakes* donde tener enormes históricos de todos sus datos. Este "síndrome de Diógenes digital" creado por falsas expectativas, por la imposibilidad de extraer valor de los datos o por la dimensión cambiante de las empresas actuales (en la que los datos de años atrás pueden no ser relevantes en el presente), es uno de los posibles motivos por lo que el tratamiento de los datos esta cambiando. Otro de los motivos para el cambio de rumbo del *Big Data* está relacionado con la *V* de Velocidad, hoy en día no solo es importante la capacidad de ingestar rápidamente los datos, sino la capacidad de poder procesar y obtener decisiones o actuar en tiempo

real a partir de los datos, aportando valor al negocio. En este escenario se vuelve más importante la velocidad que el volumen de datos, esto es lo que se denomina *Fast Data*.

Dentro del *Fast Data* es habitual el uso de BBDD *in-memory*, de buses de eventos y de tecnologías de procesamiento capaces de procesar los eventos en tiempo real. Como veremos posteriormente al desarrollar nuestra arquitectura, el *Fast Data* será una parte fundamental en nuestro proyecto en el que tendremos que clasificar las llamadas en tiempo real y tomar decisiones (o alarmar) en función de las mismas.

Observando de nuevo la Figura 2.12 podemos ver este cambio en la tendencia hacia el *Fast Data* a partir del 2012 cuando aparece la tecnología Kafka y en los años posteriores con la incorporación de diferentes herramientas para el procesamiento de eventos como son *Spark Streaming* o *Flink*.

2.3.2. Arquitecturas *RealTime*

La evolución que hemos visto en el apartado anterior, con la explosión del *Big Data* y la irrupción del *Fast Data*, hace necesaria la incorporación en las empresas de arquitecturas de procesamiento de datos en tiempo real que, como cualquier otra arquitectura de datos, sean capaces de ingestar, procesar y permitir la explotación y análisis de los datos. La diferencia fundamental en las arquitecturas *RealTime* y las arquitecturas de datos tradicionales son el **volumen de los datos a tratar** y la **capacidad para hacerlo en tiempo real**.

Como veremos, no existe una arquitectura que se adapte a todos los casos de uso (*one-size-fits-all*) y, según la necesidad, será necesario aplicar una u otra.

Arquitectura Lambda λ

La arquitectura lambda, representada por la letra griega λ, fue presentada en 2011 por Nathan Marz en un artículo publicado en su blog titulado "*How to beat the CAP theorem*" [21].

El propósito de Nathan Marz cuando crea su arquitectura, como indica el

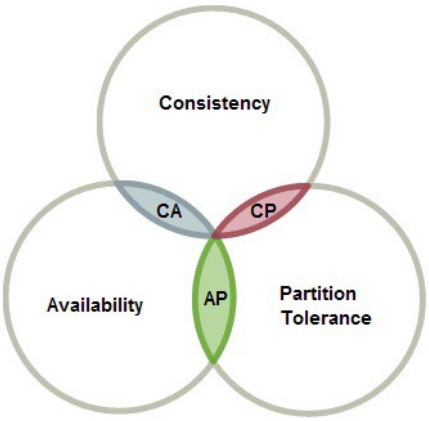

Figura 2.13: Teorema CAP. Fuente [25]

título del artículo, es batir el teorema CAP popularizado por la irrupción de las bases de datos NoSQL. Este teorema, ilustrado en la Figura 2.13, viene a decir que si queremos tener tolerancia a particiones (imprescindible para bases de datos distribuidas necesarias para el *Big Data*), tenemos que optar entre consistencia, asegurar que el dato que leemos es el último que hemos escrito, o disponibilidad, que la base de datos se encuentra siempre lista.

El método propuesto para conseguir este objetivo con grandes cantidades de datos se basa en los siguientes principios:

- Una capa *batch* eventualmente consistente de una manera extrema, en la que las escrituras tardan siempre unas pocas horas en estar disponibles. Eliminando algunos problemas complejos con los que tratar como la concurrencia o las reparaciones de lectura.

- Reducir las operaciones CRUD (***C****reated,* ***R****ead,* ***U****pdate,* ***D****elete*) en la capa *batch* por únicamente CR, tratando los datos como objetos inmutables. Esto nos soluciona el problema de la consistencia, ya que de este modo un dato existe o no existe, pero no puede tener varias versiones.

- Una capa *realtime* que se encarga de los datos de las últimas horas (los que no están disponibles en la capa *batch*)

- Las querys atacan a ambas capas de forma simultanea realizando un *merge* de los datos.

