Departamento de Sociología
Facultad de Ciencias Sociales
Universidad de Chile

MEMORIA PARA OPTAR AL TÍTULO PROFESIONAL DE SOCIÓLOGO

DISCUSIÓN ONLINE SOBRE LA SEQUÍA EN CHILE

BÚSQUEDA DE ENTIDADES PROTAGONISTAS DE LA DISCUSIÓN A TRAVÉS DE MINERÍA DE DATOS EN TWITTER

Autor: Joaquin Quiroz Gajardo
Profesor Guía: Fernando Campos
12 de mayo, 2021

Agradecimientos

En primer lugar, agradezco a la Facultad de Ciencias Sociales y a todas las personas que la integran, por haberme formado y educado todos estos años. Agradezco al profesor Fernando Campos y Victoria Cruzat por ayudarme a construir mi tesis, por los consejos, por orientarme y por estar siempre disponibles para brindarme su ayuda. Al mismo tiempo, agradezco a mi familia y colegas por su apoyo durante este proceso.

Índice

1. Resumen

Este trabajo proporciona un análisis basado en Minería de Datos, sobre textos pertenecientes a la discusión respecto de la sequía en Chile, a través de los comentarios expuestos en la red social Twitter, recopilados entre abril y diciembre de 2020. Los modelos producidos para este análisis reconocen a las entidades nombradas en la temática de la sequía y categorizan las opiniones que los usuarios dan a estas entidades. La investigación propone comprender la discusión en torno a la sequía como una exploración metodológica, respecto de las técnicas de Procesamiento de Lenguaje Natural, la Extracción de Características y la Minería de Datos desde las Ciencias Sociales. De este modo, los resultados aumentarán la comprensión sociológica sobre la opinión pública respecto de la mega sequía que atraviesa Chile, mientras ofrece un método efectivo para trabajar sobre grandes volúmenes de información presentes en redes sociales. Se realizó una clasificación de tipo TASBA a través de una versión en español del modelo BERT, obteniendo un accuracy del 93,65%. La investigación concluye que la metodología propuesta tiene un desempeño suficiente para justificar su implementación en campo del debate de la sequía en redes sociales.

2. Palabras Claves

Twitter, natural language processing, targeted aspect base opinion mining, transferring learning, transformers, sentence base opinion mining, BERT.

3. Introducción

Las Ciencias Sociales enfrentan nuevas oportunidades para avanzar en la comprensión de los problemas sociales. Con los avances recientes en el campo del Procesamiento de Lenguaje Natural se incorporan nuevas metodologías que permiten procesar datos de texto automáticamente y aumentan el rango de datos cubiertos por los estudios (Deisenroth, Faisal y Ong, 2020).

Las tareas típicas de Procesamiento de Lenguaje Natural utilizadas en las Ciencias Sociales son el análisis de sentimientos, el análisis de reputación y la identificación de tópicos en redes sociales (Singhal et al. 2001). Mientras tanto, las técnicas de recuperación de características, dedicadas a la búsqueda de metadatos en documentos digitales (Singhal et al., 2001), no se usan comúnmente en las Ciencias Sociales (Ribeiro, Batista y Língua Falada, 2013) y es en esta área en la que se adentrará esta investigación. Así, esta tesis trata una de esas tareas correspondiente a la Minería de Datos a nivel de frase, esto es, el proceso de detectar aspectos y opiniones de entidades[1] en una frase, siendo posible detectar diferentes aspectos de una entidad (sus atributos), y la opinión existente sobre este aspecto.

Durante los últimos años se ha producido una mejora significativa en el campo de Procesamiento de Lenguaje Natural, debido al desarrollo de arquitecturas que facilitan el proceso de transferencia de aprendizaje, es decir, utilizando modelos pre-entrenados. De esta manera, los modelos de lenguaje natural aprenden nuevas tareas más rápido y con menos ejemplos de entrenamiento (Brown et al., 2020).

La desventaja que la transferencia de aprendizaje tenía en los modelos de lenguaje natural es que solo operaba en una pequeña parte del modelo, su primera capa, y el entrenamiento del resto del modelo comienza desde cero, lo cual era ineficiente. Sin embargo, en los últimos años, los modelos de lenguaje han logrado pre-entrenar más partes del modelo a través de la creación de mecanismos de atención y a las diferentes variantes de este enfoque. Esta aproximación permitió la creación de modelos completamente pre-entrenados. Es decir, crear una representación del lenguaje de alto nivel, permitiendo que

[1] Estas entidades pueden ser una variedad de cosas, desde personas hasta términos biológicos específicos. Pueden ser abstractos o existir en la realidad, pero requieren un nombre distintivo (Nadeau and Sekine 2007).

los modelos pre-entrenados logren resultados al nivel del estado de arte en muchos problemas de Procesamiento de Lenguaje Natural (Ruder, 2018).

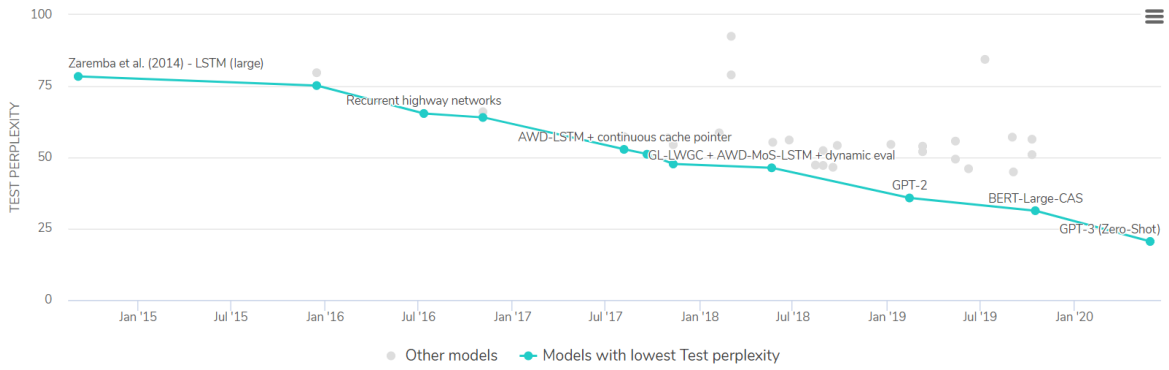

Fig. 1. Benchmarks de los Language Model (PapersWithCode 2020).

La Figura 1 muestra la mejora de los language models. Al implementar estos modelos con bases de datos de tamaño moderado, se produce un desempeño similar.

De esta manera, esta investigación busca detectar y clasificar automáticamente las entidades nombradas dentro de una muestra de tuits que abordan el tema de la sequía en Chile. Esto, a través de un proceso conocido como Aspect Extraction. Lo que hace posible relevar el apoyo, la neutralidad o bien, el rechazo expresado en una opinión.

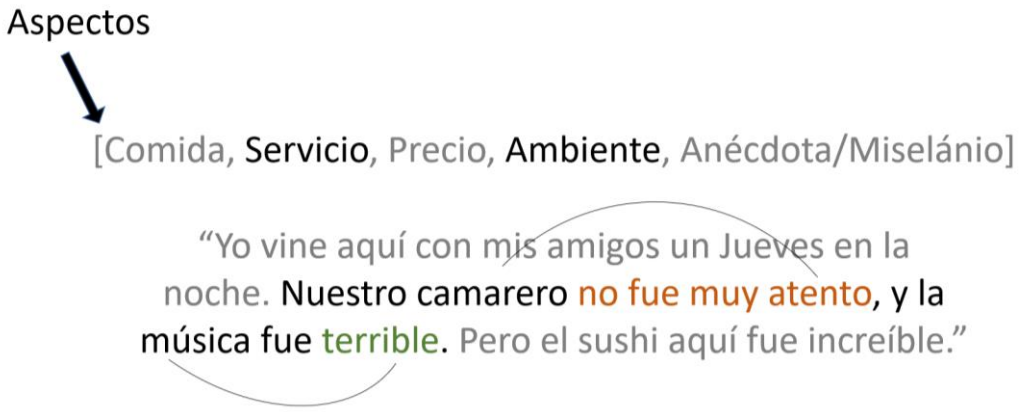

Fig. 2. Ejemplo de Aspect-Based Opinion Mining inspirado en Min (2018).

Entre los beneficios de la extracción de datos de Internet se encuentran la menor inversión de mano de obra humana, por tanto, la reducción de costos, la simplificación de problemas logísticos y minimización del tiempo requerido para la producción de información.

En este contexto, el valor de esta tesis no solo se encuentra en los resultados obtenidos en la comprensión de este problema socioambiental particular, sino que presenta una metodología específica de recolección y procesamiento de datos automáticos, complementada con una interpretación específica y contextualizada del sentido. De esta manera, la complementariedad de ambas técnicas que se llevará a cabo durante esta investigación puede ser de gran utilidad para futuras investigaciones. En este sentido, el desarrollo de la tesis presenta explicaciones detalladas de los conceptos y procedimientos llevados a cabo, con el objetivo de que el documento final sirva como un recurso útil para otros investigadores de las Ciencias Sociales que aborden problemas de procesamiento del lenguaje natural.

En cuanto a la sequía, es un tema preocupante ya que el déficit de lluvias se extiende de 2010 a 2019, generando una escasez de agua en Chile continental sin precedentes históricos. La prolongación del período de sequía ha tenido muchas consecuencias que afectan a los habitantes de los territorios. Ejemplos de ello son las poblaciones de bajos ingresos sin agua para uso sanitario, la limitación de la producción agrícola, la eliminación de puestos de trabajo, el aumento de la conflictividad en las localidades con industrias que realizan el uso intensivo del agua y la desconfianza de la población hacia la clase política, debido a los vínculos que existen entre sectores empresariales y miembros de la clase política. Debido a esto, se decide centrar esta investigación en determinar la opinión expresada a las entidades relevantes sobre el tema de la sequía.

