UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

CARRERA DE CIENCIA DE LA COMPUTACIÓN

ANÁLISIS DE LOS MÉTODOS DE RECOLECCIÓN DE TEXTOS SARCÁSTICOS

TESIS

Para optar el título profesional de Licenciada en Ciencia de la Computación

AUTORA:

Andrea Velásquez Gushiken (ORCID: 0000-0002-0039-3663)

ASESOR

Cristian José López del Álamo (ORCID: 0000-0002-2568-650X)

Lima - Perú

14 de abril de 2023

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

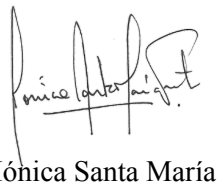
DIRECCIÓN DE INVESTIGACIÓN

DECLARACIÓN JURADA

Yo, Mónica Cecilia Santa María Fuster identificada con DNI No 18226712 en mi condición de autoridad responsable de validar la autenticidad de los trabajos de investigación y tesis de la UNIVERSIDAD DE INGENIERIA Y TECNOLOGIA, DECLARO BAJO JURAMENTO:

Que la tesis denominada "Análisis de los métodos de recolección de textos sarcásticos" ha sido elaborada por la señorita Andrea Velásquez Gushiken, con la asesoría de Cristian José López del Álamo, identificado con DNI N°29708892 , y que se presenta para obtener el título profesional de licenciado en Ciencia de la computación, ha sido sometida a los mecanismos de control y sanciones anti plagio previstos en la normativa interna de la universidad, encontrándose un porcentaje de similitud de 0%.

En fe de lo cual firmo la presente.



Dra. Mónica Santa María Fuster
Directora de Investigación

En Barranco, el 19 de diciembre de 2022

Agradecimientos:

A mi asesor, el Dr. Cristian López, por su conocimiento, su guía en temas que no conocía a profundidad y soportar mis horarios no convencionales e ideas de última hora.

A mi profesora en el curso de tesis, la Dra. Yamilet Serrano, por sus sugerencias en cómo estructurar la redacción y por ayudarme a desenredar las partes que parecían trabalenguas.

Por último, a mi familia, por su apoyo moral y por escucharme aunque no entendieran el tema, porque a veces los problemas se resuelven tratando de explicarlos.

Índice general

Pág.

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La detección de sarcasmo es un obstáculo particularmente complicado de resolver dentro del Procesamiento de Lenguaje Natural. En los últimos años se han propuesto mejoras en la arquitectura y funcionamiento de los modelos que buscan resolver el problema.

No obstante, se ha dejado del lado la importancia de los textos sarcásticos que se utilizan para entrenarlos y, con ella, los métodos de recolección de estos textos. Los métodos tradicionales producen *datasets* sesgados, con errores y ruidosos, y no distinguen entre los dos tipos de sarcasmo: intencional y percibido. Por ello, en la presente investigación, se analiza cuantitativamente el impacto que tienen los métodos de recolección de *datasets* sarcásticos en inglés en los modelos de detección de sarcasmo.

Con este fin, se hace uso de *datasets* públicos y se generan dos nuevos *datasets* con el método de Supervisión Reactiva (Shmueli *et al.*, 2020) para analizar el impacto de los distintos métodos de recolección en el desempeño de modelos de detección de sarcasmo. Se realiza una comparación detallada de los métodos, entrenando modelos en el estado del arte con un *dataset* representativo de cada uno de ellos.

Los resultados sugieren que es posible obtener mejores resultados en los modelos de detección de sarcasmo utilizando un método que provea un *dataset* limpio y el mismo tipo de sarcasmo que el que se quiere detectar. A su vez, confirman los descubrimientos realizados en investigaciones anteriores, y abren el camino a trabajos futuros.

Palabras clave:
Procesamiento de Lenguaje Natural, Análisis de Sentimientos, Detección de sarcasmo, Tipos de sarcasmo, Sarcasmo intencional, Sarcasmo percibido, Métodos de recolección de *datasets*, Supervisión Distante, Anotación Manual, Recolección Manual, Supervisión Reactiva.

ABSTRACT

Analysis of Sarcasm Collection Methods

Sarcasm Detection is a particularly complex setback in Natural Language Processing. In the last years, there have been improvements in the architecture and functionality of models that try to solve the problem.

However, the importance of the sarcastic texts used to train the models has been left aside, as well as their collection methods. The traditional methods generate biased and noisy datasets with errors, and do not differentiate the two types of sarcasm: intentional and perceived. In consequence, the current investigation does a quantitative analysis on the impact that collection methods of sarcastic datasets in English have on sarcasm detection models.

For that purpose, the investigation uses public datasets and generates two new datasets with the Reactive Supervision method (Shmueli *et al.*, 2020) to analyze the impact of the collection methods on the performance of sarcasm detection models. It makes a detailed comparison of the methods, training state-of-the-art models with a representative dataset of each one of them.

The results suggest that it is possible to obtain better models using a method that provides a clean dataset and the type of sarcasm to be detected. At the same time, they confirm the findings made by previous investigations and open a path to future works.

Keywords: Natural Language Processing, Sentiment Analysis, Sarcasm Detection, Types of Sarcasm, Perceived Sarcasm, Intended Sarcasm, Collection Methods, Distant Supervision, Manual Anotation, Manual Collection, Reactive Supervision.

Capítulo 1

Motivación y Contexto

1.1 Introducción

Durante la última década, el incremento exponencial de datos generados en el mundo ha impulsado el análisis automatizado de la información. Frente a esta situación, el Procesamiento del Lenguaje Natural (PLN) ha ido adquiriendo importancia, ya que permite extraer información subjetiva de textos de manera automática, labor que anteriormente no podía ser realizada por computadoras (Cambria y White, 2014).

El PLN es aplicado en distintos campos como el procesamiento de datos, inteligencia empresarial, robótica, medicina, lenguaje humano, entre otros, siendo la detección de sentimientos uno de sus objetivos principales. Dentro de los obstáculos en esta área se encuentra el sarcasmo, cuya detección es particularmente complicada por su carácter contradictorio y uso extendido (Poria *et al.*, 2020).

Investigaciones previas han propuesto modelos para la detección del sarcasmo como BERT (Devlin *et al.*, 2019) y RoBERTa (Liu *et al.*, 2019b). Además de los avances que se han desarrollado y aplicado en el diseño de los modelos, estos han sido entrenados y puestos a prueba con un *corpus* de comentarios sarcásticos en inglés disponibles en *datasets* ampliamente aceptados en el área (Khodak *et al.*, 2018, Oprea y Magdy, 2019b, Ptáček *et al.*, 2014, Riloff *et al.*, 2013).

Sin embargo, los métodos de recolección utilizados presentan limitaciones: datos sesgados, en ocasiones incorrectos y, ante todo, que no toman en cuenta la diferencia entre el *sarcasmo intencional* y el *sarcasmo percibido*. En el primero de estos dos tipos se garantiza que el emisor buscaba ser sarcástico, mientras que en el segundo el receptor percibió

sarcasmo, indistintamente de las intenciones del emisor. El estudio realizado por Oprea y Magdy (2019a) sugiere que el desempeño de los modelos varía entre ambos tipos, siendo menor al detectar sarcasmo percibido. Pese a ello, los modelos mencionados al principio, *BERT* y *RoBERTa*, no son los únicos que cometen este error, sino que, en su mayoría, los modelos actuales de detección de sarcasmo solo han sido evaluados con *datasets* generados con métodos tradicionales, que únicamente buscan un símbolo que indique sarcasmo como *"/s"* en *Reddit* (Khodak *et al.*, 2018) y *"#sarcasm"* en *Twitter* (Ptáček *et al.*, 2014), o etiquetan los textos manualmente (Riloff *et al.*, 2013) sin considerar tipos distintos de sarcasmo y generando datos sesgados.

En esta investigación, se analiza el impacto de los métodos de recolección de *datasets* sarcásticos en inglés en el desempeño de los modelos en el estado del arte. Para ello, se comparan los métodos tradicionales de Supervisión Distante, Anotación manual y Recolección Manual con el método de Supervisión Reactiva (Shmueli *et al.*, 2020), el más reciente y que toma en consideración los tipos de sarcasmo y el contexto en *tweets*. En los experimentos, se ha entrenado un modelo con *datasets* recolectados con métodos distintos y, posteriormente, se validó con *datasets* de sarcasmo intencional y percibido para comparar resultados.

El presente documento se encuentra dividido en capítulos y estos en secciones. En las siguientes secciones se detalla el problema a tratar en la investigación y la justificación de la misma. Posteriormente, el segundo capítulo explica conceptos clave, tales como los tipos de sarcasmo o los métodos de recolección y el tercer capítulo explora estudios relacionados. A continuación, el cuarto y el quinto capítulo exponen la metodología utilizada en los experimentos y los resultados obtenidos, junto con un análisis de ellos. Finalmente, el capítulo 6 muestra un balance de la investigación y proporciona puntos de mejora para trabajos futuros.

1.2 Descripción del Problema

Estudios en el área de detección de sarcasmo han propuesto modelos donde sobresale un modelo mejorado de BERT llamado RoBERTa.

Sin embargo, ambos modelos sufren una seria limitación debido a que han dejado del lado el problema de la calidad de los *datasets* utilizados para entrenar los modelos. Entendiéndose como calidad a datos no sesgados, correctos y que consideren los tipos de sarcasmo.

Mientras que en el diseño de los modelos se emplean técnicas novedosas, los *datasets* utilizados (Khodak *et al.*, 2018, Oprea y Magdy, 2019b, Ptáček *et al.*, 2014, Riloff *et al.*, 2013) han sido recolectados con métodos desactualizados que dejan del lado el sarcasmo intencional y sarcasmo percibido, y en ocasiones resultan en datos de baja calidad. Para una mejor explicación de la diferencia entre estos tipos, se utilizará un comentario hipotético de *Reddit*: *"Claramente un ejemplo de inteligencia humana."*. En el comentario, el sarcasmo es de tipo percibido porque es percibido por el receptor. Si el comentario incluyera el símbolo *"/s"*, usado para denotar sarcasmo en *Reddit*, el sarcasmo es intencional debido a que existe una confirmación que el emisor tiene la intención de ser sarcástico.

Los métodos catalogados en el presente estudio como desactualizados y sus respectivos problemas se listan a continuación.

- **Supervisión Distante**: Datos sesgados y a veces erróneos (Davidov *et al.*, 2010). Solo genera sarcasmo intencional.
- **Anotación Manual**: Datos sesgados y datos erróneos en menor medida que en la Supervisión Distante. Además, el trabajo manual conlleva un alto costo que origina *datasets* de tamaño reducido. Solo genera sarcasmo percibido (Joshi *et al.*, 2016a).

- **Recolección Manual**: Costo aún más alto que la Anotación Manual con una leve posibilidad de datos erróneos (Oprea y Magdy, 2019b). Solo genera sarcasmo intencional.

Las deficiencias mencionadas de estos métodos no han sido consideradas durante el desarrollo de modelos de detección de sarcasmo. Ello no puede ser ignorado puesto que cada método recolecta un tipo de sarcasmo distinto y, si el modelo no ha sido entrenado con el tipo requerido en la situación en la que será utilizado, su desempeño puede ser distinto al reportado en los estudios. Además, si el método de recolección utilizado es deficiente, los datos obtenidos no serán aptos para entrenar un modelo para aplicaciones reales.

1.3 Justificación

La mayoría de estudios en la detección de sarcasmo se han centrado en diseñar o mejorar modelos para la detección de sarcasmo experimentando con técnicas conocidas en el área de PLN, como la arquitectura de *transformers* y RCNNs (Devlin *et al.*, 2019, Liu *et al.*, 2019b, Potamias *et al.*, 2020). Mientras tanto, la calidad de los *datasets* con los que se entrenan a los modelos ha sido relegada.

Investigaciones previas sugieren que el sarcasmo intencional y percibido deben tratarse como fenómenos distintos (Joshi *et al.*, 2016a, Oprea y Magdy, 2019b). Un análisis de 1832 *tweets* del *dataset* de Riloff, recolectado con Supervisión Distante y verificado con Anotación Manual, muestra un 28 % de *tweets* percibidos incorrectamente. Más aún, 58 % de los *tweets* con *hashtag* fueron percibidos como no sarcásticos (Tabla 1.1), cuando la intención era sarcástica (Oprea y Magdy, 2019b).

Considerar los tipos de sarcasmo durante el entrenamiento del modelo es necesario para generar textos que sean percibidos como sarcásticos y reconocer textos con intención sarcástica.

La investigación detallada en el presente documento permitirá analizar el impacto que tienen los métodos de recolección en el desempeño de los modelos. Asimismo, se contrastará la efectividad de los métodos para entrenar modelos que detecten sarcasmo percibido con el intencional.