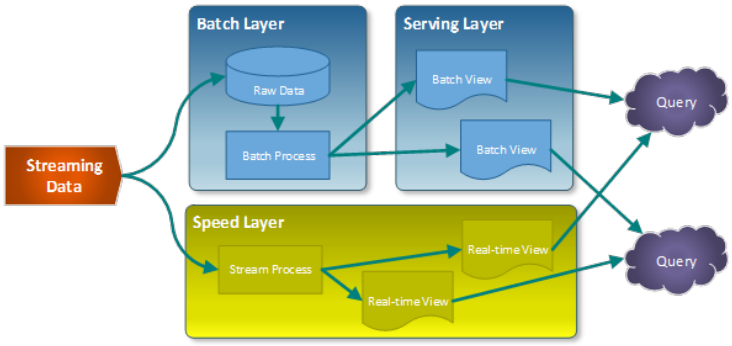

Figura 2.14: Arquitectura Lambda definida por Nathan Marz. Fuente [10]

Podemos ver un esquema de la arquitectura en la Figura 2.14. Esta arquitectura, aparte de resolver los problemas de consistencia y disponibilidad, tiene algunas ventajas:

- Disponer de todos los datos en un único punto (capa *batch*) pudiendo realizar cualquier tipo de consulta sobre los mismos.
- Al utilizar los datos como un ente inmutable facilita las auditorias.
- Según Marz, evita el error humano en la capa *batch* (en parte también por usar datos inmutables), y cualquier error en la capa *realtime* sería subsanado en pocas horas en la capa *batch*.

Arquitectura Kappa κ

En su artículo "Questioning the Lambda Architecture"[16], Jay Kreps cuestiona la arquitectura Lambda propuesta por Nathan Marz y propone una simplificación de la misma, basada en su experiencia en LinkedIn trabajando con Kafka y Samza. Esta simplificación se denomina arquitectura Kappa y viene representada por la letra griega κ.

Kreps describe la complejidad que supone en una arquitectura Lambda mantener idénticos procesos en *realtime* y *batch*. También expone que se

menosprecia la capacidad de la capa *realtime* (probablemente por la madurez del procesamiento *realtime* con respecto al *bacth*) y opina que es posible realizar el mismo procesamiento, incluso reprocesar el histórico de datos, en la capa *realtime*.

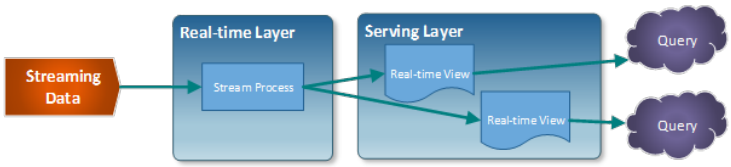

Figura 2.15: Arquitectura Kappa definida por Jay Kreps. Fuente [10]

Con esta premisa, en el artículo se presenta la arquitectura que observamos en la Figura 2.15, en la que existe un único flujo de procesamiento *realtime* para todo el modelo. La simplicidad de Kappa con respecto a Lambda es tal que el propio Kreps afirma que puede ser una idea demasiado simple para merecer una letra griega.

2.4. Trabajos anteriores

Una vez abordado el estado del arte desde diferentes puntos de vista, es importante tener una visión de los trabajos anteriores que se han realizado con objetivos similares y sus resultados. De este modo entenderemos si existe una justificación para nuestro trabajo, los problemas a los que podemos enfrentarnos y las expectativas que podemos gestionar.

En este apartado nos vamos a centrar en la aplicación de técnicas de Procesamiento del Lenguaje Natural a un *Call Center*. Encontramos varios artículos interesantes que hacen hincapié en el valor de la información y el conocimiento que puede extraerse de un *Call Center*, ya que se trata de un intermediario importante entre el cliente y la empresa. Entre estos trabajos podemos destacar:

- "Metodología para estimar el impacto que generan las llamadas realizadas en un call center en la fuga de los clientes utilizando técnicas de text mining" [30]: Que como su nombre indica, investiga si existe

relación entre las llamadas realizadas al *Call Center* y la pérdida de clientes. El trabajo, al igual que el proyecto que intentamos abordar, parte de las llamadas transcritas a texto y, aunque se utiliza como base para la modelización de temas LDA, se apoya en etiquetas existentes en las llamadas para validar los resultados.

- "Customer voice sensor: A comprehensive opinion mining system for call center conversation" [18]: Se trata de un trabajo más basado en el análisis de sentimientos, pero se encuentra realizado con llamadas de los clientes a una operadora de telecomunicaciones (en este caso China Telecom) al igual que nuestra fuente de información.