4. Antecedentes

En este apartado se presentan los aspectos necesarios para contextualizar el estudio. Los antecedentes del conflicto por la sequía en Chile sirven para tener el contexto que permitirá la interpretación de los comentarios realizados en los tuits y seleccionar las categorías adecuadas. También será examinada la cultura de Twitter y las características en las que opera la plataforma, pues sus particularidades moldearon la discusión y el cómo se tratan temas políticos.

4.1 La disputa por el agua en Chile

En este apartado se hablará de lo que es la mega-sequía en Chile, revisando las causas y consecuencias del fenómeno, junto con las políticas de gestión del agua del país, presentando las medidas y propuestas para enfrentar este problema tanto en el presente como en el futuro.

La mega-sequía de Chile es el período de sequía que experimentó Chile desde la región de Coquimbo hasta la región de Los Ríos entre 2010 y 2019, en donde hubo un déficit de lluvia de alrededor del 30% anual (Montes 2020). El carácter excepcional de la sequía corresponde a su extensión, ya que las sequías plurianuales existentes desde que existen registros no habían pasado previamente los tres años (CR 2015). En este caso, la mega-sequía fue un fenómeno causado por una combinación de factores antropogénicos y naturales.

Entre los factores más relevantes que provocaron la sequía en el territorio continental chileno, según (CR 2015), se encuentran:

1) Los cambios antropogénicos, definidos como los cambios provocados por el debilitamiento de la capa de ozono y los gases de efecto invernadero.
2) El movimiento decenal del Pacífico, que son cambios cíclicos en la temperatura del Pacífico que permiten que los sistemas frontales lleguen al continente. En períodos fríos, los sistemas frontales llegan en menor cuantía al continente, favoreciendo por tanto la sequía.

3) Finalmente, los fenómenos del niño y la niña representan cambios de temperatura en el Pacifico tropical. La niña representa el período de bajas temperaturas, lo que favorece que estos sistemas frontales no lleguen al continente, provocando sequía. En contraste, el niño tiene el efecto contrario, dado que representa temperaturas más altas que permiten el superávit de lluvias en el territorio continental chileno.

Los cambios antropogénicos y el ciclo decenal del Pacífico son los responsables de la mega-sequía que vivió Chile, teniendo el niño y a la niña un efecto neutral dado que lo prolongado de la sequía provocó que ambos efectos se neutralizaran (Vega 2021). Si bien los cambios antropogénicos solo tienen el 25% de la responsabilidad de la mega-sequía (mientras que el movimiento decenal del país tiene el 50% de la culpa), es el factor que desequilibra el ciclo natural de las lluvias en el territorio nacional, agravando el déficit de precipitaciones, permitiendo que pase de 15% a 30% (CR 2015).

Las mega-sequías son un fenómeno extraordinario en el territorio nacional, ya que ocurren naturalmente cada 300 años, mientras que, debido a cambios antropogénicos, estos episodios ocurrirán cada 20 años en el país, según se puede proyectar (CR 2015). Como era de esperar, la flora y fauna del continente y de la zona costera no están adaptadas para enfrentar sequías tan prolongadas, generando una disminución del 50% en los sedimentos entregados a la orilla del mar y un aumento del área afectada por incendios forestales en un 70% (CR 2015).

Esta mega-sequía representa y representará un gran desafío para Chile, una muestra significativa de ello es que 400 mil habitantes se encuentran sin acceso a agua para uso doméstico, los cuales se caracterizan por ser habitantes de bajos ingresos (Carvajal et al. 2013; Valencia, 2019). Varias de las industrias más esenciales del país dependen directamente del agua, como las industrias minera, hidroeléctrica y forestal. Las empresas mineras dentro de Chile se encuentran generalmente en el Gran Norte, por lo que, aunque no se ven afectadas directamente por la mega-sequía. Este sector se ve envuelto frecuentemente en conflictos con diferentes comunidades por el acceso al agua debido a lo escaso del recurso en esta zona del país (Larraín 2006).

Entre las medidas adoptadas por el país se encuentran (Arellano 2017) la entrega de agua a través de camiones cisterna a los habitantes sin acceso para uso doméstico, la entrega de fondos para el sector forestal para combatir el período de sequía, subsidios para la gran agricultura y la agricultura campesina, y campañas de CONAF para combatir los incendios forestales. Considerando las medidas que se llevan a cabo a nivel público, privado y ciudadano (CR 2015), las estrategias para combatir este fenómeno tienden a ser principalmente en favor de la industria forestal, dejando en segundo lugar el acceso a agua de uso sanitario, y en tercer lugar otras medidas para mejorar la capacidad de resiliencia del país.

Uno de los factores más importantes para entender el impacto de la sequía en la sociedad chilena es el régimen de propiedad del agua establecido el año 1981. Esta reforma provocó que el agua sea administrada íntegramente por el mercado, a través de cesiones sin costo o mediante subasta pública (Rivera et al.2016). Los argumentos a favor de este régimen de propiedad del agua son que el mercado distribuirá eficientemente los recursos hídricos del país. Los críticos de este sistema afirman que el mercado genera una posesión especulativa y permite su uso en cultivos que, por su escala y tipo de cultivo, atentan contra la capacidad de resiliencia de los territorios y dejan a poblaciones de escasos recursos sin agua para uso doméstico (Herrera et al.2019).

Otro problema de este régimen de derechos de agua es el acaparamiento. El 1% de los propietarios posee el 79% de los derechos de agua en términos de volumen. En cuanto a la propiedad, el 71% de estos derechos son para riego, el 8,3% para uso sanitario y el 0,6% para uso minero (Correa-Parra, Vergara-Perucich, and Aguirre-Nuñez 2020). La gran cantidad de derechos dedicados al riego, que serían suficientes para abastecer de agua a 243 millones de familias, han llevado que se generen propuestas para gestión del agua en este sector, como promover el cultivo de especies con bajo consumo de agua, y favorecer la agricultura familiar campesina, dado que este tipo de producción tiende a hacer un menor uso del recurso (Correa-Parra, Vergara-Perucich, and Aguirre-Nuñez 2020).

El volumen de transacciones muestra una tendencia a mantener los derechos de agua, solo el 1.54% de las escrituras han sido transferidas en el mercado. Mientras tanto, los

propietarios han renunciado a sólo 82 de los 133,459 derechos emitidos (DGA 2020a), debido a los costos mínimos de mantenimiento.

El aumento de la regulación que incorporó la reforma del 2005 al código de aguas prometió una menor demanda y posesión especulativa de los derechos de agua mediante la formación de subastas de derechos y comisiones que sancionen el mantenimiento ocioso del agua. Sin embargo, el aumento de la sequía que experimenta Chile generó el efecto contrario. Las adjudicaciones de derechos aumentaron sin que hubiera un cambio respecto de la tendencia anterior (Herrera et al.2019).

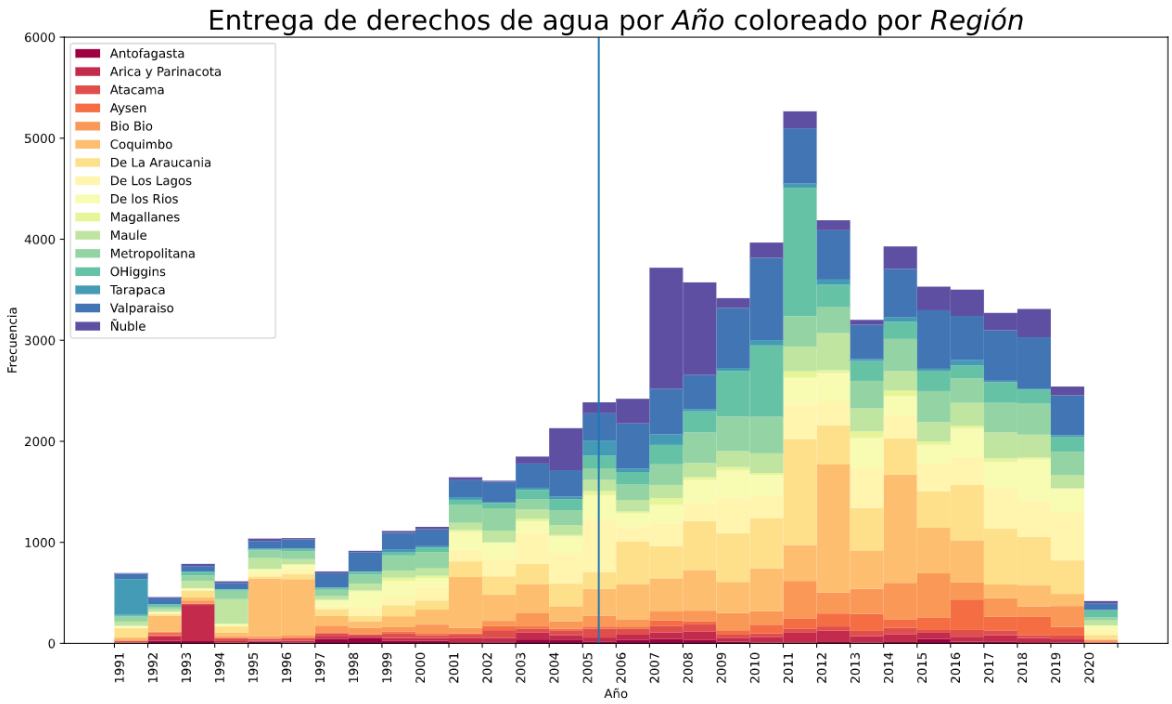

Fig. 3. Elaboración propia a partir de datos disponibles en (DGA 2020a).

Los problemas hídricos que vive el país no tienen precedentes históricos, debido a la magnitud y duración de las sequías que se avecinan. Por tanto, la gestión del agua se convertirá en un tema prioritario en el debate público a lo largo del tiempo. También se prevé que el país adoptará estándares estrictos en el tiempo para aprobar megaproyectos con uso intensivo de agua, avanzando en la evaluación del costo-beneficio de cada proyecto.

4.2 El papel de Twitter en los movimientos sociales.

Twitter es una red social de microblogging que se caracteriza por publicar mensajes de pequeño tamaño. Twitter tiene una larga historia de apertura hacia aplicaciones externas, lo que explica su crecimiento y la gran cantidad de investigaciones académicas alrededor de esta red social.