TABLA 1.1: Distribución 1832 *tweets* analizados del *dataset* de Riloff *et al.* (2013). Se muestra el número de *tweets* con y sin el *hashtag "#sarcasm"* y el número de *tweets* percibidos como sarcásticos y no sarcásticos.
Traducido del inglés de Oprea y Magdy (2019b)

	Con "#sarcasm"	**Sin "#sarcasm"**
Percib. sarcásticos	345	26
Percib. no sarcásticos	486	975

1.4 Objetivos

1.4.1 Objetivo general

El objetivo del presente estudio es analizar cuantitativamente el impacto que tienen los métodos de recolección de *datasets* sarcásticos en inglés en los modelos de detección de sarcasmo.

1.4.2 Objetivos específicos

- Contrastar las métricas de desempeño entre modelos entrenados con *datasets* de métodos diferentes (tradicionales y Supervisión Reactiva). Aplicado en el Experimento A.
- Analizar el impacto del método de recolección en la detección de textos sarcásticos por tipo: intencional y percibido. Aplicado en el Experimento B.
- Comprobar que el efecto del uso de los *datasets* de entrenamiento no se limite a un modelo en específico. Aplicado en Experimentos A y B.

Capítulo 2

Marco Teórico

En el capítulo anterior se realizó una introducción al tema de la investigación y la estructura del documento. A continuación, el capítulo 2 explica conceptos importantes sobre los métodos de recolección y la detección de sarcasmo necesarios para entender a profundidad los experimentos diseñados y sus resultados.

2.1 Sobre el sarcasmo y tipos

El sarcasmo es una forma de ironía con la finalidad de burla o crítica y carácter sutil. Al ser parte de la ironía, busca comunicar una idea contraria a su significado literal, dando indicios de ello mediante gestos, un tono de voz específico, un tono hostil o cuando el contexto es incompatible con la idea. (Bagate y Ramadass, 2020) Sin embargo, puede ser confundido fácilmente por personas con un trasfondo sociocultural diferente al emisor o con falta de conocimientos sobre el tema(Joshi *et al.*, 2016b).

2.1.1 Tipos de sarcasmo

Debido a la subjetividad en la interpretación del sarcasmo, surgen dos tipos de sarcasmo:

- **Intencional**: El fenómeno de sarcasmo intencional ocurre cuando el comentario es sarcástico para el emisor.
- **Percibido**: El sarcasmo percibido es sarcástico para el receptor, incluso si el emisor no buscaba serlo.

2.2 Métodos de recolección de *datasets* sarcásticos

Con el fin de recopilar textos sarcásticos para entrenar modelos de detección de sarcasmo, han surgido distintos métodos manuales y automáticos, los cuales se presentan a continuación en orden cronológico. Las fuentes más populares utilizadas son *Twitter* y *Reddit*. Entre ambos, *Twitter* es más utilizado debido a que los *tweets* no suelen necesitar un contexto detallado y tienen una longitud limitada (280 caracteres).

2.2.1 Métodos tradicionales

2.2.1.1. Supervisión Distante

Es el método más antiguo y simple. Etiqueta textos basándose en parámetros fijos, es decir, si sigue un patrón o contiene una palabra o frase específica se considera como sarcástico. Entre las palabras más utilizadas se encuentran los *hashtags "#sarcasm"*, *"#sarcastic"*, *"#not"* y, a veces, *"#irony"* en *Twitter* (Davidov *et al.*, 2010) y el símbolo "/s" en *Reddit* (Khodak *et al.*, 2018).

2.2.1.2. Anotación Manual

Consiste en clasificar un conjunto de textos sarcásticos manualmente. Los textos pueden ser recolectados mediante Supervisión Distante o aleatoriamente. En la mayoría de casos se prefiere la primera, ya que garantiza una mayor probabilidad de encontrar comentarios sarcásticos. De combinarse ambos métodos, el método se vuelve híbrido (Riloff *et al.*, 2013).

2.2.1.3. Recolección Manual

Al igual que el método anterior, requiere de trabajo manual. Se apoya en un conjunto de personas a los que se les solicita crear textos sarcásticos. (Oprea y Magdy, 2019b)

El procedimiento utilizado en los métodos tradicionales ocasiona que solo se obtengan textos con un solo tipo de sarcasmo. La Tabla 2.1 muestra los tipos de sarcasmo característicos de cada método, los cuales servirán como base la experimentación (B) del presente trabajo.

Método	**Tipo de sarcasmo**
Supervisión Distante	Intencional
Anotación Manual	Percibido
Recolección Manual	Intencional

TABLA 2.1: Tipos de sarcasmo por método de recolección.

2.2.2 Supervisión Reactiva

Propuesto por Shmueli *et al.* (2020), la Supervisión Reactiva, es el método más reciente y el más elaborado. Utiliza los comentarios hechos en réplica a otros en *Twitter*, en los cuales se examina el contenido en busca de una reacción que indique si un *tweet* es sarcástico. Por ejemplo, si se encuentra una réplica con la oración "Ella estaba siendo sarcástica" (traducido del inglés *"She was being sarcastic*") se puede deducir que un *tweet* anterior escrito por el destinatario de la réplica es sarcástico. La Figura 2.1 muestra un ejemplo ilustrado de cómo se obtiene el *tweet* sarcástico a partir de una reacción.

El pseudocódigo del algoritmo utilizado por el método de Supervisión Reactiva se muestra en Algoritmo 1. Como se observa, el algoritmo devuelve un conjunto S de *tweets* sarcásticos, que se recolectan a partir de hilos, donde un hilo es una secuencia de réplicas de *tweets*.

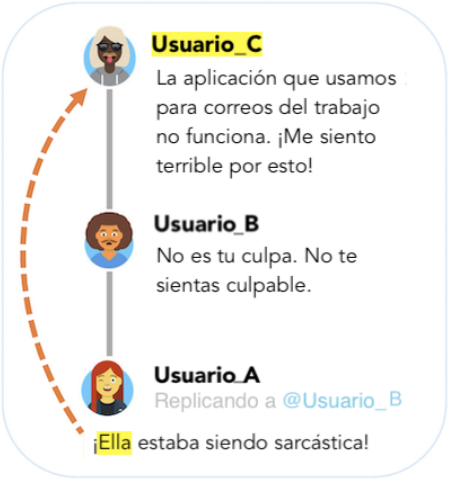


FIGURA 2.1: Un hilo de *Twitter* obtenido con Supervisión Reactiva. El método identifica el *tweet* del Usuario C como sarcástico tras ser referenciado por el pronombre en 3ra persona en el *tweet* indicador (abajo). (Traducido del inglés de Shmueli et al. (Shmueli *et al.*, 2020))

Primero, para empezar a buscar estos hilos, se buscan *tweets candidatos* con la frase *"being sarcastic"* ("siendo sarcástico" en español, línea 3) con la API de *Twitter* y se itera sobre ellos para determinar si se trata de un *indicador* (líneas 3-4). Cada *indicador* se clasifica según la persona gramatical utilizada, con la que se obtiene una expresión regular (*regexp*) que señala al *tweet* sarcástico (líneas 5-11). Para ilustrar qué es y cómo se clasifica un *indicador*, se hará referencia al ejemplo mencionado anteriormente. El *tweet* con la frase "Ella estaba siendo sarcástica" es un *indicador*, pues señala a un *tweet* sarcástico y se clasificaría como 3ra persona gramatical por hacer uso del sujeto "Ella".

Retomando el algoritmo, en las líneas 12 a 13, el *indicador* se descarta y se prosigue al siguiente de no encontrarse una persona gramatical.

Una vez obtenida la expresión regular, se recorre el hilo desde el *indicador* (el último cronológicamente) hasta el primer *tweet*, también llamado tweet raíz (línea 14).

Luego, se recuperan los autores de los *tweets* y se construye una secuencia de letras que describen a los usuarios en el hilo. Para ello se asigna una letra a cada usuario

en el orden que aparecen, empezando por la `A` que es asignada al usuario del *indicador* (línea 15).

Finalmente, para verificar si el *tweet* sarcástico existe y obtenerlo, se compara con la expresión regular de las líneas 6 a 11. De coincidir, el grupo de captura de la expresión indica en qué posición se encuentra el *tweet* sarcástico y se añade al conjunto S (líneas 16-17).

Algorithm 1 Método de Supervisión Reactiva.
Traducido del inglés de Shmueli et al. (Shmueli *et al.*, 2020). Expresiones regulares en notación PCRE.

1: **Resultado:** Conjunto S de *tweets* sarcásticos
2: $S \leftarrow \{\}$
3: $candidatos \leftarrow$ ***Buscar***('*being sarcastic*')
4: **for** $indicador$ **in** $candidatos$ **do**
5: **switch** ***Clasificar(***$indicador$***)*** **do**
6: **case** 1ra persona **do**
7: $regexp \leftarrow$ `^A[^A]*` **`(A)`** `[^A]*$`
8: **case** 2da persona **do**
9: $regexp \leftarrow$ `^AA*` **`(B)`** `A*$`
10: **case** 3ra persona **do**
11: $regexp \leftarrow$ `^AA*B[AB]*` **`(C)`** `[AB]*$`
12: **case** desconocido **do**
13: continue
14: $\{t_n(= indicador), t_{n-1}, ..., t_1\} \leftarrow$ ***Recorrer(***$indicador$***)***
15: $\{a_n, a_{n-1}, ..., a_1\} \leftarrow$ autores($\{t_n, t_{n-1}, ..., t_1\}$)
16: **if** $i \leftarrow$ ***Coincide***$(regexp, \{a_n, a_{n-1}, ..., a_1\})$ **then** ▷ el índice *i* del arreglo de autores indica la posición del *tweet* sarcástico
17: $S \leftarrow S \cup \{t_i\}$
return S

2.2.2.1. Expresiones regulares

Recordando el ejemplo de la Figura 2.1, la secuencia de usuarios correspondiente sería "ABC" y la expresión regular a utilizar sería `A[^A]*(A)[^A]*\$`, al estar en 3ra persona. En este caso, la expresión coincide y se obtiene el *tweet* sarcástico, sin embargo,

si el *tweet* del Usuario_C no existiera, el *indicador* se descartaría junto con el hilo. Asimismo, si el *indicador* se encuentra en 1ra o 2da persona pero no existe un segundo *tweet* del Usuario_A o solo existe un usuario respectivamente, la expresión no coincide y el hilo se descarta.

Las expresiones regulares tampoco coinciden en casos ambiguos, es decir, cuando existe más de un *tweet* del usuario al que le debería pertenecer el *tweet* sarcástico. En el ejemplo anterior, la expresión no coincidiría si hubieran dos *tweets* del Usuario_C.

2.2.2.2. Tipos de *tweet*

El método de Supervisión Reactiva etiqueta a los *tweets* sarcásticos referidos con 1ra persona como sarcasmo intencional y con 2da y 3ra persona como sarcasmo percibido. Adicionalmente, si se encuentra disponible, es posible recolectar el *tweet* que ocasionó el *tweet* sarcástico, llamado *tweet* provocador, y el *tweet* que falló en identificar el sarcasmo, llamado *tweet* distraído. El *tweet* provocador se consigue si el *tweet* sarcástico es una réplica y el *tweet* distraído se obtiene si existe exactamente un *tweet* entre el *tweet* sarcástico y el *tweet indicador* y pertenece a un usuario diferente a los que escribieron los primeros dos.

2.3 Métricas de desempeño

Las siguientes métricas de desempeño han sido utilizadas en la tarea de detección de sarcasmo realizada en el *workshop SemEval* del 2018 (Van Hee *et al.*, 2018) y son ampliamente conocidas en el campo de *Machine Learning*. Para obtener las siguientes métricas, se utilizan los valores de la matriz de confusión obtenida al entrenar o validar el modelo. Los valores contrastan las predicciones realizadas por el modelo con la etiqueta real, resultando en la cantidad de Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN).

2.3.1 *Accuracy*

El *accuracy*, denotado por *acc*, mide la probabilidad que una predicción sea correcta sin discriminar entre VPs y VNs, y se calcula de la siguiente manera:

$$acc = \frac{VP + VN}{VP + VN + FP + FN}. \tag{2.1}$$

La fórmula consiste en dividir la cantidad de Verdaderos entre el total de datos. El *acc* es la métrica principal utilizada para evaluar el desempeño en *datasets* balanceados. Cabe resaltar que el peor valor de *accuracy* no es 0, sino 0.5 pues significa máxima incertidumbre.

2.3.2 *Sensitivity*

El *sensitivity*, denotado por *sens*, mide la probabilidad que la predicción positiva de una muestra positiva sea correcta. El *sensitivity* es la medida principal en situaciones donde una predicción negativa incorrecta puede tener consecuencias catastróficas. Se calcula de la siguiente manera:

$$sens = \frac{VP}{VP + FN}. \tag{2.2}$$

Es posible maximizar esta métrica si un modelo predice todas las muestras como positivas, puesto que se cumpliría que $FN = 0$.