- "Topic mining for call centers based on A-LDA and distributed computing" [14]: En este caso, se realiza una modelización de temas sobre los datos del *Call Center* de *China Central Television*. En este proceso de modelización se utiliza una mejora del modelo LDA llamada A-LDA que utiliza no solo el corpus de la llamada, si no también algunas propiedades externas como el tiempo de llamada o el número de origen.

- "Author-topic based representation of call-center conversations" [24]: Este último artículo que comentamos, parte también de los datos generados a través de un *Performance of Automatic Speech Recognition* (ASR), el trabajo pone de manifiesto la pobre calidad de estas transcripciones automáticas. Por este motivo, propone una modificación de LDA basada en el modelo *Author Topic*, utilizando además del corpus del texto información del tema de la conversación.

Probablemente, dentro de las empresas, existen muchos más trabajos destinados a la explotación de esta información y que no se encuentren publicados, por lo que podemos concluir que es un campo que despierta interés y que se trata de una información con potencial que nos permita, entre otros objetivos, comprender las necesidades de los clientes de una compañía.

Por otro lado, observamos que en la mayoría de los casos, aunque la base de la modelización sea el uso de LDA, debido al ruido o a otros factores como

la ausencia de información semántica, se utilizan modificaciones del modelo que incorporan etiquetas o propiedades externas al corpus para mejorar la clasificación.

Por último, pensamos que, aún con estos antecedentes, existe la necesidad de abordar este proyecto, debido a la diferencia entre unos datos y otros, por factores como el idioma o las distintas necesidades de cada empresa. Además nuestro proyecto tiene como objetivo final la integración de este modelo en un proceso de la compañía que sea capaz de aplicarlo en tiempo real y alertar en caso de anomalías para poder tomar decisiones.

Capítulo 3

Arquitectura y tecnologías

El objetivo de este capítulo es tener una visión global de lo que será nuestro proyecto, antes de entrar en más profundidad en cada una de las partes. El proyecto que presentamos en este documento consta de dos partes bien diferenciadas. Por un lado, una parte que podemos considerar "de laboratorio", en la que realizaremos las labores más analíticas sobre los datos disponibles: extracción, procesamiento, estudio y creación de modelos; la cuál presentamos en la sección 3.1. Y por otro lado, una parte "de explotación", en la que trataremos de poner en valor los resultados de la primera parte en el mundo real y que será tratada en la sección 3.2.

El capítulo se completará con la sección 3.3, dedicada al mantenimiento del proyecto una vez llevado a producción y con la sección 3.4, en la que se describirán a grandes rasgos todas la tecnologías usadas en el proyecto.

3.1. Modelado

La primera parte de nuestro proyecto pretende conseguir el objetivo principal propuesto en la sección 1.3, **clasificar las llamadas**. Para ello tendremos que analizar y entender los datos que poseemos de las transcripciones y posteriormente construir los modelos necesarios (supervisados o no supervisados) que nos permitan clasificar las llamadas.

Para lograr este objetivo necesitaremos disponer de un *datalake* en el que

se almacene todo el histórico de transcripciones de las llamadas, ya que para el entrenamiento de muchos modelos será necesario utilizar un amplio histórico para su entrenamiento. Este *datalake* será un sistema de archivos HDFS perteneciente a una plataforma Hadoop Hortonworks. El procedimiento de carga de las transcripciones en HDFS se encuentra ya realizado y queda fuera del alcance del proyecto. Debido a que no tenemos ningún requisito temporal para la ingesta de las transcripciones, este no será un elemento crítico de nuestro proyecto.

En cuanto a la plataforma que utilizaremos para realizar la analítica, puede ser independiente del entorno de almacenamiento siempre que podamos transferir los datos a la misma. A lo largo del proyecto hemos trabajado con dos entornos diferentes:

- Entorno **Spark**: Ubicado en un clúster Hadoop Hortonworks, nos ha permitido procesar grandes cantidades de datos en un espacio de tiempo muy reducido, gracias a los beneficios de la programación distribuida. Sin embargo, no se trata del entorno ideal para el entrenamiento de modelos, principalmente para modelos *deep learning*.

- Entorno **GPU**s: Una única máquina con diferentes *GPU*s NVIDIA. Este entorno nos ha permitido entrenar los modelos de *deep learning* con una rapidez pasmosa (comparado con los entornos con CPUs). Ha sido la plataforma principal de analítica durante este proyecto, aunque el preprocesado se viera penalizado con respecto al entorno Spark.