Una innovación esencial implementada en la red fue el hashtag, que permite mantener la discusión en la red de una manera eficiente, lo que ha hecho que sea adoptada en muchas otras redes sociales. En términos simples, el hashtag es el concepto que consiste en etiquetar grupos o temas de redes sociales para que sea seguido rápidamente, permitiendo a las personas participar en una conversación global (Black 2018). Esta innovación permite la propagación de ideas políticas en el llamado "hashtag activism", algunos ejemplos son: #MeToo, #OWS (Occupy Wall Street) y #BlackLivesMatter (Black 2020), y ejemplos nacionales son #PatagoniaSinRepresas, #nomasafp, #chiledesperto, #NoEstamosEnGuerra, #EvacionMasiva y #NoEsSequíaEsSaqueo.

Fig. 4. Manifestación en medio de una subasta de derechos de agua no consuntivos en Ñuble, enero 2020, extraído en González (2020).

Además, la incorporación del retuit facilitó la comunicación de ideas dentro de la red. Este fue inicialmente inventado por los usuarios que hacen una copia del comentario de otros usuarios, citando el mensaje al poner el nombre de usuario original. Por otro lado, el trending topic permite a los usuarios debatir temas contingentes, convirtiéndose en un

campo de disputa política. Las cuentas verificadas consolidan la incorporación de figuras políticas dentro de la red con cuentas emblemáticas como las de Hugo Chávez y Donald Trump. Si bien en Chile no hay usuarios políticos tan característicos de la red social, las cuentas de Camila Vallejo y Sebastián Piñera destacan por su número de seguidores.

Existe una gran cantidad de conocimiento sobre cómo funciona la comunicación dentro de las redes de microblogging como Twitter. Una de las nociones centrales es la gran homofilia típica de las redes sociales, ya que las personas tienden a comunicarse y consumir contenido de personas e instituciones que piensan de manera similar a ellas (Yardi y Boyd 2010). Además, en las discusiones marcadas por un evento de alto impacto, hay evidencia de que las emociones de los tuits tienden a polarizarse (Yardi y Boyd 2010), o, en otras palabras, este efecto de homofilia aumenta. Por lo mismo, los mensajes emitidos con una fuerte carga emocional tienden a ser más virales entre los usuarios que piensan de manera similar, en contraste con los argumentos del tipo racional que reciben menos difusión (Brady et al.2017). Así podemos señalar que tanto las emociones como la opinión expresada en Twitter tienden a ser más polarizada y exagerada de lo que ocurre fuera de la red.

La relevancia de los datos de Twitter para las Ciencias Sociales radica en la tendencia cultural actual de tratar problemas políticos dentro de esta red, lo que la hace adecuada para tratar el problema de la sequía. La gente "retuitea" para difundir el tuit a una nueva audiencia, entretener, comentar o agregar información al tuit retuiteado para mostrar su apoyo a la opinión publicada, entre otras razones (Huang, Thornton y Efthimiadis 2010). Además, esta red social juega un papel fundamental en los movimientos sociales contemporáneos, contribuyendo a difundir la protesta política en muchas ciudades al convertirse en un vehículo de comunicación de eventos políticos planificados, además de ser un medio de comunicación de noticias e información ignorada por los medios tradicionales (Jost et al.2018).

La cada vez mayor disposición de información está permitiendo que grupos organizados intervengan en resultados electorales a través de prácticas como lo son la propagación de fake news en las redes sociales, con un medido cálculo de cuáles son las informaciones y

grupos demográficos más susceptibles de ser creídas y que tendrán un mayor impacto en el resultado electoral (Aral and Eckles 2019).

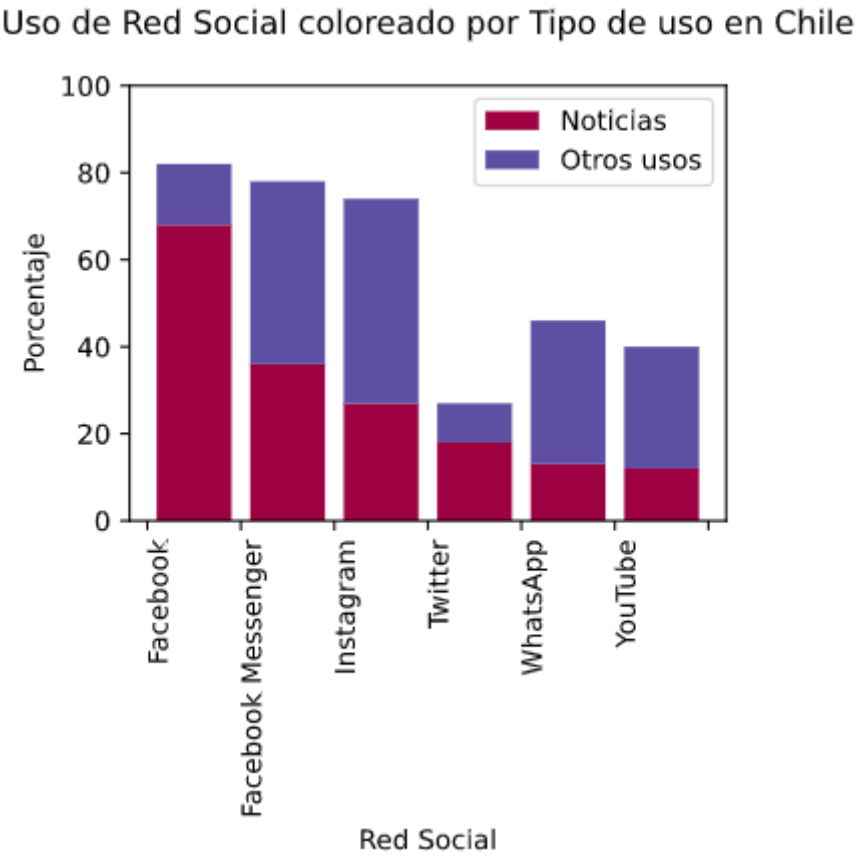

Fig. 5. Elaboración propia a partir de los datos disponibles en Nic (2018).

Como se puede ver en la figura 5, Twitter está orientado a transmitir noticias. Alrededor de 2/3 de los usuarios lo usan para consumir noticias (Nic et al.2018), en comparación con la proporción de usuarios de otras redes sociales que consumen noticias en sus redes.

Con respecto a la discusión de los problemas de sequía dentro de la plataforma, la evidencia muestra que cuando ocurren eventos de sequía, hay un aumento significativo en la discusión que trata este tema (Wagler y Cannon 2015). La red es sensible a las noticias sobre la sequía y, por lo tanto, apropiada para esta investigación. Esta sensibilidad en la red se suma a la conocida relación entre la actividad realizada dentro de las redes sociales y las

acciones políticas fuera de la red tanto a nivel internacional (Yamamoto, Kushin y Dalisay 2013) como en Chile (Scherman, Arriagada y Valenzuela 2014).

Por otra parte, a pesar de haberse estudiado en profundidad cómo las redes sociales promueven la propagación de las protestas sociales con una planificación de bajo costo y una adopción significativa por parte de los movimientos sociales de estas tecnologías de la comunicación, aún se comienza a comprender cómo esto altera el comportamiento organizacional y las consecuencias de estos cambios. Algunos cambios que los grupos activistas experimentan debido a las redes sociales son (Hodges and Stocking 2016): las múltiples interacciones, que les permiten mezclar el repertorio de acción, interactuando con diferentes grupos con enfoques personalizados; las redes sociales tienen la capacidad de facilitar la formación de coaliciones debido a la interacción individualizada; y los usuarios de las redes sociales pueden difundir y promover rápidamente conversaciones y contenido, que funcionan como refuerzo ideológico respondiendo a las personalidades e ideales políticos de los usuarios de las redes sociales.

Finalmente, debemos considerar que las plataformas de redes sociales tienden a desviar las discusiones hacia la confrontación y las emociones negativas (De-Wit, Brick, and Van Der Linden 2019). Debido a que este tipo de contenido más desafiante genera una reacción emocional en las personas más rápido, tendiendo a ser más compartido. Las redes sociales presentan contenidos que alteran la percepción de los usuarios, polarizándolos (De-Wit, Brick, and Van Der Linden 2019), por tanto, al analizar los tuits, hay que prestar atención a la posible existencia de un exceso de contenido crítico sobre contenido propositivo que postule soluciones al problema.

5. Problematización

Esta investigación busca explorar en los avances recientes que ha vivido el campo de Procesamiento de Lenguaje Natural mediante la implementación de un estudio de caso, respecto a la comunicación social relativa a la sequía en Chile, mostrando las implicancias de utilizar estos modelos en el campo de la Sociología. Adicionalmente, la investigación ofrece una perspectiva novedosa para generar antecedentes dentro del campo de los estudios ambientales, contribuyendo a la comprensión de la opinión pública que tienen las entidades. Las entidades pueden ser cualquier objeto, persona, lugar, fecha, concepto, etcétera que se pueda nombrar (Nadeau and Sekine 2007), por lo tanto, requiere un nombre único. Para esta investigación, solo los lugares, las instituciones públicas, privadas y las localidades serán relevantes para detectar.

Este estudio permitirá probar las técnicas de procesamiento de lenguaje natural con los datos generados para esta investigación, la cual posee objetivos más allá del testeo de los modelos. Esta alternativa se considera mejor que utilizar datos ya creados para testear los modelos, pues obligara a que se generen los problemas y soluciones que tienen las investigaciones sociológicas reales, permitiendo que la implementación de estas técnicas por otras investigaciones posea la facilidad de una experiencia previa realista.

Esta investigación busca implementar una técnica de inteligencia artificial supervisada que se asemeje a los desafíos de la investigación en Sociología. Se entiende a partir de esta intención la no utilización de una base de datos ya creada para implementar los modelos, ya que busca adentrarse en los desafíos que conlleva la construcción de la base de datos. La fase de construcción y limpieza de bases de datos es la más exigente en términos de tiempo. Por lo tanto, las estrategias de recopilación y clasificación utilizadas en el estudio pueden ser útiles en investigaciones futuras.