2.3.3 *Specificity*

A diferencia del *sensitivity*, el *specificity*, denotado por *spec*, mide la probabilidad que la predicción negativa de una muestra negativa sea correcta. Su definición es:

$$spec = \frac{VN}{VN + FP}. \tag{2.3}$$

Es posible maximizar esta métrica si un modelo predice todas las muestras como negativas, puesto que se cumpliría que $FP = 0$.

2.3.4 *Balanced Accuracy*

El *balanced accuracy*, denotado por *balacc* es la media aritmética del *sensitivity* y el *specificity*. Nótese que al combinar ambas métricas, no es posible que un modelo obtenga el valor máximo de 1 si todas sus prediciones son iguales: si fueran positivas, el *sensitivity* sería 1, pero el *specificity* sería 0 y viceversa si fueran negativas. En dichos casos, el *balanced accuracy* resultaría en su peor valor, 0.5. Esta definido por:

$$balacc = \frac{sens + spec}{2}. \tag{2.4}$$

Mientras que el *accuracy* se utiliza para *datasets* balanceados, el *balanced accuracy* se utiliza en la presente investigación como principal métrica para *datasets* desbalanceados.

2.4 Modelos de detección de sarcasmo

Los modelos en el estado del arte de detección de sarcasmo hacen uso de la arquitectura de *transformers*. Los *transformers*, propuestos por Vaswani *et al.* (2017), son un modelo paralelizable no recurrente basado en una arquitectura *decoder-encoder*. La paralelización y ausencia de recurrencia reducen el tiempo de entrenamiento significativamente, en contraste con las redes recurrentes de *long short-term memory* (LSTM) usadas previamente. Mientras las redes LSTM se enfocan en retener información y propagarla hasta los últimos pasos en una red recurrente, los *transformers* utilizan el mecanismo

de atención, buscando palabras claves y dependencias entre la secuencia de símbolos de entrada y salida.

El *encoder* de los *transformers* es utilizado en el modelo de representación de lenguaje de *Bidirectional Encoder Representations from Transformers* (BERT). BERT apila un conjunto de *encoders* [1] para generar *word embeddings* dependientes del contexto. Los modelos pre-entrenados de BERT pueden ser utilizados en distintas tareas de PLN, incluyendo la detección de sarcasmo, tras un proceso de *fine tuning*[2] (Devlin *et al.*, 2019).

Posteriormente, mediante una optimización de hiperparámetros, una mayor cantidad de pasos, *batches* y corpus de texto más extensos en el entrenamiento, el modelo de *Robustly Optimized BERT Pretraining Approach* (RoBERTa) consigue mejorar el desempeño de BERT (Liu *et al.*, 2019b).

Finalmente, a partir de RoBERTa, Potamias *et al.* (2020) proponen RCNN-RoBERTa, que mejora el desempeño de los modelos mencionados al agregar una capa de una red LSTM entre RoBERTa y la capa de activación de salida.

2.5 Consideraciones finales

En este capítulo se han explicado los dos tipos de sarcasmo y los métodos de recolección con los que se trabajan a lo largo de los experimentos. De los métodos resalta el de Supervisión Reactiva por su algoritmo novedoso y capacidad de diferenciar entre tipos. Asimismo, cabe resaltar que las relaciones entre métodos y tipos de sarcasmo expuestas en la Tabla 2.1 se utilizan en el experimento B.

[1] BERT_{base} utiliza 12 encoders y BERT_{large}, 24

[2] Re-entrenar un modelo con un nuevo *dataset* para utilizarse en una tarea específica.

Por otro lado, las métricas de desempeño son usadas en los resultados para evaluar a los modelos de detección de sarcasmo mencionados, servirán para obtener datos cuantitativos en los experimentos. El *accuracy* se utilizará para *datasets* balanceados, mientras el *balanced accuracy* para *datasets* desbalanceados.

Capítulo 3

Trabajos Relacionados

En el presente capítulo se analizan los trabajos relacionados más resaltantes que tratan sobre los métodos de recolección y su relación con el desempeño de los modelos de detección de sarcasmo. Asimismo, se señalan sus limitaciones y una apreciación de sus estudios.

3.1 Tipos de *corpus*

En Joshi *et al.* (2016b) se hace una recopilación de *corpus* agrupados por pequeña extensión, gran extensión y otros. Bagate y Ramadass (2020) agrega un nuevo grupo de diálogos.

Dentro del grupo de gran extensión, otros y diálogos se encuentran foros de discusión (Lukin y Walker, 2017), publicaciones en *Reddit* (Wallace *et al.*, 2015), extractos de *Google Book Search* (Kreuz y Caucci, 2007), entre otros. Al contener conversaciones y varias oraciones además de la oración sarcástica, son utilizados para modelos que requieren contexto como Hazarika *et al.* (2018). Sin embargo, al ser más difíciles de conseguir, son menos frecuentes.

En contraste, los *corpus* de pequeña extensión son los más abundantes y populares en el área. En ambos *surveys*, este tipo se compone por *tweets*, debido a su sencilla obtención y que contienen solo una oración sarcástica, usualmente identificable sin necesidad de contexto. En la sección siguiente se puede evidenciar su popularidad en que los estudios sobre métodos de recolección utilizan principalmente *tweets* para la creación de un *corpus*.

3.2 Métodos de recolección

3.2.1 Supervisión Distante

La investigación de métodos para la recolección de textos sarcásticos es una fase previa indispensable para el área de la detección de sarcasmo textual. Por ello, se encuentra presente desde los inicios de esta área.

El primer método propuesto, Supervisión Distante, fue utilizado y evaluado por Davidov *et al.* (2010). En su estudio se recolectaron 5.9 millones de *tweets* y de ellos se examinaron 1500 con el *hashtag "#sarcasm"*, con los que descubrieron que los usuarios utilizaban el *hashtag* para clarificar que otro *tweet* era sarcástico o como marcador para encontrar el *tweet*.

El método es rudimentario y no captura sarcasmo percibido, solo intencional. Los mismos autores del estudio advierten que el *hashtag* suele ser colocado cuando el sarcasmo es casi imperceptible o carece de contexto y que ello produce datos sesgados. Sin embargo, estas deficiencias aparentemente han sido desestimadas a cambio de la simplicidad y capacidad del método de crear un *corpus* de gran tamaño (Ghosh *et al.*, 2015, Ghosh y Veale, 2017, Khodak *et al.*, 2018, Ptáček *et al.*, 2014), lo que a conducido a trabajos recientes (Cai *et al.*, 2019, Liu *et al.*, 2019a, Pelser y Murrell, 2019, Potamias *et al.*, 2020, Wu *et al.*, 2018) a entrenar y verificar modelos sin considerar el tipo de sarcasmo y la baja calidad del *dataset*.

3.2.2 Anotación Manual

En años posteriores a la creación de la Supervisión Distante, se publicaron *datasets* generados con el método de Anotación Manual (Abercrombie y Hovy, 2016, Riloff *et al.*, 2013), el cual utiliza anotadores humanos para una clasificación más fiable.

Riloff *et al.* (2013) presentan los *tweets* sin contexto alguno a sus tres anotadores y para asegurar que su clasificación sea correcta, se les pide revisar un mismo conjunto de 200 *tweets*, de donde se obtiene una alta concordancia[1].

Las precauciones que toman los autores para evitar sesgos en sus clasificaciones son insuficiente, ya que el contexto cultural de los anotadores puede ser un factor que afecte sus resultados (Joshi *et al.*, 2016a, Poria *et al.*, 2017). En su investigación, no proporcionan información alguna sobre los anotadores. Además, mencionan que si el *tweet* era originalmente sarcástico, pero para reconocerlo era necesario tener un contexto adicional, se catalogaba directamente como “no sarcástico".

Por su lado, Abercrombie y Hovy (2016) piden a 60 voluntarios anglosajones clasificar 650212 *tweets* y obtienen una concordancia muy baja. La concordancia es aún menor cuando se les provee el contexto de la conversación en *Twitter* e información del autor[2].

En el mismo estudio se admiten desventajas del método, entre las que se encuentran que los anotadores pueden haber sido deshonestos o que el sarcasmo es una tarea muy complicada para humanos.

A la Anotación Manual hay que añadir que suele utilizar *tweets* obtenidos mediante Supervisión Distante y, en consecuencia, propaga el sesgo que podría existir en ellos.

Por otro lado, el trabajo de catalogar miles de textos es abrumante para una pequeña cantidad de investigadores, por lo que a veces en la Anotación Manual se hace uso de voluntarios en plataformas de *crowdsourcing*. Estas aceleran y facilitan el proceso

[1] Se compara las clasificaciones entre pares de anotadores y se obtiene un promedio de coeficiente Kappa de Cohen de 0.81. Este coeficiente mide las concordancias entre anotadores, tomando en cuenta que pudieron ocurrir por azar. Un valor negativo equivale a concordancia inversa, de 0 que no hay concordancia y 1 que hay una concordancia total.

[2] Se utiliza el coeficiente Alfa de Krippendorff, utilizada también para medir concordancias. Los valores negativos, 0 y 1 tienen el mismo significado que Kappa de Cohen. En el estudio se obtuvo 0.35 para *tweets* anotados sin contexto y 0.18 para *tweets* con contexto e información del autor y audiencia.

de anotación, mas abren paso a estafadores que corrompen los datos con clasificaciones aleatorias. Con el fin de contrarrestar sus efectos, se acostumbra tomar medidas para identificar datos sospechosos. En el *dataset* utilizado para el ejercicio 11[3] del taller de Evaluación Semántica del 2015, como paso previo a la publicación de los textos a clasificar, 1025 *tweets* de 12 000 fueron etiquetados por investigadores del equipo. Los *tweets* fueron luego expuestos al público y los resultados de los usuarios que tuvieran datos muy diferentes fueron descartados. Pese a ello, los investigadores reconocen que es imposible eliminar por completo estos datos y que una pequeña cantidad pudo haberse filtrado o que clasificaciones honestas se podrían haber eliminado (Ghosh *et al.*, 2015).

3.2.3 Recolección Manual

En Oprea y Magdy (2019b) se utiliza por primera vez el método de Recolección Manual. En él se solicita a voluntarios a crear *tweets* sarcásticos, en contraste con la Anotación Manual donde se les solicita clasificar. De esta manera se elimina el sesgo presente desde la Supervisión Distante y, asumiendo que los voluntarios son honestos, se garantiza que los textos estén anotados correctamente.

No obstante, al igual que la Supervisión Distante, la Recolección Manual solo produce sarcasmo intencional y el trabajo manual necesario es aún mayor que el de Anotación Manual. Debido a su costo significativo, iSarcasm es el único *dataset* publicado generado con este método.

3.2.4 Supervisión Reactiva

El método más reciente y prometedor soluciona problemas existentes en los métodos anteriores. La Supervisión Reactiva aprovecha las reacciones de los mismos usuarios y permite etiquetar tomando en cuenta el contexto, conocido por los usuarios participantes

[3]El ejercicio 11 corresponde al Análisis de Sentimiento de Lenguaje Figurado en *Twitter*

de la conversación. Asimismo, elimina el sesgo ocasionado por un sarcasmo muy sutil, puesto que ha sido reconocido por otros usuarios.

Ante todo, la característica más resaltante de la Supervisión Reactiva es extraer el tipo de sarcasmo de cada texto, que ninguno de los métodos vistos considera. Cabe resaltar que el método incluso captura una mayor cantidad de *tweets* que la Supervisión Distante: el estudio afirma que la velocidad de recolección promedio (312 *tweets* por día) fue comparada con la de Supervisión Distante (171 *tweets* por día) y tuvo una mejora del 82 % (Shmueli *et al.*, 2020).

Aun cuando es la opción más confiable para la creación de un *dataset* sarcástico, solo el estudio de Plepi y Flek (2021) lo ha utilizado para comparar tipos de sarcasmo, el cual será expuesto en la sección siguiente.

3.3 Comparaciones entre tipos de sarcasmo

Mientras que la Supervisión Distante recolecta textos con sarcasmo intencional, el método de Anotación Manual recolecta sarcasmo percibido. Dicho contraste posibilitó investigar a mayor profundidad la diferencia entre los tipos de sarcasmo en los estudios presentados a continuación.

Abercrombie y Hovy (2016) compararon el efecto de *datasets* de *tweets* sarcásticos con réplicas generados con ambos métodos en un modelo de Regresión Logística, teniendo como variables al contexto [4] proporcionado y datos balanceados y desbalanceados. En los resultados (Figura 3.1) se muestra una tendencia a un menor desempeño al entrenar con Anotación Manual. Este es uno de los primeros estudios con métodos tradicionales que muestran la importancia de diferenciar entre tipos de sarcasmo.