El modelo presentado en nuestra arquitectura es un elemento sometido al cambio y, además de por posibles mejoras en los hiperparámetros o por la tecnología, debe re-entrenarse conforme se vayan recibiendo datos nuevos en el histórico, ya que es lógico pensar que la temática de las consultas variarán a lo largo del tiempo debido por ejemplo al lanzamiento de nuevos productos.

En los capítulos **??**, **??** y **??** trataremos en más detalle toda la parte de modelado, desde el análisis de los datos hasta la creación de modelos supervisados y no supervisados.

3.2. Explotación

La segunda parte del proyecto tendrá como meta llevar a producción el resultado de la etapa de modelado. Teniendo en cuenta que uno de nuestros objetivos consiste en clasificar las llamadas en tiempo real, el primer paso que debemos abordar es seleccionar un modelo de arquitectura de procesamiento en tiempo real de entre las opciones vistas en la sección 2.3.2: Lambda y Kappa. Lo que intentamos en este paso, ya que no existe una solución ideal que se adapte a todos los casos de uso, es encontrar la solución que mejor se adapte a nuestro proyecto.

En nuestro caso, disponemos de todas las llamadas en tiempo real, que nos llegarán en forma de eventos. Además cada llamada podrá ser tratada de forma individual y no debemos preocuparnos por eventos que lleguen con retraso. Estos son los motivos principales que nos han llevado a decidirnos por la arquitectura Kappa propuesta por Jey Kreps, debido a la mayor simplicidad tanto a la hora de elaborar la capa de servicio, como a la hora de mantener un único desarrollo *real-time*.

En la figura 3.1 podemos ver la arquitectura global de la solución propuesta. Esta arquitectura consta de dos capas principales. Por un lado, una capa de *streaming* o capa rápida donde se realizará todo el procesamiento en *streaming* de las transcripciones de las llamadas conforme vayan llegando. Por otro lado, una capa de servicio que permite a los usuarios explotar la información procesada en la primera capa.

El *core* de la primera capa será Apache Kafka, que además nos dará la posibilidad de retener los eventos y poder reprocesarlos en caso de error, siendo una solución ideal para arquitecturas Kappa. De hecho Jey Kreps, quien propuso la arquitectura Kappa es el co-fundador de Confluent, la compañía que se encuentra detrás de Apache Kafka. Los detalles de la capa de *streaming* podemos verlos con más detalle en el capítulo **??**.

El núcleo de la capa de servicio será el *stack* de Elastic. Esta solución nos permitirá ingestar los eventos desde la capa rápida en tiempo real y exponerlos a los usuarios. Además usaremos este mismo flujo para todos los eventos de monitorización. El hecho de utilizar el *stack* de Elastic nos dará

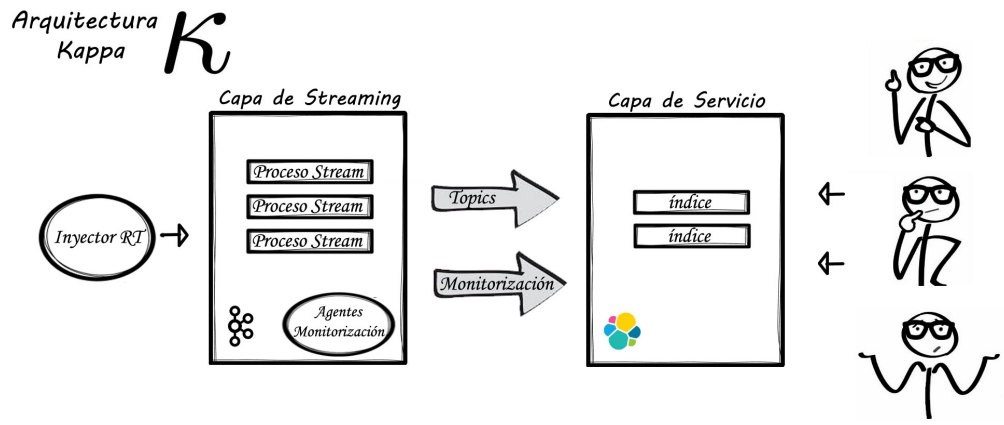

Figura 3.1: Arquitectura Kappa

posibilidad de realizar el alarmado estático y dinámico en esta misma capa. Los detalles de esta capa se describen en detalle en el capítulo **??**.