Pregunta de investigación.

Dentro de la red social de microblogging Twitter y dentro del periodo de abril a diciembre del 2020, esta investigación busca responder: ¿Cuál es la opinión respecto a las entidades nombradas en la discusión actual sobre la sequía en Chile en términos de aprobación, neutralidad o rechazo?

5.1 Objetivo General

- Desarrollar un modelo que permita medir el apoyo o el rechazo que tienen las entidades involucradas en la discusión sobre la sequía que se dan en Twitter desde abril a diciembre de 2020 en Chile.

5.2 Objetivos específicos

- Extraer de la red social Twitter una parte significativa de tuits que abordan la sequía en Chile desde abril a diciembre de 2020.
- Identificar y clasificar las entidades nombradas dentro de los tuits.
- Catalogar adecuadamente la opinión expresada sobre las entidades nombradas dentro de los tuits en términos de apoyo, neutralidad o rechazo.
- Comprender y mejorar la elección de categorías a través del análisis cualitativo, ampliando el contexto a través del cual interpretar opiniones.

6. Hipótesis

Esta investigación plantea como hipótesis que existe una imagen crítica de las empresas que pertenecen al área de producción y la institucionalidad pública que inspeccionan y otorgan el uso de los derechos de agua, junto con opiniones de apoyo a las comunidades que enfrentan eventos de sequía, tanto pequeños productores como personas con falta de agua para uso doméstico.

Dado que la gente estaba acostumbrada a un estado de sequía en relación con el período anterior al COVID-19, se espera que haya opiniones y comunicaciones positivas que resalten la mayor presencia de agua relativa al período anterior al COVID-19. En este sentido, se espera una presencia significativa de comunicaciones con mensajes positivos relativos a localidades que sufren de sequía y tienen, en los últimos meses, una presencia significativamente mayor. Sin embargo, se espera que este efecto de relativa victoria no se traduzca en empresas con actividades que dependan del uso intensivo del agua, debido al efecto relativo de la explosión social de 2019 y a las prácticas de despojo que han llevado a cabo en los territorios.

7. Relevancia

Este estudio ejecuta uno de los modelos más recientes e influyentes en el área de Procesamiento de Lenguaje Natural y lo propone como una alternativa plausible para los estudios sociológicos. Este modelo, que permite la extracción de características de texto, posee el potencial de extraer automáticamente información de documentos digitales, permitiendo registrar en tiempo real la reacción de los usuarios de redes sociales ante eventos de interés social. También aumenta el volumen de datos que se pueden procesar o clasificar dentro de una investigación al automatizar este proceso.

Como producto de esta investigación se ofrece tanto el código para aplicar este modelo como la explicación detallada de la metodología, con el fin de que sea utilizada en futuras investigaciones por aquellos investigadores que quieran implementar el modelo en otros conjuntos de datos y en relación con otros problemas.

Asimismo, en el ámbito de la sociología ambiental, genera información sobre la sequía, enfocándose en un aspecto del fenómeno que es la opinión pública sobre las entidades que participan en esta discusión. Este método permite, por tanto, ampliar la concepción que se tiene de los actores en conflictos ambientales, al incorporar la opinión pública en las comunicaciones que estos generan.

En lo referente a las limitaciones, el investigador producirá los datos de entrenamiento del modelo, por lo que la cantidad de estos datos será menor que lo deseado en la investigación, lo que disminuirá el rendimiento del modelo. Esto, dado que se requiere una gran cantidad de datos clasificados humanamente para obtener buenos resultados en este tipo de modelo.

El uso de Twitter como base de datos hace que la investigación fluctúe en torno a muchos sesgos, dado que los usuarios cuentan con muchas características comunes, como lo son la accesibilidad a internet, edad y el hecho de ser una red social con mayor presencia de hombres, por lo que las conclusiones deben tener presentes estas limitaciones.

8. Marco Conceptual

Los modelos de Procesamiento de Lenguaje Natural basados en la arquitectura de redes neuronales artificiales poseen ventajas sobre los modelos anteriores basados en estadísticas, y la probabilidad bayesiana, tales como: menor tamaño de memoria y potencia de procesamiento requeridos para ser implementados (Cho et al 2014). Esta aproximación a los modelos de lenguaje cambia y mejora rápidamente, hasta llegar a resultados igualmente auspiciadores como los obtenidos en modelos de imágenes en años anteriores (Ruder 2018). El cambio de paradigma se produjo gracias a la base de datos ImageNet, los modelos de preentrenamiento y las redes neuronales artificiales, convirtiéndose en estándares para la mayoría de los problemas de aprendizaje automático.

Ahora bien, para comprender la evolución de los modelos de Procesamiento de Lenguaje Natural, primero hay que entender que es "Next Sentence Prediction" en un "Language Model". Estos modelos consisten en predecir la siguiente palabra en una secuencia, permitiendo medir qué tan bien funciona un modelo en comparación con otros y, al mismo tiempo, entrenar modelos sin datos de entrenamiento clasificados (Liu, Cheung, and Louis 2019).

Los enfoques de probabilidad clásicos como Bag of Words asignan una probabilidad a cada palabra de pertenecer a una categoría, teniendo el modelo una dimensionalidad que corresponde a la extensión del vocabulario considerado (Dirac 2019). El problema de estos enfoques era que, en los documentos en lenguaje natural, el orden importa, existiendo la posibilidad de tener significados completamente diferentes según el orden de las palabras.

Considerando (Jurafsky y Martin 2008) el modelo de lenguaje de izquierda a derecha $p\ (x \mid c)$ donde el valor del contexto c determina la probabilidad de aparición de la palabra x podemos entender cómo funcionan los modelos de lenguaje cuando consideramos el contexto. Un bigrama, que es una metodología para modelos de lenguaje que sigue la probabilidad bayesiana, es un enfoque simplificado en cuanto a consideración del contexto refiere, porque es un modelo de lenguaje que considera solo la palabra anterior en la secuencia para hacer la predicción, $p\ (w_n \mid w_{n-1})$.

La simplicidad del bigrama los convierte en un modelo de Markov, los cuales establecen que es posible generar predicciones del futuro sin haber visto tan lejos en el pasado. Las cadenas de Markov son una secuencia de eventos probabilísticos, en los que la probabilidad de eventos depende del evento anterior. Un ejemplo de este tipo de modelo es el clima: el clima de un día depende en gran medida del clima del día anterior. El problema de los modelos de lenguaje es que requieren de más contexto sobre los eventos anteriores para hacer la predicción correcta del siguiente evento o, en este caso, requieren de más palabras previas para hacer la predicción de la siguiente palabra.

La dificultad de este enfoque, por tanto, fue el pésimo desempeño, debido a la consideración mínima del contexto. Además, las diferentes variantes de n-gramas (trigrama, cuadrigrama) cuentan, por su parte, con el problema de que eran altamente ineficientes, por tanto, imposibles de ser usadas cuando se consideran varias palabras del pasado. Esto, ya que la dimensionalidad del modelo crece de manera exponencial cada vez que se considera una nueva palabra hacer la predicción, y, en definitiva, no existe el poder de cómputo necesario para implementar dichos n-gramas.

En esta sección teórica se explican los conceptos básicos del Procesamiento de Lenguaje Natural que son necesarios para comprender las decisiones de mi investigación y las ventajas de la metodología seleccionada. El capítulo también revisará las perspectivas desde las que se realizan los estudios en torno al fenómeno de la sequía.

8.1 Fine Tuning

Fine Tuning es el proceso de utilizar una red neuronal previamente entrenada como parámetros iniciales para un nuevo modelo. El Fine Tuning ayuda a mejorar la velocidad de los modelos de entrenamiento y a superar los conjuntos de datos de tamaño pequeño. Para implementar esta técnica es necesario utilizar datos del mismo dominio de la tarea objetivo (texto, imágenes, audio, entre otros) para transferir parámetros iniciales que sean considerados relevantes en la nueva tarea. Generalmente, a este punto de partida se agrega una nueva "última capa" específica para la tarea deseada, con valores inicializados aleatoriamente (Dodge et al. 2020).

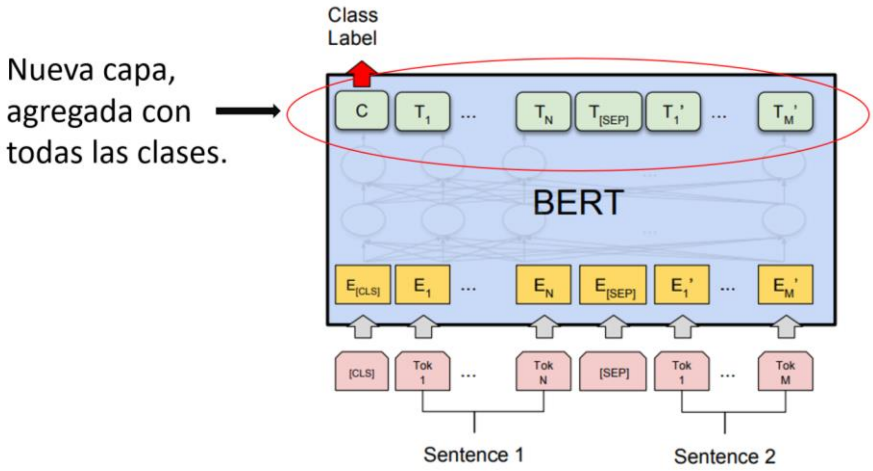

Fig. 8. Extraído de Devlin (2018).

La Figura 8 muestra la nueva capa agregada al modelo de pre-entrenado. Esta última capa necesita entrenar sus parámetros, mientras que la parte de pre-entrenada permanece con sus parámetros sin cambios, es decir, estados congelados (frozen).

Los resultados del proceso de Fine Tuning pueden variar en diferentes episodios de entrenamiento, incluso con valores de hiperparámetros fijos (Dodge et al. 2020). Mientras que los parámetros de los modelos son los pesos y sesgos (weights y biases) que posee el modelo, los hiperparámetros son los elementos que intervienen en la generación de esos valores de parámetros, como la tasa de aprendizaje, el número de unidades ocultas, el tipo de función de activación, entre otros. Es necesario probar distintos hiperparámetros para llegar al modelo más adecuado que evite el desajuste y el sobreajuste. En palabras sencillas, estos son el no aprendizaje de los patrones de los datos de entrenamiento y el memorizar, pero no generalizar la información de los datos de entrenamiento, respectivamente.