[4] Información sobre los usuarios involucrados

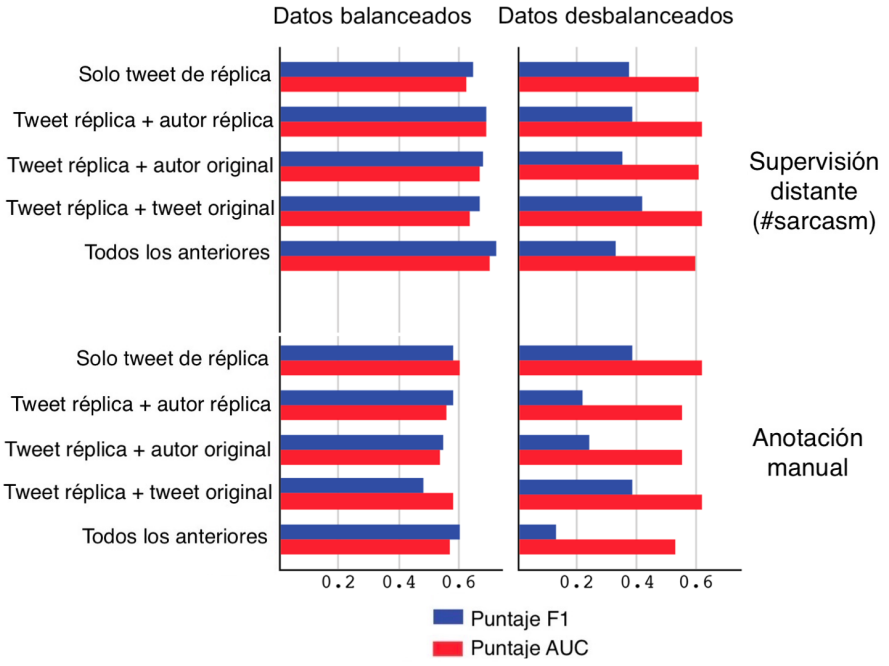


FIGURA 3.1: Comparación entre el desempeño de un modelo de regresión entrenado con *tweets* y su réplica clasificados con Supervisión Distante y con Anotación Manual, y con datos balanceados y desbalanceados. Se muestra el desempeño al proporcionar (de arriba hacia abajo) solo el *tweet* de réplica, el *tweet* de réplica y su usuario, el *tweet* de réplica y el usuario del *tweet* original, el *tweet* de réplica y el *tweet* original y todos los datos mencionados (Abercrombie y Hovy, 2016).

En este estudio solo se utilizaron métodos tradicionales pues el método de Supervisión Reactiva fue creado posteriormente y, como se mencionó en la subsección 3.2.2, la clasificación realizada por los anotadores no es muy confiable por su baja concordancia.

Asimismo, Oprea y Magdy (2019b) analizaron el *dataset* de Riloff *et al.* (2013), recolectado con Supervisión Distante y reclasificado con Anotación Manual. En los resultados encontraron una discrepancia (Tabla 1.1) entre los *tweets* con el *hashtag "#sarcasm"*, i.e. sarcasmo intencional; y percibidos como sarcásticos, i.e. sarcasmo percibido.

Por último, Plepi y Flek (2021) hacen uso de SPIRS, el *dataset* publicado junto a la propuesta de la Supervisión Reactiva. En su estudio crean, entrenan y validan un modelo para la detección sarcasmo y de tipos de sarcasmo utilizando *Graph Attention Networks* para modelar relaciones entre los usuarios de *Twitter* que escribieron o percibieron un

tweet sarcástico. Sus resultados en detección de sarcasmo muestran un error del 20 % en sarcasmo percibido y 15 % en sarcasmo intencional. En la detección de tipos de sarcasmo obtienen un *F1* de 56 % en sarcasmo percibido y 84 % en sarcasmo intencional. Como se menciona en el estudio, ello sugiere que el sarcasmo percibido es más difícil de identificar, pues suele ser confundido con sarcasmo intencional. Estas conclusiones son similares a las señaladas en los estudios previamente mencionados de Abercrombie y Hovy (2016) y Oprea y Magdy (2019b).

A pesar que los resultados de estos tres estudios muestran que existe una diferencia entre tipos de sarcasmo, mas no realizan una comparación elaborada entre todos los métodos de recolección, con diferentes *datasets* y modelos en el estado del arte. Además, no prueban si es que modelos entrenados con un tipo de sarcasmo son mejores prediciendo este tipo de sarcasmo.

3.4 Resumen del capítulo

En el capítulo se presentaron los tipos de *corpus* y *datasets* utilizados en el área de detección de sarcasmo, siendo los *tweets*, de pequeña extensión los más populares. Asimismo, se presentaron los estudios donde se propusieron los métodos de recolección por primera vez, desde el más antiguo (Davidov *et al.*, 2010) al más reciente (Shmueli *et al.*, 2020), que tiene un rol primordial en los experimentos mostrados en el siguiente capítulo. Por último, se mencionaron estudios similares al presente (Abercrombie y Hovy, 2016, Oprea y Magdy, 2019b, Plepi y Flek, 2021), que realizan observaciones sobre los tipos de sarcasmo, pero donde la comparación no es su objetivo principal y, por lo tanto, no contrastan diferentes modelos, *datasets* o toman en cuenta todos los métodos de recolección.

Capítulo 4

Metodología

En este capítulo se profundizará sobre los pasos preliminares realizados para la ejecución de los experimentos, así como una descripción de lo realizado en estos últimos. Se proveen los detalles necesarios que permitan replicar los resultados de la presente investigación.

Dado distintos *datasets* generados con métodos de recolección diferentes, en la presente investigación, se analiza su impacto en el desempeño de modelos de detección de sarcasmo. Con este propósito, se realizaron dos experimentos: el primero (A) se centra en contrastar los modelos entrenados con *datasets* de distintos métodos, mientras el segundo (B) profundiza en el contraste del desempeño de los modelos por cada tipo de sarcasmo. En ambos experimentos se utilizó más de un modelo (BERT y RoBERTa) para comprobar que los resultados encontrados no se limitan a uno en específico.

Durante los experimentos se emplearon los métodos tradicionales (Supervisión Distante, Anotación Manual y Recolección Manual) y el de Supervisión Reactiva. Asimismo, se utilizaron los modelos de BERT y de RoBERTa con una capa para clasificación de textos, red neuronal lineal, sobre la cabeza de BERT.

La metodología puede ser escrita en líneas generales de la siguiente forma.

1. Primero, se obtienen los *datasets*, sea que se encuentren disponibles públicamente en otras investigaciones (para ambos experimentos) o que se tengan que generar (para el Experimento B).

2. Segundo, se limpian los textos, reduciendo el ruido. Este paso aplica para ambos experimentos.

3. Tercero, se realiza la división de datos que se usarán en divisiones de entrenamiento, validación y prueba para entrenar modelos en el Experimento A. Además se divide el *dataset* de SPIRS para obtener dos *datasets* de sarcasmo percibido e intencional.

4. Cuarto, el *corpus* del paso anterior y el generado para el Experimento B se tokeniza. Los pasos hasta aquí son los pasos preliminares necesarios para servir de dato de entrada a los modelos.

5. Quinto, los *datasets* que fueron divididos en el tercer paso se utilizan para realizar un *fine-tuning* de los modelos y buscar hiperparámetros que optimicen su desempeño.

6. Sexto, con los modelos entrenados del paso anterior y el *corpus* generado para el Experimento B, tokenizado en el cuarto paso, se utilizan para los experimentos.

La Figura 4.1 muestra gráficamente estos pasos, los cuales se explicarán detalladamente en las siguientes subsecciones.

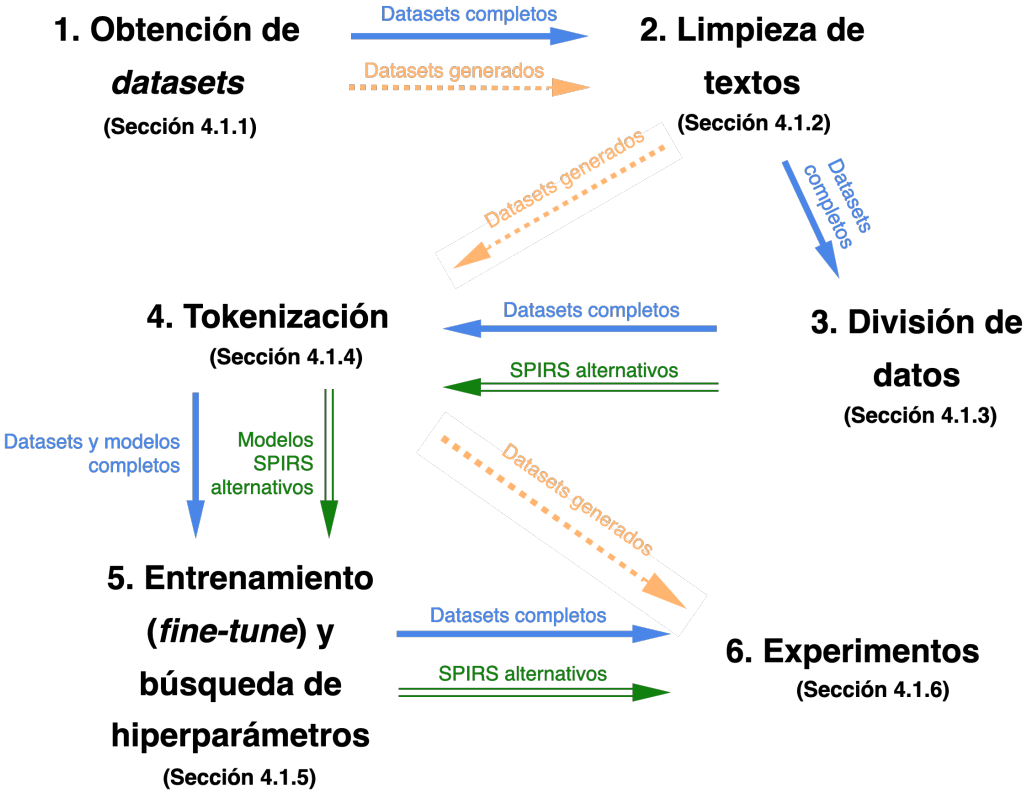


FIGURA 4.1: Resumen de la metodología. Se muestran los pasos realizados con su sección correspondiente, así como los datos necesarios para cada uno y su producto. Los primeros cuatro pasos son preliminares para poder entrenar y validar los modelos usados en los experimentos. Las flechas con línea continua en azul señalan los *datasets* completos, asi como los modelos entrenados con ellos; las flechas con doble línea en verde, los *datasets* y modelos de los de SPIRS alternativo; y las flechas punteadas en naranja, los *datasets* generados.

4.1 Descripción de la Metodología

4.1.1 Obtención de *datasets*

4.1.1.1. *Datasets* completos

Como primer paso para ambos experimentos, se eligió un *dataset* representativo por cada método. Por ser los más utilizados y de mayor tamaño, para Supervisión Distante y Anotación Manual se utilizaron los *datasets* de Ptáček *et al.* (2014) y Riloff *et al.* (2013). Mientras tanto, por ser los únicos de su clase, para Recolección Manual y Supervisión Reactiva se usaron iSarcasm (Oprea y Magdy, 2019b) y SPIRS (Shmueli *et al.*, 2020). Todos los *datasets* están disponibles públicamente en sus publicaciones respectivas.

Debido a la política de desarrolladores de *Twitter*[1], solo se proporcionan los identificadores de los *tweets*, por lo que para el presente trabajo se tuvo que solicitar acceso a su *API* para obtener los textos. Puesto que los *tweets* pueden ser eliminados por sus autores o por la plataforma misma, no todos pudieron ser recuperados. El *dataset* más afectado fue el de Riloff, cuya cantidad de *tweets* disminuyó casi a la mitad, de 3000 a 1563.

Estos *datasets* serán utilizados en el experimento A y serán llamados *datasets* completos.

4.1.1.2. *Datasets* generados

Adicionalmente, con el fin de probar los modelos entrenados con los *datasets* previamente mencionados, el experimento B requiere un *dataset* por cada tipo de sarcasmo, intencional y percibido. Tomando en cuenta que no se podían utilizar los mismos textos que en el entrenamiento, se crearon dos nuevos *datasets*.

[1]Disponible en `https://developer.twitter.com/es/developer-terms/policy`

Primero se obtuvieron *tweets* sarcásticos con el algoritmo de Supervisión Reactiva, descrito en la sección 2.2.2. Luego, los *tweets* se divieron por tipo de sarcasmo para formar las muestras positivas. Las muestras negativas se obtuvieron de los últimos textos[2] de SPIRS, que no se utilizaron para entrenar los modelos.

De esta manera, se consiguieron 1848 *tweets* con sarcasmo percibido y 1374 con sarcasmo intencional. A fin de crear *datasets* comparables, se tomó el tamaño mínimo para ambos (1374) y se rellenó con las mismas muestras negativas, de modo que la única diferencia entre los *datasets* son las muestras positivas.

Estos nuevos *datasets* serán utilizados en el experimento B y serán llamados *datasets* generados. Se encuentran disponibles públicamente en `https://github.com/a4vg/datasets-reactive-supervision`

La Figura 4.2 muestra los pasos descritos anteriormente: la obtención de textos de los *datasets* completos con la API de *Twitter* y el uso del algoritmo de Supervisión Reactiva para obtener los *datasets* generados.