Debido a la situación actual, las llamadas no se ingestan en *real-time* si no que se reciben mediante procesos *batch* cada cierto tiempo. Este escenario cambiará en el futuro por lo que se construirá un elemento de entrada a la capa rápida que actúe como inyector, generando eventos en tiempo real a partir de los datos en *batch*. Esta pieza será suprimida una vez las llamadas sean recibidas en tiempo real.

3.3. Mantenimiento e Integración Continua

Como ya hemos visto al hablar del entrenamiento del modelo, el desarrollo de este tipo de proyectos no tiene un principio y un final, si no que se trata de un proceso cíclico en el que por necesidades del negocio, por cambio en los datos o por cambio en las tecnologías, será necesario añadir mejoras o modificaciones en nuestro desarrollo. También el despliegue del *software* en producción es un proceso susceptible de futuros cambios.

Por estos motivos será necesario disponer de algún método para seguir evaluando la calidad del modelo una vez sea llevado a producción, tener la posibilidad de realizar test A/B en el futuro, y definir mecanismos que nos

permitan, tras cada cambio efectuado, poder realizar las pruebas necesarias y desplegar estos cambios de una manera totalmente automatizada y sin intervención humana.

Para conseguir estos objetivos, con la mayor agilidad posible, será necesario trabajar en un marco de trabajo *DevOps* contando con mecanismos que nos permitan realizar tanto integración, como despliegue continuos. En el capítulo **??** veremos con más detalle como aplicamos esta metodología.

3.4. Tecnologías

Al igual que la arquitectura descrita anteriormente era la encargada de responder a las necesidades de negocio, las tecnologías descritas en este apartado nos darán las piezas necesarias para poder construir esa arquitectura y dar respuesta a nuestro caso de uso.

En el proceso de selección de las tecnologías, no solo se ha tenido en cuenta la idoneidad de las mismas para el caso de uso, si no que se ha valorado también la experiencia en la misma y la disponibilidad dentro del entorno de trabajo. Esto puede provocar que en algunos casos aunque la tecnología se adapte al caso de uso, puedan existir otras soluciones más óptimas cuyo uso era menos viable, dados los plazos de ejecución del proyecto.

A continuación enumeraremos las tecnologías agrupadas en las diferentes capas que hemos comentado en el apartado de arquitectura, además añadiremos las tecnologías que se usarán para la integración y despliegue continuo.

3.4.1. Modelado

La parte de modelado la realizaremos trabajando siempre con Python 3.6 y Jupyter, trabajar en un entorno dinámico e interactivo mediante *notebooks* nos dará mucha flexibilidad para poder realizar nuestro análisis.

A continuación vemos las bibliotecas más relevantes que hemos usado, tanto en entorno Spark como en el entorno de GPUs.

- **Spark SQL y *dataframes***: Dentro del entorno de Spark, para el procesamiento de los datos trabajaremos con las bibliotecas de Spark

SQL y *dataframes*, que nos permitirán de una manera sencilla realizar transformaciones en nuestros datos de forma distribuida.

- **MLlib**: También en el entorno de Spark usaremos la biblioteca nativa MLlib, que nos permite crear algoritmos y modelos de *machine learning* sobre un cluster Spark.
- **Pandas**: Al igual que en *Spark* utilizamos las bibliotecas de SQL y *dataframes*, Pandas nos permitirá hacer el análisis y manipulación de los datos en cualquier entorno Python.
- **Keras**: Será la biblioteca que usaremos, con **Tensorflow** como backend, para crear modelos de *deep learning*.
- **Gensim**: Se trata de una biblioteca para el modelado de temas no supervisados y el procesamiento del lenguaje natural.
- **NLTK**: Son un conjunto de bibliotecas que nos ayudarán a realizar tareas de procesamiento de lenguaje natural.
- **Sklearn**: Se trata de una biblioteca para tareas de *machine learning*, que contiene multitud de funciones que nos ayudarán en nuestro desarrollo.
- **Optuna**: Por último, la biblioteca Optuna nos ayudará a optimizar los modelos supervisados que creemos. Hablaremos de ella con más detalle en la sección **??**.

3.4.2. Explotación

A continuación veremos las tecnologías y plataformas que harán posible el despliegue de nuestra capa de *streaming*. Estas tecnologías serán vistas con más detalle a lo largo del capítulo **??**.

- **Apache Kafka**: Es el *core* de la capa rápida, se trata de un bus de eventos distribuidos, a través del cual se realizará la ingesta o publicación de los eventos (llamadas). Las diferentes capas de procesamiento que requieran estos eventos se suscribirán a este Bus.