En Redes Neuronales Artificiales o Artificial Neural Network, cada neurona realiza una suma del producto de los pesos de la capa anterior, de los valores de las neuronas y de los sesgos, finalizando con una función de activación que genera la salida de las neuronas individuales.

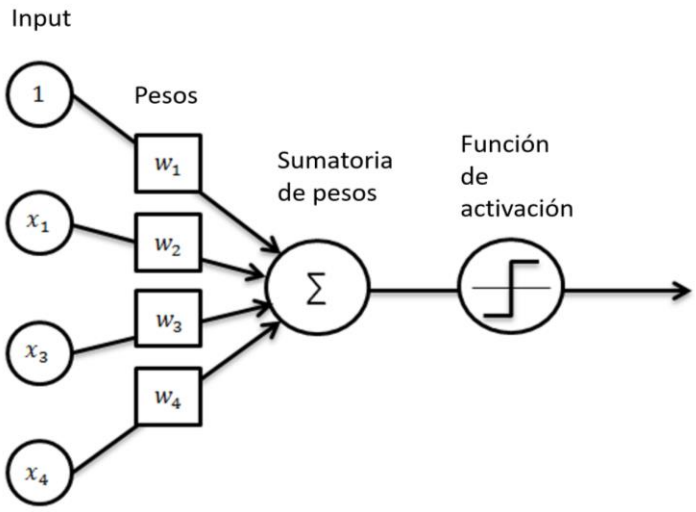

Fig. 9. Imagen extraída en Ognjanovski (2019).

La capa de entrada (o input), que se encuentra al comienzo de los modelos, posee un tamaño que depende de las características (o features) que la red artificial considera para el modelo. Estas características deben ser relevantes para incorporar información útil, por el contrario, las características irrelevantes generan ruido que perjudican el rendimiento del modelo. La capa de salida (u output) corresponde al número de clases o posibles categorías de salida. Finalmente, las redes neuronales artificiales se componen de capas ocultas (o hidden layers), que se encuentran en el medio de las capas de entrada y salida.

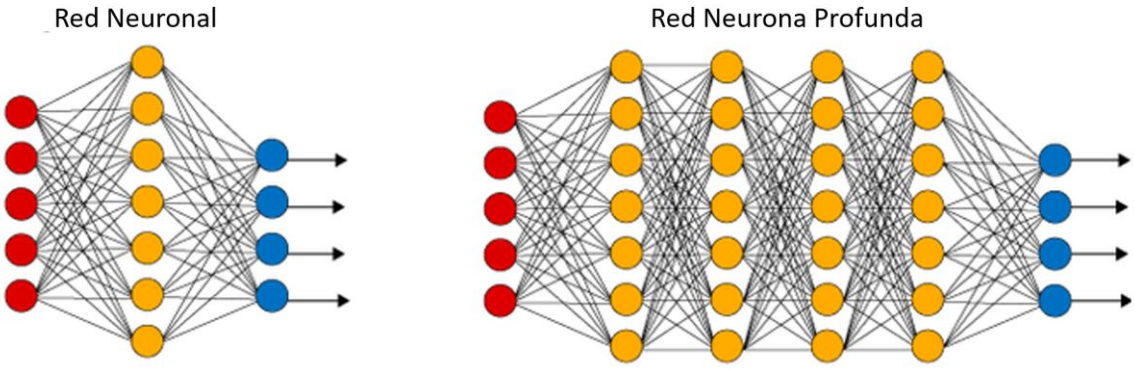

Fig. 10. Imagen extraída de DeepAI (2020).

Las redes neuronales profundas son aquellos modelos que poseen más de una capa oculta. La ventaja de aumentar el número de capas ocultas y aumentar la complejidad de los modelos es la versatilidad obtenida para manejar datos con distribuciones no obvias.

Los modelos que se utilizan en la presente investigación son modelos secuencia a secuencia, que toman como entrada secuencias de datos y generan otras secuencias de datos como salida (Alammar 2018). Las secuencias son grupos de elementos en relación. El orden de los elementos es tan esencial como el propio elemento para producir la secuencia, Como es el caso de los píxeles de una imagen, las palabras de una fase o las ondas en un audio. Además, en este tipo de modelo, la secuencia de entrada y salida puede diferir en longitud y tipo de datos. Por ejemplo, en la tarea de resumen, la secuencia de entrada es más larga que la secuencia de salida que resume la información. Otro ejemplo son modelos que generan descripciones de imágenes, en las que la entrada del modelo es un dato de tipo imagen y la salida son datos de tipo texto.

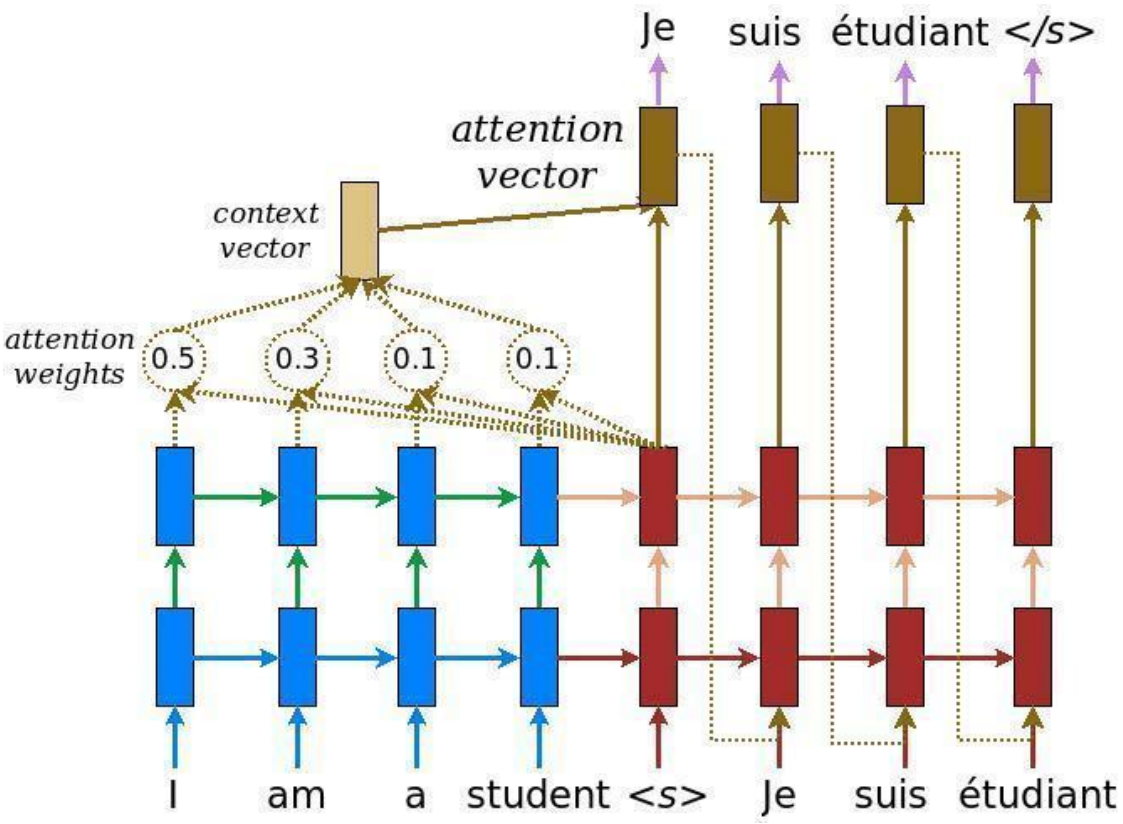

Fig. 11. Imagen extraída de TensorFlow (2020).

En la figura 11 se ve como la secuencia de entrada (frase de ingles) posee un tamaño o cantidad de palabras distinto a la secuencia de salida (frase en francés). Este modelo de secuencia a secuencia posee codificador y una parte de decodificador. El codificador procesa cada elemento en la secuencia de entrada, compilándolo en un vector llamado "contexto". Luego el vector de contexto se sincroniza con el decodificador que genera la secuencia de salida, elemento por elemento (Alammar 2018).

8.2 Reconocimiento de entidades nombradas

El reconocimiento de entidades nombradas es un problema de aprendizaje automático que busca detectar entidades dentro de un texto y darle una clasificación que puede ser predefinida o genérica.

Esta técnica forma parte de "extracción de características", que es un campo de machine learning que busca los algoritmos apropiados para extraer automáticamente información significativa de los datos digitales. El proceso de aprendizaje de estos algoritmos consiste en encontrar patrones y estructuras en los datos para optimizar los parámetros del modelo (Deisenroth, Faisal y Ong 2020), en este caso, los modelos requieren datos clasificados humanamente para aprender. Para aprender los modelos necesitan feedback sobre qué tan bueno es su rendimiento a través de una métrica (que es un hiperparámetro), la más común es la Accuracy, es decir, las predicciones correctas divididas por el número de predicciones realizadas.

$$Accuracy = \frac{True\ positive + False\ positive}{n}$$

El problema de esta medida es que puede generar que el modelo se centre en identificar correctamente una categoría cuando los datos no están equilibrados entre las diferentes categorías.