4.1.2 Limpieza de textos

Una vez que se obtuvieron los *datasets* directamente de *Twitter*, se realizó una limpieza de los textos. Esta consistió en remover las palabras que comienzan por "http" (hipervínculos), "@" (nombre de usuarios) y "#" (*hashtags*), incluyendo "#sarcasm" y similares. Ello con el fin de reducir el ruido y aumentar la legibilidad de los textos. Además, también se removieron los espacios en blanco sobrantes al principio y al final del texto.

Los textos que acabaron vacíos se eliminaron del *dataset* y los restantes se convirtieron a minúsculas.

[2]Se resalta que fueron los *últimos* textos no sarcásticos pues en la sección 4.1.3, en la creación de los *datasets* alternativos de SPIRS, se utilizan las primeras muestras sarcásticas. De esta manera, se evita utilizar las mismas muestras negativas en el entrenamiento y prueba realizadas en el experimento B.

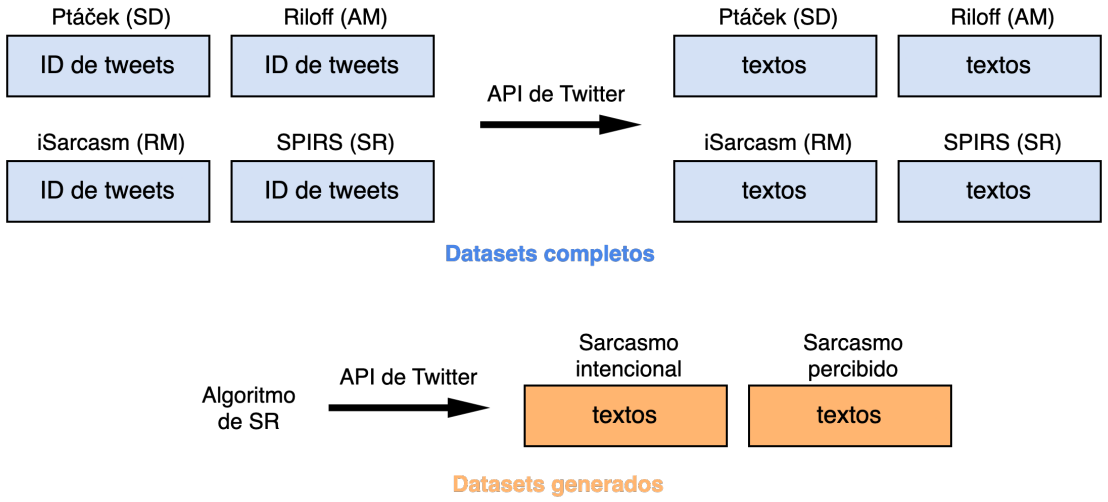


FIGURA 4.2: Obtención de *datasets* completos (en azul) y generados (en naranja). Para los primeros, se utiliza la *API* de *Twitter* para buscar los textos de los *tweets* con su ID; para los segundos se utiliza el algoritmo de Supervisión Reactiva y *Twitter* para generar *datasets* de sarcasmo intencional y percibido por separado. SD es Supervisión Distante, AM es Anotación Manual, RM es Recoleccion Manual y SR es Supervisión Reactiva.

4.1.3 División de datos

En este paso se parten los *datasets* completos en tres divisiones: entrenamiento, validación y prueba, en proporciones de 70 %, 20 % y 10 % respectivamente. Por cada división se intentó mantener los *datasets* balanceados y con la misma cantidad de muestras, sin reducir significativamente la cantidad de textos originales. La Figura 4.3 muestra de manera general las divisiones con porcentajes de los *datasets* completos, usadas en ambos experimentos.

Como se observa en la figura, en estas divisiones, los *datasets* de mayor extensión, SPIRS y Ptáček, presentan exactamente la misma configuración, es decir, el mismo número de muestras negativas y positivas. Nótese que la diferencia entre la cantidad de muestras por clase es muy similar, con tan solo unos pocos textos desbalanceados.

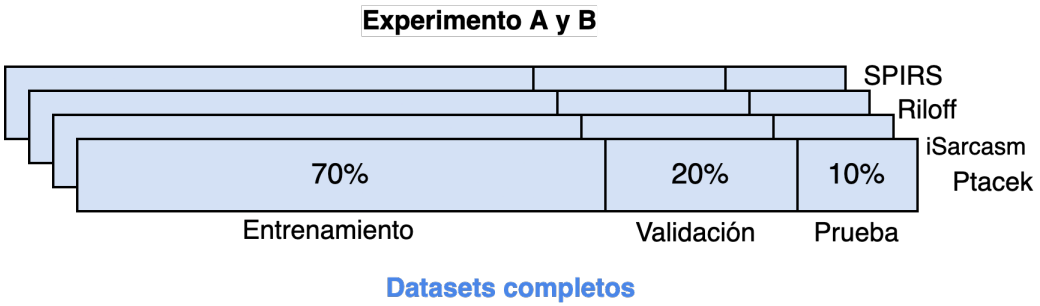


FIGURA 4.3: División de datos de los *datasets* completos (en azul).

Por el contrario, no fue posible equilibrar las clases en los *datasets* más pequeños de iSarcasm y Riloff, ya que el número de *tweets* hubiera sido muy reducido.

Las configuraciones exactas de las muestras en cada *dataset* se muestran en las Tablas 4.1. Como se mencionó previamente, el desbalance en los *datasets* más extensos de SPIRS y Ptáček (2 % de textos desbalanceados) es mucho menor que en los más pequeños de iSarcasm y Riloff (66 % y 58 % respectivamente).

El tamaño reducido de los *datasets* de iSarcasm y Riloff hace evidente una de las desventajas de los métodos de Recolección y Anotación Manual: el costo considerablemente alto de generar textos sarcásticos manualmente. Pese a que originalmente el *dataset* de Riloff tenía cerca del doble de textos, como se mencionó en la sección 4.1.1, esta cantidad continúa siendo sumamente pequeña en comparación con SPIRS y Ptáček, recolectados con métodos no manuales.

Respecto al experimento B, los *datasets* generados no son divididos pues no se utilizaron para el entrenamiento, son una división exclusivamente de prueba. Sin embargo, para lograr el objetivo de comparar los métodos de recolección por tipo de sarcasmo, se hizo uso del etiquetado de tipos en SPIRS. Así, se aprovechó el potencial de la Supervisión Reactiva y, los *tweets* de SPIRS se separaron en dos *datasets*. En la Figura 4.4 se puede apreciar el proceso realizado para esta separación, el cual será explicado a continuación.

	Muestras			
División	**N**	**P**	**D**	**Total**
Entrenamiento (70 %)	7384 (51 %)	7127 (49 %)	257 (2 %)	**14511**
Validación (20 %)	2110 (51 %)	2036 (49 %)	74 (2 %)	**4146**
Prueba (10 %)	1054 (51 %)	1019 (49 %)	35 (2 %)	**2073**
Total (100 %)	**10548**	**10182**	**368**	**20730**

(A) Configuración de muestras de SPIRS y Ptáček. Extenso y ligeramente desbalanceado.

	Muestras			
División	**N**	**P**	**D**	**Total**
Entrenamiento (70 %)	1969 (83 %)	405 (17 %)	1564 (66 %)	**2374**
Validación (20 %)	563 (83 %)	116 (17 %)	447 (66 %)	**679**
Prueba (10 %)	281(83 %)	58 (17 %)	223 (66 %)	**339**
Total (100 %)	**2813**	**579**	**2234**	**3392**

(B) Configuración de muestras de iSarcasm. Pequeño y altamente desbalanceado

	Muestras			
División	**N**	**P**	**D**	**Total**
Entrenamiento (70 %)	866 (79 %)	228 (21 %)	638 (58 %)	**1094**
Validación (20 %)	247 (79 %)	65 (21 %)	182 (58 %)	**312**
Prueba (10 %)	124 (79 %)	33 (21 %)	91 (58 %)	**157**
Total (100 %)	**1237**	**326**	**911**	**1563**

(C) Configuración de muestras de Riloff. Pequeño y altamente desbalanceado

TABLA 4.1: Configuración de muestras por *dataset*. Se indica la cantidad de muestras por clase negativa (N), positiva (P), la cantidad que representan las muestras desbalanceadas (D) y el total de muestras por división y clase. Los porcentajes mostrados junto a las cantidades de muestras son relativos a cada división.

Partiendo de los textos sarcásticos del *dataset* completo de SPIRS, se obtuvieron *tweets* con sarcasmo intencional y percibido. Al igual que con los *datasets* generados, se eligió el que contenía una menor cantidad de textos. El elegido fue el percibido y los *tweets* de sarcasmo intencional tuvieron que reducirse de 6795 a 3387.

Posteriomente, se obtuvieron los primeros 3387 *tweets* no sarcásticos de SPIRS para complementar las 3387 muestras positivas en cada tipo y formar dos nuevos *datasets* que serán referidos como "*datasets* alternativos de SPIRS" o individualmente como

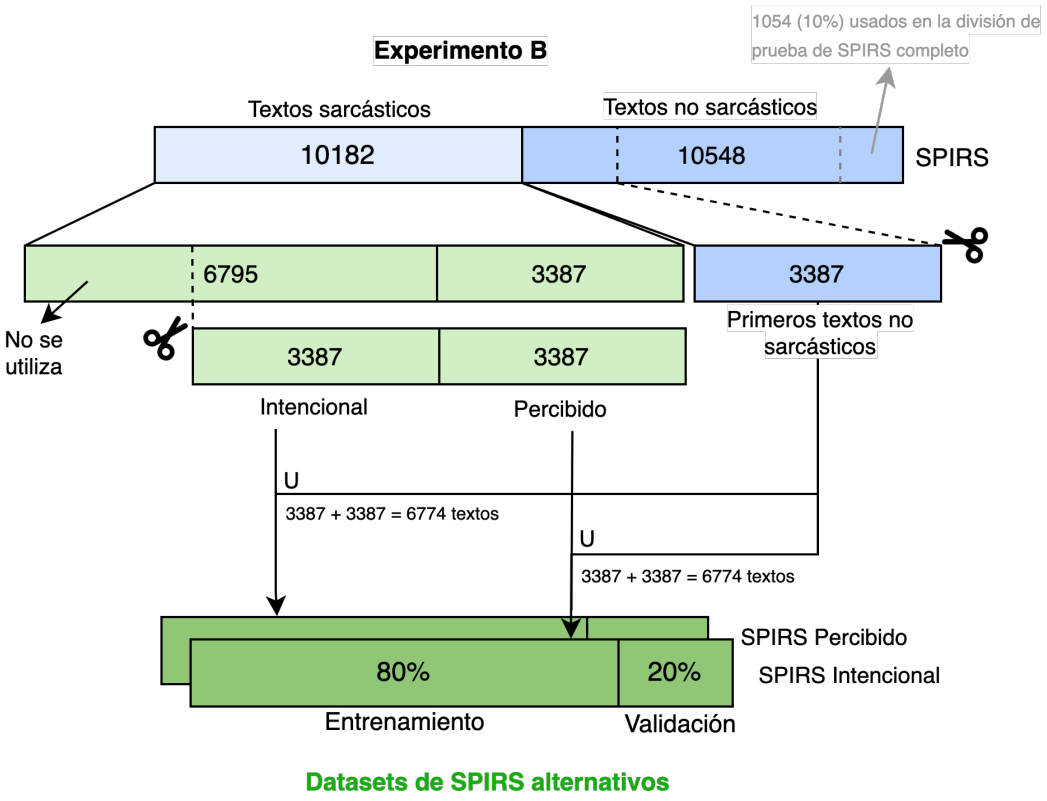


FIGURA 4.4: Proceso de división de datos de los *datasets* de SPIRS alternativos. Se extrajo una porción de textos sarcásticos y no sarcásticos del *dataset* completo de SPIRS (en azul) para crear SPIRS Percibido e Intencional (en verde). Las tijeras indican que el *dataset* se truncó.
Se utilizaron los primeros textos no sarcásticos para no repetir los textos de prueba extraídos del final de SPIRS completo (de la Figura 4.3 y Tabla 4.1).

SPIRS percibido y SPIRS intencional. Estos se utilizaron únicamente para entrenar modelos, por ello, se dividió en 80 % de entrenamiento y 20 % de validación, prescindiendo de la división de prueba. La configuración final puede verse en la Tabla 4.2.

4.1.4 Tokenización

Antes de utilizar los *datasets* en los modelos, ya sea para entrenarlos o probarlos, todos sus textos debieron ser tokenizados.

	Muestras		
División	**N**	**P**	**Total**
Entrenamiento (80 %)	2710	2710	**5420**
Validación (20 %)	677	677	**1354**
Total (100 %)	**3387**	**3387**	**6774**

TABLA 4.2: Configuración de muestras en *datasets* de SPIRS alternativos en clase positiva (P) y negativa (N). La configuración es igual tanto para SPIRS Intencional como SPIRS Percibido. No se incluye una división para prueba pues para ello se utilizaron los *datasets* generados. Los *datasets* se balancearon con muestras negativas de los primeros textos no sarcásticos de SPIRS.