- **Tensorflow Serving**: Servicio para servir modelos de *machine learning* creado dentro del ecosistema Tensorflow.
- **Kafka Stream**: A la hora de procesar la información en eventos ingestada en nuestro Bus Kafka disponemos de Kafka Stream. Una serie de bibliotecas que nos permiten construir aplicaciones y microservicios cuyo origen y destino sean un Bus Kafka.
- **Openshift**: Se trata de una plataforma de contenedores de Kubernetes creada por *Red Hat*. Todos los desarrollos de la capa de explotación serán implementados sobre contenedores.

3.4.3. Capa Servicio

La capa de servicio estará compuesta, como hemos indicado, por el *stack* de Elastic. Este *stack* posee diferentes piezas, cada una de las cuales realiza una función determinada.

- **Elasticsearch**: Se trata del núcleo del *stack*. Aunque no se trata en el sentido más estricto de una BBDD No-SQL, si no de un motor de búsqueda, Elasticsearch nos permite almacenar la información en forma de documentos json en tiempo real y realizar consultas y agregaciones sobre cualquier campo. Entre las características que podemos aprovechar de Elasticsearch para nuestro objetivo están:
  - Ingesta en tiempo real.
  - Consulta en tiempo real.
  - Disponibilidad de mecanismos de ingesta (Logstash) y consulta (Kibana).
  - Posibilidad de crear alarmas en base a consultas.
  - Posibilidad de crear *jobs* de *machine learning* que detecten anomalías en series temporales.
  - API REST: Posibilidad de realizar cualquier operación o consulta mediante API REST.

- **Logstash**: Será la pieza que nos permitirá transportar la información desde la capa de *streaming* a la capa de servicio. Logstash nos permitirá leer de Apache Kafka, realizar las transformaciones necesarias y volcar la información resultante en Elasticsearch. Además de los datos propios de la clasificación, Logtsash nos ayudará a extraer también los datos de monitorización de la capa de Streaming.

- **Kibana**: Será el frontal donde los usuarios podrán consultar sus diferentes cuadros de mando y construir nuevos de acuerdo a sus necesidades. También, gracias al módulo de *machine learning* de Elasticsearch, los usuarios podrán crear *jobs* de *machine learning* para detectar anomalías en los temas tratados y generar las alertas necesarias.

- **Beats**: Se trata de agentes ligeros que nos permitirán extraer información variada de distintas fuentes. En nuestro caso usaremos *Metricbeat* para extraer datos de monitorización de la capa de *streaming*.

3.4.4. Integración y Despliegue Continuo

Por último, para conseguir los objetivos de integración y despliegue continuos utilizaremos diversas tecnologías. Estas tecnologías, y la interacción entre ellas se verán con más detalle en el capítulo **??**.

- **BitBucket**: Será el repositorio usado para almacenar las nuevas versiones de nuestro *software* de manera que podamos tener un control de versiones. Almacenará tanto el código de nuestra capa de *streaming*, como de los modelos que necesitemos poner en producción.

- **Jenkins**: Es un servidor de integración continua *open source* que, mediante la creación de tareas, nos ayudará a realizar el *build* de nuestro software realizando de manera automática las pruebas necesarias.

- **Nexus**: Se trata de un gestor de repositorios, en nuestro caso utilizaremos repositorios de Maven. Es usual usar este tipo de gestores de repositorios en las empresas para disponer de bibliotecas propias (y no públicas) y para no tener una dependencia con ningún agente externo.

- **S2I**: Se trata de una funcionalidad de Openshift que será vital para el desarrollo del despliegue continuo. Esta funcionalidad nos permitirá crear nuevas imágenes de contenedores a partir de código o binarios propios.

Parte II

Modelado: datos, modelos y optimizaciones

Parte III

Explotación: procesamiento, visualización y alarmados

Parte IV

Conclusiones: mantenimiento y futuros trabajos

Capítulo 4

Conclusiones

4.1. Aplicación aprendido

Aunque se ha profundizado en algunos aspectos, más allá de lo aprendido a lo largo del máster, las asignaturas cursadas a lo largo de tres semestres han creado una base solida que han permitido el desarrollo de este proyecto.

En esta sección queremos hacer un repaso sobre la aplicación que todas las asignaturas cursadas a lo largo del máster, en mayor o menor medida, han tenido en este proyecto.

- **Fundamentos de la ciencia de datos**: Esta asignatura nos ha permitido entender los principios básicos del mundo de la ciencia de datos. Entre otros aspectos la hemos utilizado para entender el ciclo de vida de un proyecto de ciencia de datos. También se han tratado otros aspectos que nos han sido útiles en el desarrollo del proyecto, como son la metodología *Agile*, la calidad de los datos, etc.