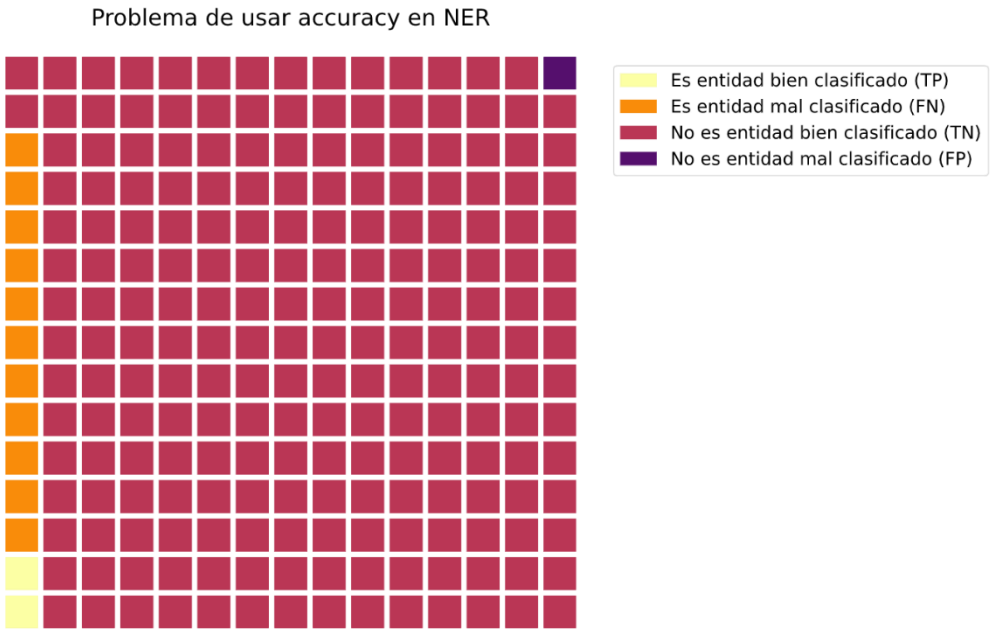

Fig. 12. Ejemplo de modelo que clasifica mal a las entidades.

En el caso del reconocimiento de entidades nombradas, la mayoría de los elementos de la secuencia no son entidades, por tanto, denotadas como "o". Modelos, como el del ejemplo anterior, representan un problema, ya que clasifican a casi todos los casos como "no-entidad", dado que esta estrategia de clasificación le permite tener una buena accuracy. Para resolver esto se requiere una medida que exija un buen rendimiento en la clasificación para todas las categorías del modelo.

El F1 score es la medida adecuada para determinar la calidad del modelo en el reconocimiento de entidades nombradas (Claeser, Kent y Felske 2018), este se define como la media armónica de la precisión y la exhaustividad, también llamados precision y recall respectivamente.

$$Precision = \frac{True\ positive}{True\ positive + \ False\ positive}$$

La precision considera todos los casos denotados como positivos y ayuda a determinar la proporción de casos denominados positivos acertados, esta medida por sí misma es útil para algunos problemas. Por ejemplo, una plataforma como YouTube Kids tiene la prioridad de

detectar si los videos no son apropiados para la plataforma, siendo no tan importante clasificar a un video apropiado como no apropiado.

$$Recall = \frac{True\ positive}{True\ positive + \ False\ negative}$$

Ahora bien, el recall es una métrica centrada en el costo de los falsos negativos, que puede ser fundamental en situaciones específicas, como la detección de enfermedades. Decir que alguien no tiene una enfermedad que sí posee es en ciertos casos mucho más dañino que decirle a alguien sano que eventualmente tiene una enfermedad. Por lo tanto, mientras que el recall es una medida que considera qué tan bien predice el modelo considerando los valores realmente positivos para medir su desempeño, la precision es una medida que toma en cuenta los valores que se pronosticaron como positivos para medir qué tan bien se clasifica el modelo.

$$F1 = 2 * \frac{Precision * Recall}{Pecision + Recall}$$

El F1 score (Valor-F) es la media armónica de precision y el recall, considera igualmente ambos parámetros. Es ideal para el problema de Reconocimiento de entidades nombradas debido a que la parte de recall de la fórmula resuelve el problema de falta de entidades clasificadas a través de castigar al modelo por poseer falsos negativos.

8.3 Aspect Based Opinion Mining

Esta investigación busca avanzar más allá de la identificación de las entidades en el texto. Adicionalmente, se busca encontrar la opinión expresada a esa entidad en la mención realizada. Para ese objetivo, la Aspect Based Opinion Mining (ABOM) es útil.

Como se muestra en (Zhang y Liu 2014), ABOM consiste en encontrar la opinión dada a una entidad, clasificándola luego en categorías personalizadas, siendo definido el aspecto como el concepto en el que se expresa la opinión. Para el caso de esta investigación, serán tres categorías positivas, neutras y negativas. Es importante entender, para futuras

investigaciones, que es posible crear otras categorías personalizadas, como confianza, orgullo, indiferencia, entre otras.

Además, es posible crear conjuntos de subcomponentes de las entidades, por ejemplo, un texto puede tener una opinión global positiva de un teléfono, pero, en el mismo texto, tener una opinión negativa de la calidad de audio del teléfono, que es un componente de la entidad. Incluso, es posible tener subcategorías de las subcategorías, como tener una opinión sobre la calidad de audio que posee el micrófono del teléfono. Sin embargo, como expresan Zhang y Liu (2014), no es conveniente incorporar muchos conjuntos de subcomponentes debido a la disminución del rendimiento de los modelos, considerando que la generación de modelos útiles es un proceso complicado en sí mismo.

Esta investigación se limitará a clasificar sin utilizar ningún subcomponente de las entidades. Esta decisión es tomada debido a las capacidades limitadas del estudio para producir datos clasificados humanamente. Sin embargo, en futuras investigaciones se recomienda utilizar conjuntos de subcomponentes, especialmente si la investigación se trata de una entidad específica.

8.4 Redes neuronales recurrentes

Recursividad, en términos generales, significa definir un objeto iterablemente, en donde la definición de la función incluye la aparición de su propia definición. Una relación recurrente es una ecuación definida recursivamente, en el arreglo de valores, cada elemento es una función que contiene el elemento anterior (Dirac 2019).

$$h_i = A(x, h_{i-1})$$

En términos de la teoría de grafos, este es un grafo dirigido, lo que significa que cada arista tiene direccionalidad.

Estas familias de modelos implementan un bucle en el que la salida de la capa anterior, denotada "hidden state" es ingresada como información para cada nueva capa, en conjunto a un nuevo elemento de la secuencia de entrada. Esto representa una mejora en consideración del contexto en modelos de lenguaje.

Por otro lado, la parte de codificación "encoder", que se encarga de procesar la secuencia de entrada “input”, posee unidades recurrentes que generan una cadena, esta recibe los "hidden state" y los "input".

Los inputs son la secuencia de palabras, estas palabras son transformadas en vectores para que puedan ser procesadas por el modelo, considerando que los modelos de Inteligencia Artificial solo pueden trabajar con números. Los textos, imágenes, audios y demás tipos de datos tienen que ser transformados a una representación vectorial de estos. El número con que se denota a cada palabra se le denomina su “token”.

Finalmente, se genera el último "hidden state", llamado "encoder stage" o contexto, el cual contiene la información de toda la secuencia $f(\vec{x}) = h_n$.

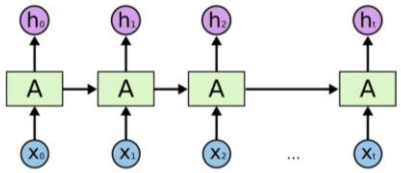

Fig. 13. Extraído de Dirac (2019).

Luego comienza la etapa de decodificación que funciona hasta que se activa un “token” que denota el final de la oración.

El problema de las Redes Neuronales Recurrentes, como se describe en (Britz 2016), es que genera una pérdida de información de los datos posicionados al comienzo de la secuencia, generando una disminución en el rendimiento de los modelos (Cho et al. 2014). Así mismo, otras limitaciones son la desaparición y la explosión de los gradientes, que son valores extremadamente grandes o pequeños que toman el "hidden state", producido por reiterados números de multiplicaciones.

$$h_n = w^n x_0 + w^{n-1} x_1 + \cdots$$

Cuando la secuencia es extensa, el problema es que el número de multiplicación es igualmente amplio, por eso, los valores que son inusualmente grandes o pequeños multiplicados repetidas veces consigo mismos terminan en valores extremos.

La programación secuencial, que es un método para procesar secuencias elemento por elemento y en orden, evita el procesamiento paralelo de la información de entrada, que es más rápido y ofrece la opción de procesar la información en múltiples máquinas. Finalmente, tenemos, sumado a la ineficiencia en el entrenamiento, un bajo rendimiento del modelo, el cual es atribuible a la incapacidad empírica de una matriz para, por si sola, capturar información de dependencias de secuencia larga o, en otras palabras, de recordar las palabras al comienzo de la secuencia procesada, olvidándose, por tanto, de lo dicho al comienzo del texto.

Long Short Term Memory Networks es una red neuronal recurrente que resuelve muchos de los problemas que presenta la versión básica de las redes neuronales recurrentes, en particular, la pérdida de las dependencias a largo plazo, dado que es una arquitectura diseñada específicamente para resolver este problema. Muchas veces, en los textos se da la situación de que las frases hacen referencia a información proporcionada en partes muy al comienzo de los textos. Por ejemplo: "Crecí en Francia ... hablo con fluidez ..."

Este problema se abordó directamente en los Long Short Term Memory Networks. El modelo propone gates (accesos) para agregar y eliminar información. Sin embargo, el problema es que se tarda mucho en entrenar y, en secuencias largas, la pérdida de información sigue ocurriendo, reduciendo así su rendimiento. Finalmente, el proceso de transferencia de aprendizaje nunca funciona bien. Por lo que esta arquitectura ha entrado en desuso durante los últimos años ante el advenimiento de nuevas alternativas.

8.5 Attention Mechanisms

Las diferentes variantes de modelos recurrentes creadas previo al desarrollo de los mecanismos de atención no han resuelto el problema de la pérdida de información en secuencias largas. En términos técnicos, el vector de contexto no deja de "olvidar" las palabras colocadas al inicio de las secuencias (Venkatachalam 2019a).

En las redes de atención “end-to-end” propuestas por Sukhbaatar et al. (2015) se resuelve este problema, al codificar la secuencia de palabras en muchos “hidden state”, así, se incorpora un hidden state para cada palabra en la oración (codificador) y las palabras

procesadas en la secuencia objetivo (decodificador) que ya se predijeron (Britz 2016). Estos hidden state se consideran en cada paso de la etapa del decodificador con atención.