Para convertir los textos al formato de entrada de BERT y RoBERTa, se tokenizaron y codificaron las palabras con `BertTokenizer` y `RobertaTokenizer` de la librería *Transformers* de *HuggingFace*[3].

El vocabulario para estos tokenizadores se obtuvo de los modelos preentrenados "*bert base uncased*" (Devlin *et al.*, 2019) y "*roberta base*" (Liu *et al.*, 2019b). Nótese que se eligió la versión *uncased* de BERT pues coincide con la conversión a minúsculas realizada durante la limpieza de textos.

Durante la tokenización, los textos son divididos en palabras que, si no fueran parte del vocabulario, son divididas en raíces, sílabas y hasta caracteres para coincidir con algún token conocido. Para caracteres desconocidos y para demarcar el inicio y final de un texto se utilizan tokens especiales que dependen del modelo. Por ejemplo, BERT utiliza `[CLS]` y `[SEP]` para el inicio y fin de un texto (Devlin *et al.*, 2019), mientras que RoBERTa utiliza `<s>` y `</s>` (Liu *et al.*, 2019b). Posteriormente, los tokens se codifican numéricamente para ser introducidos al modelo.

Los modelos requieren una misma cantidad de tokens, de manera que las codificaciones se rellenan con ceros a la derecha o se truncan de acuerdo a si faltan o se exceden. La cantidad máxima utilizada fue de 128 tokens, debido a la corta longitud de los *tweets*.

[3]Disponible en `https://huggingface.co/docs/transformers`.

4.1.5 Entrenamiento (*fine-tuning*) y búsqueda de hiperparámetros

En este paso se entrenaron los modelos de BERT y RoBERTa con los *dataset* completos y alternativos de SPIRS. Así, por cada *dataset* de entrenamiento existe un modelo equivalente en BERT y RoBERTa. Los *datasets* generados no son utilizados en este paso pues solo existen para probar los modelos una vez entrenados.

A diferencia de los modelos convencionales de *Machine Learning*, los modelos de *transformers* han sido preentrenados para propósitos generales, por lo que para un propósito específico como la detección de sarcasmo solo se tuvo que realizar un *fine-tuning*.

Los parámetros para realizar el *fine-tuning* se eligieron primero para BERT y luego fueron aplicados a RoBERTa. Se partió de los valores recomendados en Mosbach *et al.* (2021) y los predeterminados en el optimizador AdamW de la librería Pytorch (`torch.optim.AdamW`). Además, se realizaron distintas pruebas variando el *learning rate*, *dropout*, tamaño de *batch* y épocas.

Debido a que los hiperparámetros son dependientes del tamaño y los *datasets* presentaban tamaños desiguales, se agruparon en *datasets* grandes (Ptáček y SPIRS), medianos (alternativos de SPIRS) y pequeños (iSarcasm y Riloff).

4.1.6 Experimentos

Los experimentos consistieron en probar los modelos de BERT y RoBERTa entrenados con los *datasets* de prueba procedentes de la división de *datasets* públicos (Experimento A) y los *datasets* generados (Experimento B). La métrica de desempeño principal utilizada en los *datasets* balanceados fue el *accuracy*, mientras que en los desbalanceados se utilizó el *balanced accuracy*.

4.1.6.1. Experimento A

Para el experimento A solo se utilizaron los *datasets* completos, pues se intentan comparar las métricas de desempeño en cada método.

Los modelos se prueban con *datasets* del mismo método con el que fueron entrenados. Por ejemplo, el modelo entrenado con la división de entrenamiento del *dataset* de Ptáček (Supervisión Distante), se prueba con la división de prueba del mismo *dataset* de Ptáček. La Figura 4.5 muestra todos los casos probados.

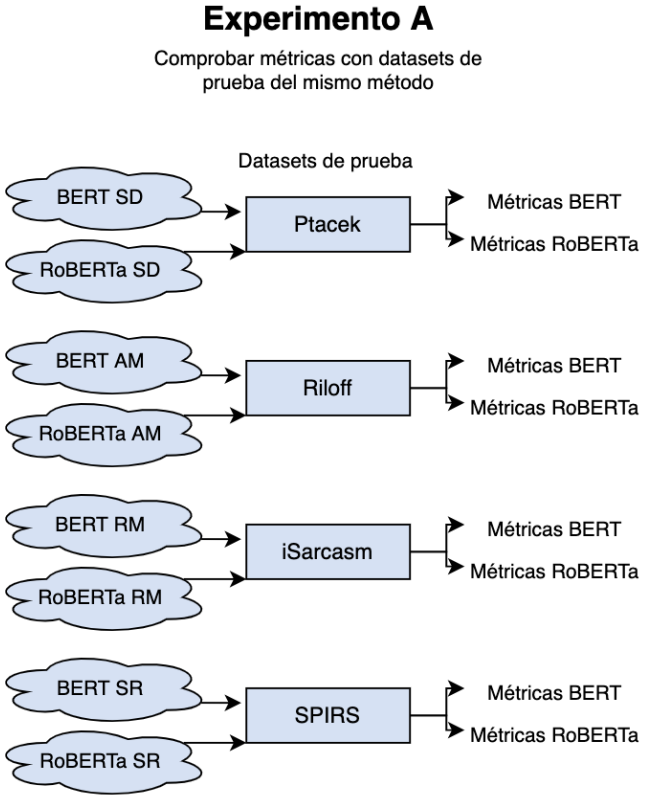


FIGURA 4.5: Casos probados en el experimento A. SD es Supervisión Distante, AM es Anotación Manual, RM es Recoleccion Manual y SR es Supervisión Reactiva.

4.1.6.2. Experimento B

En el experimento B se busca contrastar el desempeño de un modelo al detectar el tipo de sarcasmo con el que fue entrenado y con el que no fue entrenado. Para ello se usaron los modelos entrenados con los *datasets* completos de métodos tradicionales y los *datasets* alternativos de SPIRS.

Cada modelo se prueba con los dos *datasets* generados. La Figura 4.6 muestra los casos probados y señala el tipo de sarcasmo recolectado por los métodos.

4.2 Alcances y Limitaciones

De acuerdo con el objetivo principal de la presente investigación (sección 1.4.1), la prioridad al entrenar los modelos no fue el de lograr las mejores métricas o un ajuste perfecto para un solo modelo, sino el mantener hiperparámetros similares entre *datasets* para realizar una comparación justa.

No obstante, como se mencionó anteriormente, mantener un conjunto idéntico de hiperparámetros para todos los *datasets* no es posible debido a sus distintos tamaños. Esta diferencia de tamaños se debe a la naturaleza de los mismos métodos: los *datasets* recolectados con métodos manuales (Anotación y Recolección Manual) son inherentemente más costosos que los automatizados (Supervisión Distante y Reactiva).

4.3 Resumen del capítulo

En el capítulo se brindaron datos y explicaron los pasos preliminares para realizar los experimentos A y B. Los primeros pasos se centraron en la recolección, limpieza y división de datos para formar los *datasets* que fueron utilizados para entrenar y probar modelos. De esta manera, se obtuvieron los *datasets* completos, usados para el entrenamiento

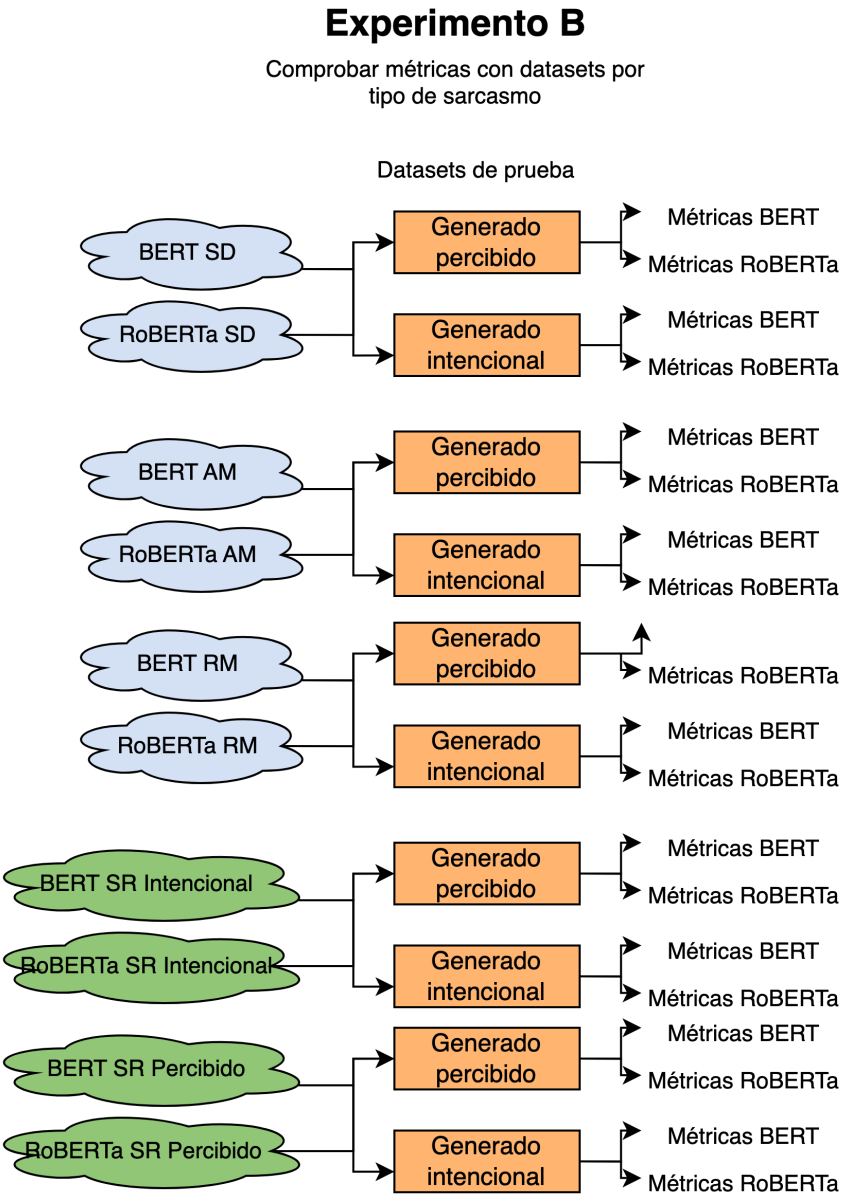


FIGURA 4.6: Casos probados en el experimento B. SD es Supervisión Distante, AM es Anotación Manual, RM es Recoleccion Manual y SR es Supervisión Reactiva.

de modelos necesarios para ambos experimentos; y los *datasets* generados y alternativos de SPIRS para el experimento B.

Los últimos pasos se centraron en la experimentación, buscando los hiperparámetros de entrenamiento que favorezcan a un grupo de *datasets* y finalmente probando los modelos en los experimentos. Los resultados de dichos pasos se encuentran a continuación, en el siguiente capítulo.

Capítulo 5

Experimentaciones y Resultados

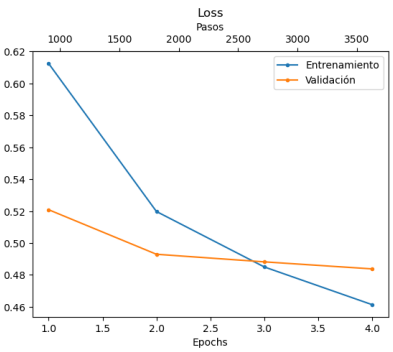
En el presente capítulo se expondrán los resultados obtenidos al efectuar los pasos descritos en el capítulo anterior. Primero, se proveen los parámetros utilizados en el entrenamiento, junto a un análisis de los impedimentos encontrados en el grupo de *datasets* pequeños.

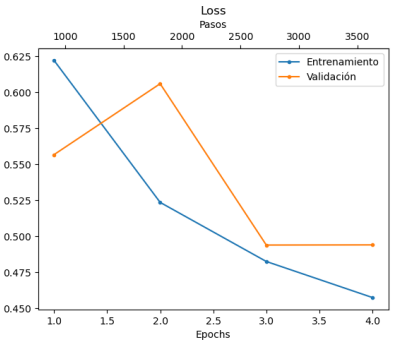
Posteriormente, se presentan las métricas obtenidas en los experimentos y una explicación de ellas. Por último, se incluye una subsección de discusión, donde se formulan posibles implicaciones de los resultados.