- **Tipología y ciclo de vida de los datos**: Esta asignatura es fundamental para entender el ciclo de vida de los datos. En ella hemos tratado los diferentes tipos de datos que podemos encontrar y las formas que tenemos de obtenerlos.

- **Arquitecturas de bases de datos no tradicionales**: En esta asignatura se trabajó con diferentes modelos de bases de datos NoSQL que

han sido fundamentales para el desarrollo del proyecto. En nuestro proyecto hemos trabajados con *Elasticsearch* en la capa de servicio, que aunque se trate de un motor de búsqueda presenta muchas similitudes con una BBDD NoSQL documental. Además se ha trabajado en la capa de *streaming* con Apache Kafka que interiormente posee una BBDD RocksDB, que se trata de una NoSQL clave-valor.

- **Estadística avanzada**: Aunque no se han aplicado los modelos estadísticos vistos en esta asignatura, han sido útiles los conceptos vistos para las fases de preprocesado y análisis de los datos.

- **Minería de datos**: En esta asignatura se trabajaron a nivel práctico los métodos *core* de la minería de datos. Para nuestro proyecto han sido de utilidad conceptos de preparación de datos, de *clustering* y de evaluación de modelos.

- **Modelos avanzados de minería de datos**: En esta asignatura se profundiza más en los modelos de minería de datos y empezamos a aplicar las redes neuronales, que son la base de nuestros modelos supervisados, además tratamos temas que nos han sido útiles como la combinación de clasificadores.

- **Deep learning**: El contenido de esta asignatura ha sido vital para los modelos de aprendizaje supervisado. En ella hemos profundizado en el mundo de las redes neuronales y los diferentes tipos: convolucionales, recurrentes, etc.

- **Análisis de sentimientos y redes sociales**: En esta asignatura se trabajaron las bases del procesamiento del lenguaje natural que ha sido el *core* de nuestro proyecto.

- **Visualización de datos**: en esta asignatura trabajamos los conceptos que existen detrás de una buena visualización. Estos conceptos se han intentado aplicar en nuestro proyecto, tanto en la fase de análisis como en la capa de servicio y visualización. Las visualizaciones interactivas se han creado en nuestro caso a través de Kibana.

- **Diseño y construcción del data warehouse**: Aunque no se han aplicado directamente los conocimientos de esta asignatura en el desarrollo del proyecto, sí ha sido necesario acceder al *datawarehouse* de la empresa para la obtención de datos.
- **Análisis de datos en entornos big data**: Aunque no hemos cursado esta asignatura, sí hemos aplicado algunos conceptos tratados en ella y relacionados con el uso de un ecosistema Hadoop. Principalmente HDFS, Spark y Hive.

4.2. Líneas de trabajo futuras

- Métricas Optimización:
- Etiquetas monitorización:
- Profundizar estudio no supervisado:
- Llevar a producción modelo no supervisado.
- Adaptar llamadas tiempo real, eliminar inyector.
- Disponer llamadas

4.3. Caso de negocio

Lo que me contó Nacho ver si encajarlo aquí o en la introducción.

4.4. Agradecimientos

Un *handicap* a la hora de realizar el proyecto dentro de una gran empresa ha sido el hecho de trabajar con unos plazos tan ajustados. Aspectos como la autorización en el acceso a la información, el acceso a diferentes entornos, la intercomunicación entre áreas, etc. requieren unos tiempos que pueden retrasar la ejecución de un trabajo de este tipo.

Esta última sección tiene como objetivo dar las gracias a todas las personas que han hecho posible la realización de este proyecto en tiempo y forma.

A Antonio Fernández Gallardo

A Willy Gavilán Montegro y Carolina Bouvard Nuño por permitirme realizar el proyecto dentro de Telefónica.

A Jorge Ayuso Rejas y Rus Mesas Javega por su ayuda a la hora de obtener los datos y por permitirme usar con ellos el equipo con GPUs usados por el entrenamiento.

A Laura Canga García por dejarme usar el Bus Kafka y a José Ramón Fernández Acosta por ayudarme, incluso a deshoras, con la creación y usuarios del bus.

A Pablo Palomares Darocas por su ayuda a la hora de acceder y usar las plataformas *DevOps*: Bitbucket, Nexus y Jenkins.

A Nacho Charfolé Sancho por darme su visión más allá de la parte técnica.