Como se expresa en Weng (2018), el decodificador opera en la posición t, donde el vector de contexto c_t es la suma de todos los hidden state en la secuencia de entrada, y la secuencia ya decodificada y_{t-1} está siendo tomada en consideración, generando la secuencia de salida hasta producir el end-of-sequence token.

$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$

El modelo de alineación $\alpha_{t,i} = aling\ (x_i, y_t)$, asigna un valor $\alpha_{t,i}$ que mide cuánto considera el hidden state en cada salida, al par de entrada en la posición i y la salida en la posición t, (x_i, y_t) en función de qué tan bien coincide el modelo. Atención es el promedio ponderado de los valores (Dontloo 2018).

$$c = \sum_j a_j h_j$$

Donde $\sum_j\ a = 1$

La matriz de score de alineación permite ver la relación entre las oraciones fuente y objetivo.

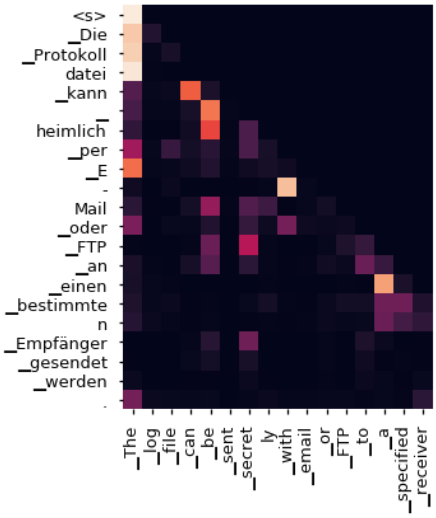

Fig. 14. Ejemplo de traducción extraído de Klein (2017).

Cómo se puede ver en la figura 14, se presta una atención diferente a las palabras del idioma original por cada palabra traducida, porque hay palabras que son más importantes que otras para continuar con la traducción. El sistema tiene la opción de prestar atención de palabras en el pasado y futuro de la palabra traducida.

El codificador mapea la secuencia de entrada, una representación simbólica, para generar una representación continua (Vaswani et al. 2017). Las salidas del decodificador dependen del objetivo del modelo, y pueden ser de otro tipo de dato, como pasar audio a texto como se ve en la figura número 15.

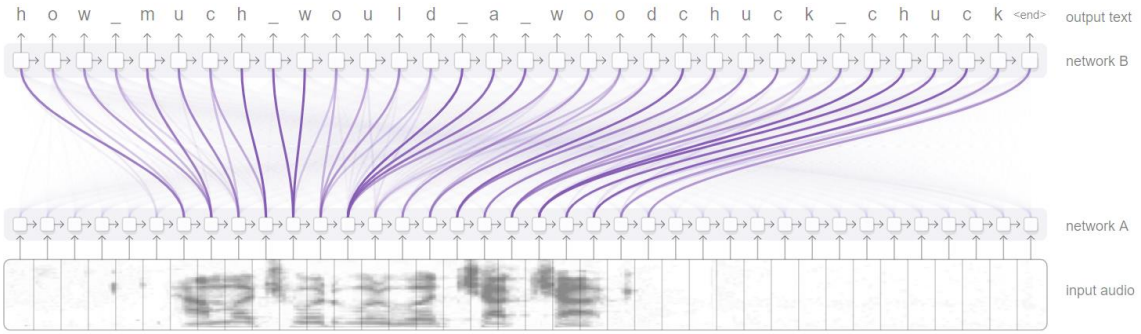

Fig. 15. Extraído de Olah y Carter (2016).

La Figura 15 muestra la conversión de audio en texto, que son datos de diferente tipo. Esta arquitectura de atención está inspirada en los Mecanismos de Atención Visual que tiene el cerebro para enfocar la información recuperada por los ojos. Esto es enfocándose en un punto en particular, prestando menos atención a otras partes de la imagen (Britz 2016), siendo esta la razón por la cual los mecanismos de atención se centran en un subconjunto de la información que reciben.

El problema que trae consigo que la función considere todos los hidden state para cada paso de tiempo del decodificador es que aumenta considerablemente el número de cálculos necesarios para procesar las secuencias (Venkatachalam 2019a). Esto debido al número de cálculos necesarios para obtener el vector de probabilidad alfa (Dontloo 2018).

$$e_{ij} = a(s_i, h_j), \quad a_{i,j} = \frac{exp\,(e_{i,j})}{\sum_{k=1}^{T_x} exp\,(e_{i,k})}$$

Siendo h la secuencia del codificador y s la secuencia del decodificador. El problema de eficiencia de la atención fue que, considerando una largo m para la secuencia del codificador y un largo n para la secuencia del decodificador. Se requiere pasar por esta red $m * n$ veces para adquirir todas las puntuaciones de atención e_{ij}.

Las aproximaciones recurrentes, como los primeros mecanismos de atención, tienen la desventaja de alinear sus posiciones con el tiempo de cálculo para generar la secuencia de "hidden state". Por eso se generó la necesidad de procesar la información por pasos. El proceso de entrenamiento se vuelve una cadena en la que es necesario esperar a que cada etapa oculta pase a la siguiente, excluyendo la posibilidad de paralelismo dentro de los ejemplos de entrenamiento, lo que significa procesar simultáneamente toda la secuencia de entrada (Vaswani et al. 2017). Este problema fue resuelto por los Transformers.

8.6 Transformers

Los transformers, propuestos originalmente en Vaswani et al. (2017), son una arquitectura que elimina la necesidad de generar un hidden state para cada elemento de la secuencia de entrada. Incorporando una nueva metodología para calcular el key k, que representa la secuencia de entrada, el query q, que representa la última secuencia decodificada, y el value v que es una nueva representación de la secuencia de entrada.

Estos tres conceptos provienen de los retrieval systems, que son sistemas que recuperan elementos. En el contexto de Internet, estos elementos generalmente son contenidos, por ejemplo, en los videos de YouTube. Una "query", en el contexto de YouTube, es la solicitud de videos, evaluados con las "keys" de los videos (título, descripción, etiquetas de video, entre otros). El sistema recupera los "value" de los videos que mejor coinciden. En el contexto de la atención, el h representa el "value" (Dontloo 2019).

Los Transformers son diferentes a los enfoques de atención presentados antes, para calcular el score $s_{j,i}$.

$$e_{ij} = f(s_i)\, g(h_j)^t$$

El producto escalar de k y q, en este caso $k = f(s_i)$ y $q = g(h_j)$ respectivamente, dan una puntuación de cómo se presta atención a cada secuencia de entrada dado el decodificador

obtenido. La suma del vector de valor da el resultado final y permite implementar el self-attention tanto en el codificador como en el decodificador. (Venkatachalam 2019b). Ahora solo es necesario calcular $g(h_j)$ unas m veces y $f(s_i)$ unas n veces (Dontloo 2019).

Así, los Transformers se pueden entender como una arquitectura que combina matemáticas convolucionales con el mecanismo de atención, inicialmente propuesto en Vaswani et al. (2017). La propuesta original consta de un conjunto de seis codificadores y seis decodificadores.

Todos los codificadores están compuestos por las mismas partes, estas son, un mecanismo de self-attention, que opera sobre cada uno de los componentes de la secuencia de entrada y una red neuronal de retroalimentación.

Cada una de las entradas se multiplica por tres capas diferentes (vector de query, vector key y vector de value) que presenta una dimensión más pequeña (64 en comparación con las 512 dimensiones de las incrustaciones de palabras).

En el artículo original Vaswani (2017) plantea un enfoque puramente basado en los mecanismos de atención. El codificador posee seis capas idénticas, que al mismo tiempo poseen como subcapas un mecanismo de self-attention de múltiples cabezales y una red de alimentación hacia adelante (feed forward neural networks) conectadas a la salida de la parte del self-attention. Cada una de las cabezas de atención realizan los mismos cálculos, siendo posteriormente concatenados (Dontloo 2019).

$$MultiHead\ (Q, K, V) = Concat(head_1, \dots, head_h)W^O$$

donde $head_i = Attention\ (QW_i^Q, KW_i^K, VW_i^V\ )$

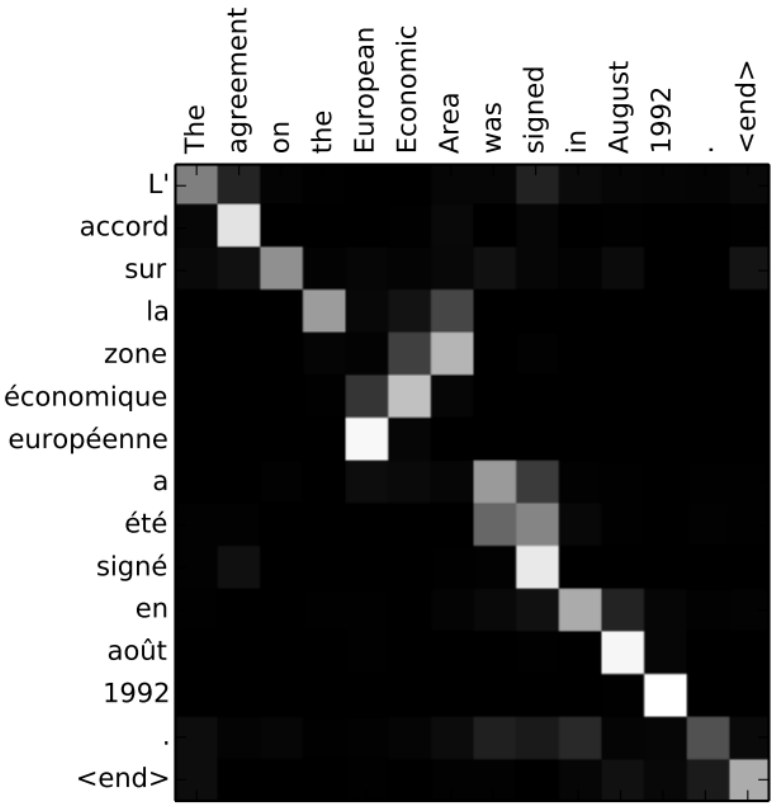

Fig. 17. extraído de Vaswani (2017).