5.1 Búsqueda de hiperparámetros

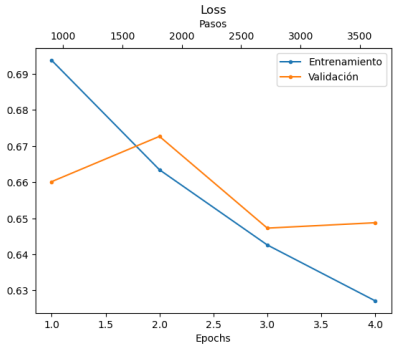
La Tabla 5.1 presenta los hiperparámetros seleccionados luego de múltiples pruebas enfocadas en conseguir el mejor ajuste para todos los *datasets* en el grupo. Estos fueron aplicados en el entrenamiento de los modelos que se usaron en los experimentos A y B.

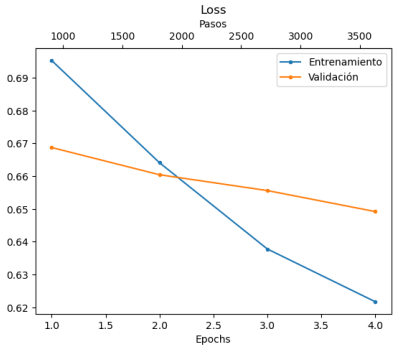
En las Figuras 5.1, se puede observar el valor del *loss* obtenido durante el entrenamiento. Como se aprecia, el ajuste en los modelos entrenados con el grupo de *datasets* grandes y medianos (Figuras 5.1a, 5.1b, 5.1e, 5.1f) es más óptimo que el de los *datasets* pequeños. Las líneas azul y naranja, que muestran el *loss* de entrenamiento y validación respectivamente, decrecen y convergen. Por el contrario, la escasa cantidad de textos en el grupo de *datasets* pequeños ocasionó distintos problemas, los cuales serán explicados en la siguiente subsección.



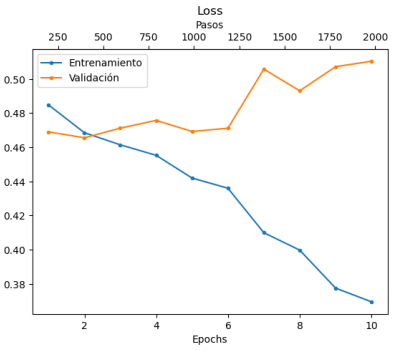


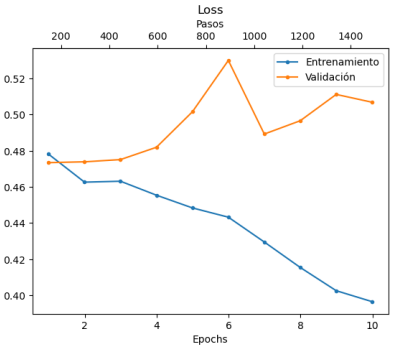
(A) Ptacek. BERT arriba y RoBERTa abajo.



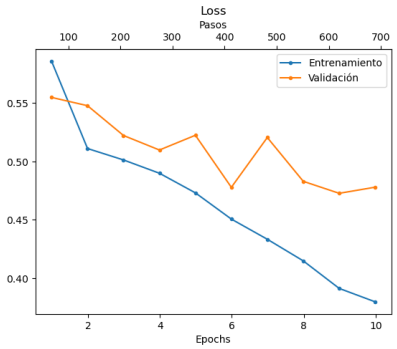


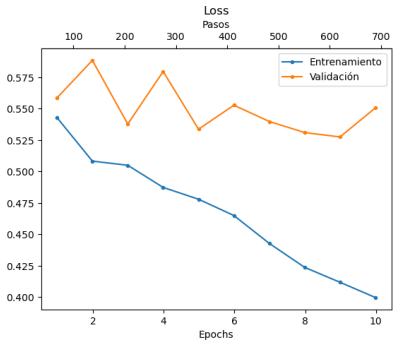
(B) SPIRS. BERT arriba y RoBERTa abajo.



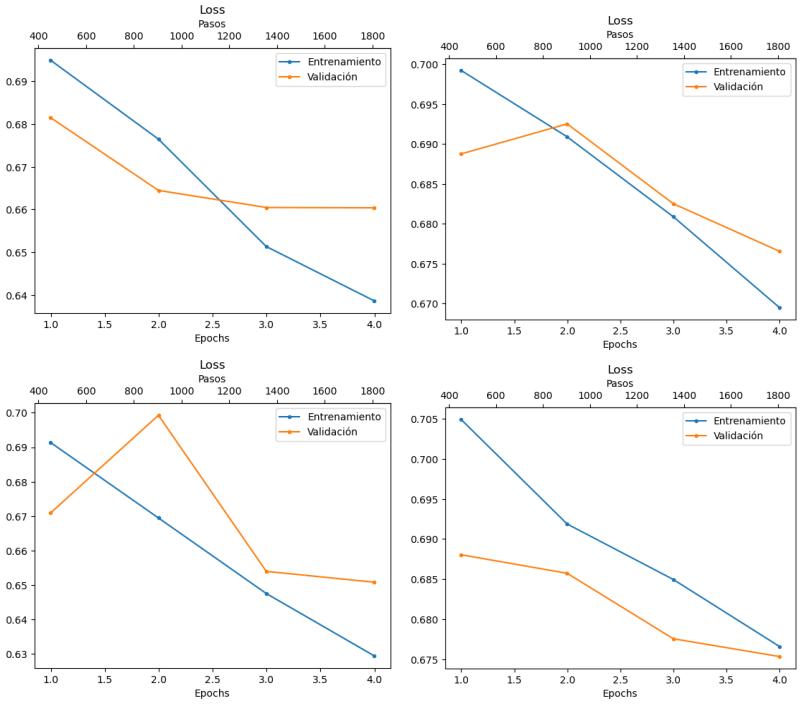


(C) iSarcasm. BERT arriba y RoBERTa abajo.





(D) Riloff. BERT arriba y RoBERTa abajo.



(E) SPIRS Intencional. BERT arriba y RoBERTa abajo.

(F) SPIRS Percibido. BERT arriba y RoBERTa abajo.

FIGURA 5.1: *Loss* del *dataset* de entrenamiento y validación.

Hiperparámetros	***Datasets* grandes**	***Datasets* medianos**	***Datasets* pequeños**
Épocas	4	4	10
Batch size	16	12	12
Dropout	0.1	0.1	0.1
Weight decay	0.01	0.01	0.01
Warmup ratio	0.06	0.06	0.06
Adam β_1 y β_2	0.9, 0.999	0.9, 0.999	0.9, 0.999
Adam ϵ	1e-08	1e-08	1e-08
Learning rate	4e-06	3e-06	4e-06
Learning rate schedule	Lineal con warmup	Lineal con warmup	Lineal con warmup

TABLA 5.1: Hiperparámetros utilizados en el entrenamiento. Los *datasets* se agrupan en grandes (Ptáček, SPIRS), medianos (SPIRS Intencional y Percibido) y pequeños (Riloff y iSarcasm)

5.1.0.1. Problemas encontrados en el grupo de *datasets* pequeños

Pese a que en la selección de hiperparámetros se intentó reducir el *overfitting*, la cantidad limitada de muestras positivas en el grupo de *datasets* pequeños impidió un entrenamiento óptimo.

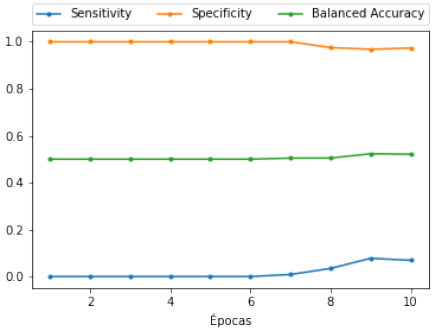
La cantidad de épocas tuvo que ser aumentada a 10 antes de evidenciar un aprendizaje. Sin embargo, como se puede ver en las Figuras 5.1c y 5.1d, la diferencia de *loss* entre los *datasets* de entrenamiento y validación aumenta en cada época, lo cual indica un alto *overfitting*.

A fin de reducir la complejidad del modelo y, por lo tanto, el *overfitting*, se utilizaron sin éxito las técnicas de regularización de *dropout* y *weight decay*, detalladas a continuación.

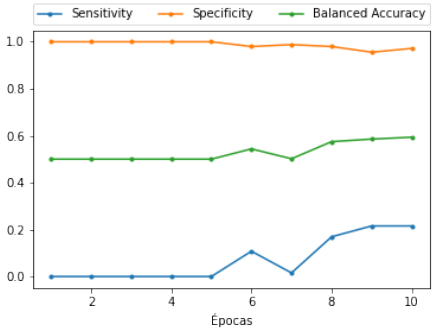
- El *dropout* causa que algunas neuronas se desactiven aleatoriamente durante el entrenamiento (Srivastava *et al.*, 2014). A mayor valor, mayor probabilidad de desactivación. En la presente investigación se incrementó de 0.1 a 0.3 y 0.5, sin embargo el modelo incurrió en *underfitting* y comenzó a predecir la misma clase.
- El *weight decay* es un valor que multiplica la suma del cuadrado de los pesos de una red y se suma al *loss*, logrando penalizar la complejidad del modelo (Krogh y Hertz, 1991). Se utilizó un valor de 0.1 en lugar de 0.01, mas no se evidenció una mejora en el *overfitting* o en las métricas del modelo.

Por otro lado, el gran desbalance de los *datasets* y las escasas muestras positivas ocasionaron métricas deficientes. El *balanced accuracy*, la métrica más importante en *datasets* desbalanceados, tiene un valor de 0.5 en promedio, el peor posible.

En las Figuras 5.2, se desglosa el *balanced accuracy* en *sensitivity* y *specificity*, mostrando que el *sensitivity* es menor. Esta es la medida enfocada en las predicciones verdaderas positivas, la clase con menos muestras.



(A) iSarcasm. BERT arriba y RoBERTa abajo.



(B) Riloff. BERT arriba y RoBERTa abajo.

FIGURA 5.2: Métricas de *balanced accuracy*, *sensitivity* y *specificity* durante el entrenamiento del grupo de *datasets* pequeños

5.2 Resultados del Experimento A

Las Tablas 5.2 y 5.3 muestran los resultados obtenidos para el grupo de *datasets* grandes y pequeños correspondientes al Experimento A. Debido al balance y desbalance de los *datasets*, la métrica mostrada para el primer grupo es el *accuracy* y el *balanced accuracy* para el segundo grupo.

	BERT	RoBERTa
	Accuracy	***Accuracy***
Ptáček (Sup. Dist.)	0.79	0.76
SPIRS (Sup. React.)	0.64	0.62

TABLA 5.2: Resultados de las pruebas en el grupo de *datasets* grandes. Dado que son *datasets* balanceados, se utiliza la métrica de *accuracy*.

	BERT			RoBERTa		
	Bal. Acc.	***Sens.***	***Spec.***	***Bal. Acc.***	***Spec.***	***Sens.***
iSarcasm (Rec. Man.)	0.51	0.07	0.95	0.50	0.02	0.99
Riloff (Anot. Man.)	0.56	0.15	0.98	0.56	0.12	1.00

TABLA 5.3: Resultados de las pruebas en el grupo de *datasets* pequeños. Dado que son *datasets* desbalanceados, se utiliza la métrica de *balanced accuracy*, que combina el *sensitivity* (*sens*) y *specificity* (*spec*).

En los *datasets* grandes el *accuracy* es mayor en Ptáček que en SPIRS. Este resultado se discutirá posteriormente en la subsección 5.3.1.

Mientras tanto, en los *datasets* pequeños, el *balanced accuracy* muestra un desempeño deficiente: el mejor valor fue de tan solo 0.56 en los modelos entrenados con Riloff. Este resultado es esperado, puesto que en el entrenamiento se obtuvieron métricas similares y con *overfitting* debido a la reducida cantidad de textos.

Además, el desbalance de los *datasets* también afectó el desempeño. Al separar el *balanced accuracy* en *specificity* y *sensitivity*, se puede apreciar un gran contraste, donde el primero es mucho mayor. El *specificity* es la métrica que se enfoca en la clase negativa, que representa alrededor del 80 % del *dataset* de iSarcasm y Riloff, como se mostró en la Tabla 4.1.

El modelo entrenado con Riloff muestra un *balanced accuracy* mayor, sin embargo no es posible compararlo con iSarcasm por la diferencia de tamaños de los *dataset*.

En ambos grupos de *datasets*, los modelos mantuvieron las mismas tendencias. Sin embargo, RoBERTa, a pesar de ser una versión mejorada de BERT, tuvo un desempeño menor. Ello podría deberse a que la elección de hiperparámetros durante el entrenamiento se realizó basándose en BERT, y este suele utilizar un *learning rate* mayor a RoBERTa (Devlin *et al.*, 2019, Liu *et al.*, 2019b).

5.3 Resultados del Experimento B

La Tabla 5.4 muestra los resultados de la pruebas realizadas con los modelos entrenados con el *dataset* de Ptáček y los *datasets* alternativos de SPIRS. Además, la tabla contrasta las métricas de las pruebas con los *datasets* generados y resalta la prueba con mayor desempeño.