Bibliografía

[1] O. Abdelwahab and A. Elmaghraby. Deep learning based vs. markov chain based text generation for cross domain adaptation for sentiment classification. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 252–255, Julio 2018.

[2] Toni Lozano Bagén Anna Bosch Rué, Jordi Casas Roma. *Deep Learning Principios y Fundamentos*. 2018.

[3] David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 17–24, Cambridge, MA, USA, 2003. MIT Press.

[4] David M. Blei and John D. Lafferty. Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pages 147–154, Cambridge, MA, USA, 2005. MIT Press.

[5] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[7] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-

step data mining guide. Technical report, The CRISP-DM consortium, Agosto 2000.

[8] Michael Cox and David Ellsworth. Application-controlled demand paging for out-of-core visualization. In *Proceedings of the 8th Conference on Visualization '97*, VIS '97, pages 235–ff., Los Alamitos, CA, USA, 1997. IEEE Computer Society Press.

[9] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation - Volume 6*, OSDI'04, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.

[10] Jesús Domínguez. De lambda a kappa: evolución de las arquitecturas big data. `https://www.paradigmadigital.com/techbiz/de-lambda-a-kappa-evolucion-de-las-arquitecturas-big-data/`, Abril 2018. Último acceso 2019-10-13.

[11] Christine Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.

[12] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. *SIGOPS Oper. Syst. Rev.*, 37(5):29–43, October 2003.

[13] Yoav Goldberg. *Neural network methods in natural language processing*. Morgan & Claypool publishers, 2017.

[14] Wenming Guo, Lihong Liang, and Tianlang Deng. Topic mining for call centers based on a-lda and distributed computing. *Concurrency and Computation: Practice and Experience*, 29(3):e3776, 2017. e3776 CPE-15-0479.R1.

[15] Tom Hope, Itay Lieder, and Yehezkel S. Resheff. *Learning TensorFlow: a guide to building deep learning systems*. OReilly Media, 2017.

[16] Jay Kreps. Questioning the lambda architecture. `https://towardsdatascience.com/`

cap-theorem-and-distributed-database-management-systems-5c2be977950e, Julio 2014. Último acceso 2019-10-14.

[17] Douglas Laney. 3D data management: Controlling data volume, velocity, and variety. Technical report, META Group, February 2001.

[18] P. Li, Y. Yan, Chaomin Wang, Zhijie Ren, Pengyu Cong, Huixin Wang, and Junlan Feng. Customer voice sensor: A comprehensive opinion mining system for call center conversation. In *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 324–329, Julio 2016.

[19] F. Long, K. Zhou, and W. Ou. Sentiment analysis of text based on bidirectional lstm with multi-head attention. *IEEE Access*, 7:141960–141969, 2019.

[20] A. S. M. Ashique Mahmood. *Literature Survey on Topic Modeling.* 2013.

[21] Nathan Marz. How to beat the cap theorem. http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html, Octubre 2011. Último acceso 2019-10-14.

[22] Yingjie Miao. Reactive ldap library. https://medium.com/kifi-engineering/reactive-lda-library-d495ed2a6342. Último acceso 2019-12-18.

[23] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[24] M. Morchid, R. Dufour, M. Bouallegue, and G. Linarès. Author-topic based representation of call-center conversations. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 218–223, Diciembre 2014.

[25] Syed Sadat Nazrul. Cap theorem and distributed database management systems. https://towardsdatascience.com/cap-theorem-and-distributed-database-management-systems-5c2be977950e, Abril 2018. Último acceso 2019-10-14.

[26] Michael Nguyen. Illustrated guide to lstm's and gru's: A step by step explanation. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85 Septiembre 2018. Último acceso 2019-10-13.

[27] Christopher Olah. Understanding lstm networks – colah's blog. https://colah.github.io/posts/2015-08-Understanding-LSTMs/, Agosto 2015. Último acceso 2019-10-13.

[28] Stacey Ronaghan. Deep learning: Common architectures. https://medium.com/@srnghn/deep-learning-common-architectures-6071d47cb383, Agosto 2018. Último acceso 2019-10-13.

[29] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[30] Jullian Sepúlveda. *Metodolgía para estimar el impacto que generan las llamadas realizadas en un call center en la fuga de los clientes utilizando técnicas de text mining.* PhD thesis, Universidad de Chile, 2015.

[31] A. M. Turing. *Computing machinery and intelligence.* Blackwell for the Mind Association, 1950.

[32] Kevin Warwick and Huma Shah. *Turings imitation game: conversations with the unknown.* Cambridge Univeristy Press, 2016.

[33] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg

Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.