	BERT		**RoBERTa**	
	Accuracy		***Accuracy***	
	Intencional	**Percibido**	**Intencional**	**Percibido**
Ptáček (Sup. Dist.)	**0.575**	0.504	**0.585**	0.473
SPIRS Int. (Sup. Reac.)	**0.610**	0.529	**0.600**	0.517
SPIRS Per. (Sup. Reac.)	0.530	**0.616**	0.524	**0.608**

TABLA 5.4: Resultados de los modelos entrenados con los *datasets* de Ptáček y alternativos de SPIRS al ser probados con los *datasets* generados por tipo de sarcasmo.

Como puede verse, el modelo de Ptáček y SPIRS intencional muestran un mejor desempeño al probarse con el *dataset* intencional; mientras que el modelo de SPIRS percibido destaca en la prueba con el *dataset* percibido. Al igual que la investigación de Abercrombie y Hovy (2016), estos resultados sugieren que existe una diferencia al entrenar modelos con *datasets* recolectados para un tipo específico de sarcasmo.

Además, en todas las pruebas, ambos modelos de SPIRS obtuvieron un mayor desempeño que el modelo de Ptáček, lo cual podría deberse a la calidad superior de los

datasets recolectados con Supervisión Reactiva. El ruido y sesgo presente en el *dataset* de Ptáček puede haber sido perjudicial para el modelo.

Por otro lado, como se ve en la Tabla 5.5 los modelos de iSarcasm y Riloff mantuvieron un desempeño deficiente. El modelo entrenado con el *dataset* de Riloff, incluso, obtuvo métricas peores a las obtenidas en el experimento A. Además, el modelo de iSarcasm entrenado con BERT fue el único que obtuvo un *balanced accuracy* diferente a 0.50.

No obstante, aunque deficiente y de manera definitivamente no concluyente, el modelo de BERT de iSarcasm muestra un desempeño ligeramente mejor al probarse con el tipo de sarcasmo intencional, con el que fue entrenado.

	BERT		**RoBERTa**	
	Balanced Accuracy		***Balanced Accuracy***	
	Intencional	**Percibido**	**Intencional**	**Percibido**
iSarcasm (Rec. Man.)	0.52	0.49	0.50	0.50
Riloff (Anot. Man.)	0.50	0.50	0.50	0.50

TABLA 5.5: Resultados de los modelos entrenados con los *datasets* pequeños al ser probados con los *datasets* generados por tipo de sarcasmo.

5.3.1 Discusión de resultados

En el experimento A, el desempeño del modelo entrenado con el *dataset* de Ptáček fue mayor que el de SPIRS. El *dataset* del primero fue generado con el método de Supervisión Distante, el cual es el más utilizado en las propuestas de modelos de detección de sarcasmo (Cai *et al.*, 2019, Liu *et al.*, 2019a, Pelser y Murrell, 2019, Potamias *et al.*, 2020, Wu *et al.*, 2018).

Sin embargo, las desventajas de la Supervisión Distante, como el ruido, datos erróneos y el sesgo, lo convierten en un método no confiable. El método de Supervisión Reactiva, por el contrario, refleja condiciones más reales, por lo que los resultados podrían sugerir que el desempeño de los modelos reportado en las investigaciones debería ser menor en la práctica.

Esta idea se refuerza en el experimento B, donde el modelo entrenado con Ptáček se prueba con los *datasets* generados y su *accuracy* disminuye en ambos modelos para ambos tipos de sarcasmo.

5.3.2 Resumen del capítulo

Los resultados presentados en el capítulo sugieren que los tipos de sarcasmo tienen un impacto significativo en el desempeño de los modelos de detección de sarcasmo y reafirman los estudios realizados por Abercrombie y Hovy (2016) y Oprea y Magdy (2019a). Estos muestran que al tener en cuenta el tipo de sarcasmo para el que será utilizado el modelo, se puede mejorar su desempeño. Por ejemplo, si el modelo se empleará para detectar sarcasmo intencional, se debe entrenar con el mismo tipo de sarcasmo.

Además, muestran que las tendencias no son específicas a un modelo, pues se mantienen en BERT y RoBERTa. Ambos modelos se encuentran en el estado del arte del área.

En la experimentación también se encontraron obstáculos ocasionados por las desventajas de los mismos métodos de recolección. El tamaño y desbalanceo de los *datasets* pequeños impidieron llevar a cabo un entrenamiento fructífero y obtener conclusiones definitivas.

En el capítulo siguiente, se realizará una evaluación final de la presente investigación.

Capítulo 6

Conclusiones y Trabajos Futuros

6.1 Conclusiones

La presente investigación realiza un análisis cuantitativo de los métodos de recolección y su impacto en el desempeño de los modelos de detección de sarcasmo. Para realizar un contraste y evaluación de en qué medida han sido afectados estos modelos, se trazan tres objetivos específicos.

El primero consiste en contrastar los métodos de recolección entre ellos, entrenando un modelo con *datasets* de cada uno; y el segundo en analizar la importancia de los tipos de sarcasmo obtenidos por los métodos. Ambos objetivos fueron abordados por los experimentos A y B, respectivamente. En el experimento A, se logró contrastar el desempeño de los modelos entrenados con un *dataset* de cada método y en el B, se logró validar cada modelo con un *dataset* por cada tipo de sarcasmo, a excepción de los modelos de métodos manuales, por tener *datasets* muy pequeños y desbalanceados.

El procedimiento utilizado por la Supervisión Reactiva permite obtener *datasets* más cercanos a la realidad que la Supervisión Distante. Teniendo esto en cuenta y a partir de los resultados en el Experimento A, que muestran un menor desempeño en el *dataset* de Supervisión Reactiva (diferencia de *accuracy* de 0.15 puntos en BERT y 0.14 en RoBERTa), es posible, aunque no definitivo, que el desempeño de modelos en el estado del arte que utilizan la Supervisión Distante sea menor en la práctica que el reportado en sus estudios.

Por otro lado, en el Experimento B se encontró que el modelo entrenado con el *dataset* de SPIRS Intencional, de Supervisión Reactiva, tuvo un mejor desempeño detectando sarcasmo intencional que el modelo que utilizó Supervisión Distante (diferencia de *accuracy* de 0.035 puntos en BERT y 0.015 en RoBERTa), que aunque se compone del mismo tipo de sarcasmo, tiene datos sesgados y ruidosos.

El tercer objetivo busca asegurarse que los comportamientos encontrados no sean específicos a un modelo. Para alcanzarlo, todas las pruebas en los experimentos se realizaron con dos modelos, el de BERT y el de RoBERTa.

La limitación más importante fue encontrada durante el entrenamiento, donde se hicieron evidentes las desventajas de los métodos de Anotación Manual y Recolección Manual. La dependencia de ambos métodos en el componente manual elevan significativamente el costo e impide generar *datasets* extensos. La cantidad reducida de textos ocasiona un *overfitting* y un *balanced accuracy* promedio de 0.5.

Pese a esta limitación, los resultados obtenidos logran indicar que el tipo de sarcasmo obtenido con un método de recolección es un factor importante para el desempeño del modelo: un modelo entrenado con un *dataset* que contenga un tipo específico será mejor para predecir sarcasmo del mismo tipo. El *accuracy* promedio en Ptáček es mejor por 0.0915 puntos detectando sarcasmo intencional, SPIRS Intencional por 0.082 y en SPIRS Percibida es mejor por 0.085 detectando sarcasmo percibido. Así, si se toma en cuenta el tipo de sarcasmo que tendrá que ser predecido por el modelo, se podría mejorar el desempeño sin necesidad de aplicar hiperparámetros diferentes o implementar nuevas arquitecturas.

Además, se comprueban los estudios previos de Abercrombie y Hovy (2016) y Oprea y Magdy (2020), donde se encontró que los tipos de sarcasmo deberían ser tratados como fenómenos diferentes al momento de intentar detectar sarcasmo.

Por otra parte, se advierte sobre el tamaño reducido de los *datasets* generados con los métodos manuales, puesto que entrenar modelos con ellos conllevan a un alto *overfitting* y un desempeño bajo.

Para finalizar, se espera que esta investigación resalte la importancia de los datos utilizados para entrenar modelos de detección de sarcasmo, ya que esta usualmente es opacada por el entrenamiento o arquitectura del modelo. Asimismo, el estudio busca aportar un mayor conocimiento en el área poco desarrollada de los métodos de recolección de textos sarcásticos. Con este mismo propósito, se proporcionan posibles trabajos futuros a continuación.

6.2 Trabajos futuros

Trabajos futuros que continúen con la línea de investigación podrían utilizar modelos y otros *datasets* generados con los métodos de recolección de Supervisión Distante y Reactiva. De esta manera, se verificaría si las tendencias de desempeño de los modelos en este estudio se mantienen.

Respecto a los métodos de Anotación y Recolección Manual, ya que los textos son evaluados o creados por un humano, se podrí

a comparar la calidad de sus *datasets* con los generados por la Supervisión Reactiva. Sin embargo, ello requerirá *datasets* con un tamaño similar a el de SPIRS.

Por último, observando los resultados de Abercrombie y Hovy (2016) se aprecia que las diferencias entre tipos de sarcasmo se acentúan mientras datos como información del autor y de la audiencia son incorporados en el entrenamiento del modelo. Por ello, un aspecto interesante que también debería ser investigado es proporcionar datos contextuales a los modelos y verificar si las diferencias entre métodos incrementan.

REFERENCIAS BIBLIOGRÁFICAS

Abercrombie, G. and Hovy, D. (2016). Putting Sarcasm Detection into Context: The Effects of Class Imbalance and Manual Labelling on Supervised Machine Classification of Twitter Conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113, Berlin, Germany. Association for Computational Linguistics.

Bagate, R. and Ramadass, S. (2020). Different Approaches in Sarcasm Detection: A Survey. pages 425–433.

Cai, Y., Cai, H., and Wan, X. (2019). Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Cambria, E. and White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, USA. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Reyes, A., and Barnden, J. (2015). SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter.

Ghosh, A. and Veale, T. (2017). Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, Copenhagen, Denmark. Association for Computational Linguistics.

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., and Mihalcea, R. (2018). CASCADE: Contextual Sarcasm Detection in Online Discussion Forums.

Joshi, A., Bhattacharyya, P., Carman, M., Saraswati, J., and Shukla, R. (2016a). How Do Cultural Differences Impact the Quality of Sarcasm Annotation?: A Case Study of Indian Annotators and American Text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99, Berlin, Germany. Association for Computational Linguistics.

Joshi, A., Bhattacharyya, P., and Carman, M. J. (2016b). Automatic Sarcasm Detection: A Survey.

Khodak, M., Saunshi, N., and Vodrahalli, K. (2018). A Large Self-Annotated Corpus for Sarcasm. *arXiv:1704.05579 [cs]*.

Kreuz, R. and Caucci, G. (2007). Lexical Influences on the Perception of Sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York. Association for Computational Linguistics.

Krogh, A. and Hertz, J. (1991). A Simple Weight Decay Can Improve Generalization. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann.

Liu, L., Priestley, J., Zhou, Y., Ray, H., and Han, M. (2019a). A2Text-Net: A Novel Deep Neural Network for Sarcasm Detection.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Lukin, S. and Walker, M. (2017). Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2021). On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. Technical Report arXiv:2006.04884, arXiv.

Oprea, S. and Magdy, W. (2019a). Exploring Author Context for Detecting Intended vs Perceived Sarcasm. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859.

Oprea, S. and Magdy, W. (2019b). iSarcasm: A Dataset of Intended Sarcasm.

Oprea, S. V. and Magdy, W. (2020). The Effect of Sociocultural Variables on Sarcasm Communication Online. *arXiv:2004.04945 [cs]*.

Pelser, D. and Murrell, H. (2019). Deep and Dense Sarcasm Detection. *arXiv:1911.07474 [cs]*.

Plepi, J. and Flek, L. (2021). Perceived and Intended Sarcasm Detection with Graph Attention Networks. *arXiv:2110.04001 [cs]*.

Poria, S., Cambria, E., Hazarika, D., and Vij, P. (2017). A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. *arXiv:1610.08815 [cs]*.

Poria, S., Hazarika, D., Majumder, N., and Mihalcea, R. (2020). Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *arXiv:2005.00357 [cs]*.

Potamias, R. A., Siolas, G., and Stafylopatis, A.-G. (2020). A Transformer-based approach to Irony and Sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Ptáček, T., Habernal, I., and Hong, J. (2014). Sarcasm Detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Shmueli, B., Ku, L.-W., and Ray, S. (2020). Reactive Supervision: A New Method for Collecting Sarcasm Data. *arXiv:2009.13080 [cs]*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. page 30.

Van Hee, C., Lefever, E., and Hoste, V. (2018). SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*.

Wallace, B. C., Choe, D. K., and Charniak, E. (2015). Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., and Huang, Y. (2018). THU_ngn at SemEval-2018 Task 3: Tweet Irony Detection with Densely connected LSTM and Multi-task Learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56, New Orleans, Louisiana. Association for Computational Linguistics.