La figura 17, extraída del artículo original de Vaswani (2017), muestra el mecanismo de atención que opera sobre los diferentes elementos de la secuencia de entrada. En este contexto, el self-attention es un mecanismo que se enfoca en diferentes posiciones de una sola secuencia (Vaswani et al. 2017). En el caso del codificador, el mecanismo de auto atención opera sobre toda la secuencia de entrada. La parte del decodificador es similar al codificador. Sin embargo, incorpora una nueva subcapa, que despliega mecanismos de atención sobre la salida generada en la etapa codificada (el vector con el contexto). En el caso del decodificador, la self-attention opera sobre la secuencia de salida ya generada (Vaswani et al. 2017).

8.7 BERT (Bidirectional Encoder Representations from Transformers)

Los modelos de lenguaje generan parámetros de modelo útiles para el fine tuning, esto es usar modelos pre-entrenados los cuales son entrenados por lo general mediante la resolución de tareas sobre datos no etiquetados. Por lo general, utilizan la tarea de predicción de la siguiente oración, que utiliza la información del lado izquierdo de la oración para realizar una predicción de la siguiente palabra de la secuencia.

Sin embargo, debido a que BERT es un modelo bidireccional, tiene que innovar para poder realizar el entrenamiento al modelo pre-entrenado que será posteriormente utilizado mediante el proceso de fine tuning en otras aplicaciones. La característica bidireccional significa que el proceso de entrenamiento implica predecir la palabra que sigue una secuencia (de izquierda a derecha) y predecir la palabra anterior en una secuencia (de derecha a izquierda) al mismo tiempo.

Para entrenar BERT, se usan dos tareas: (1) Modelo de lenguaje enmascarado, que predice palabras en una oración, reemplazando el 15% de estas con un token enmascarado, teniendo como contexto los tokens hacia atrás y adelante sin máscara; y (2) una predicción de la siguiente oración, que toma en consideración el contexto entre oraciones al predecir si una oración viene después de otra oración con una proporción de verdadero y falso del 50% (Palacio 2020).

Esta forma de entrenamiento permite obtener mejores resultados que los modelos anteriores de similar tamaño como GPT 1 y ELMo (Devlin et al. 2018). El artículo original Devlin et al. (2018) especifica cuatro formas de procesar las oraciones.

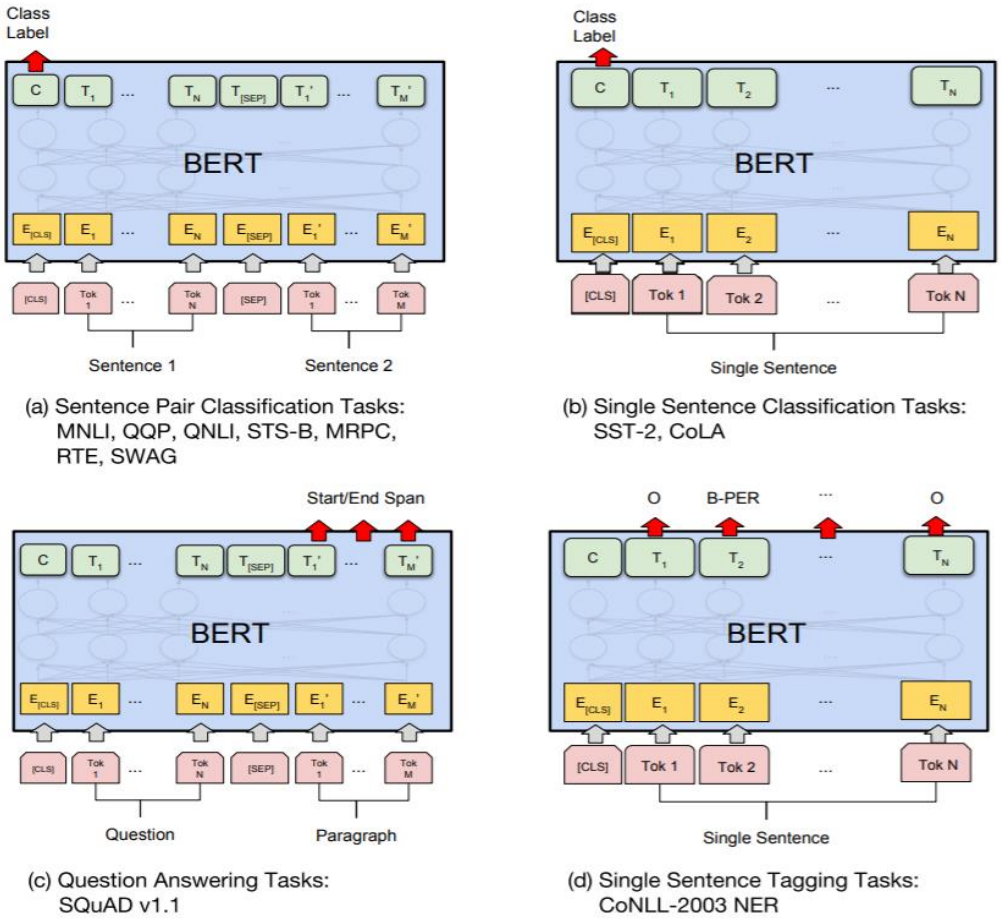

Fig. 18. Extraído de Devlin (2018).

Dos de las tareas que el modelo BERT realiza después del fine tuning son sobre las oraciones por sí mismas, que consisten en tareas de clasificación (como análisis de sentimiento) y etiquetado (como identificación de entidades). Otra de las tareas permite la generación de otras secuencias (como la Traducción) y el trabajar con pares de secuencias (como determinar lo correcto de una respuesta en una pregunta de desarrollo). Los usuarios de los modelos pre-entrenados pueden adaptar el modelo para su aplicación deseada.

8.8 Percepciones de mega-sequía

Las sequías son fenómenos con diferentes aristas desde las que analizar. A las causas físicas de la sequía se han sumado las políticas de despojo que se han extendido por toda América Latina, por lo que existen diferentes perspectivas de estudio. Uno esencial es el

análisis de los fenómenos físicos (Panez Pinto 2018). Esta perspectiva permite comprender los efectos de la sequía más alejados del campo de acción de las comunidades individuales, limitándose a la dimensión material del espacio. Por esta razón, el fenómeno se suele denominar un problema y no un conflicto, sin análisis político.

Por otro lado, la perspectiva característica de las Ciencias Sociales es donde el análisis proviene de un terreno a-espacial (Panez Pinto 2018). En esta interpretación, el espacio se conceptualiza como un fenómeno social, tomando como focos principales (Panez Pinto 2018): a) las relaciones simbólicas con el agua, y b) la asimetría de poder de los actores, desde donde el análisis de las políticas privatizadoras juega un papel protagónico. El aspecto en el que se limita este análisis es en su enfoque relacional, subrayando la materialidad en la que ocurren las relaciones espacializadas.

Finalmente, está la perspectiva que cuestiona la relación entre sociedad y naturaleza (Panez Pinto 2018), desde la cual se pretende superar la separación físico-simbólica, entendiendo el agua como producto-productor de relaciones sociales. Esto nos permite comprender las relaciones de poder en el proceso de producción tecno-social que organiza el flujo de recursos, a partir de los cuales sirven como muestra de la distribución social del poder.

Desde esta última perspectiva, se adoptan las críticas características de las comunidades latinoamericanas en resistencia, como la fetichización del agua, donde la perspectiva de los derechos humanos conduce a la mercantilización, haciendo que la cultura se desprenda de la naturaleza extraída (Panez Pinto 2018).

Los despojos del modelo de desarrollo extractivista han provocado una gran sensibilidad en las comunidades afectadas, llegando a la idea de que el aumento de la demanda de agua sería la principal causa de la sequía (Bolados García et al. 2018). El incremento más significativo de la actividad productiva dependiente del agua se produjo desde la década de los noventa (Bolados García et al. 2018). En este periodo, Chile inició una política para convertirse en una potencia alimentaria mientras experimentaba una expansión considerable en la inversión minera privada (Bolados García et al. 2018), acompañado de una expansión de las políticas que favorecieron la incorporación de grandes empresas. Además de la extracción de agua, se sumaron los programas de estímulo y crédito a la agricultura, que solo llegarán al 4% para la agricultura familiar campesina, junto con los

proyectos de la comisión nacional de riego, orientados principalmente a la agricultura de exportación (Bolados García et al. 2018).

Estos cambios en el mercado agrícola se produjeron mientras la agricultura familiar campesina tenía que hacer frente a la sequía, con la consiguiente pérdida de ganado y plantaciones, lo que provocó que las personas terminaran obligadas a trabajar en otros sectores, como el minero y el inmobiliario (Bolados García et al. 2018).

En cuanto a las percepciones de la población frente a los eventos de sequía según (Aldunce et al. 2017), se destaca su reconocimiento en torno a los cambios climáticos, con la disminución de las lluvias y la diferencia de paisajes siendo los cambios más reconocidos. En las causas que llevaron a la sequía, el aumento de la demanda destaca como el factor más mencionado, mientras que alusiones al cambio climático natural y el antropogénico reciben pocas observaciones. Entre los impactos que tuvo la sequía, la sequedad del paisaje fue el factor más citado, y en cuanto a la oferta laboral, se reconoce una disminución de la actividad económica de la agricultura y el turismo. Existieron menciones de las poblaciones que vieron afectada su capacidad de subsistencia y se vieron obligadas a migrar, junto con los conflictos surgidos entre vecinos por el manejo del recurso hídrico, con algunas menos menciones sobre la pérdida de fauna.

Esta condición sugiere que la opinión pública sobre el tema de la sequía se concentra "principalmente" en los efectos actuales sobre las comunidades que vieron deterioradas sus métodos de subsistencia. Los cambios más vívidos son los observados en el panorama y los cambios en la oferta laboral, junto con las consecuencias que afectan principalmente a las poblaciones de bajos ingresos e indígenas, pues son las principales preocupaciones y, en menor medida, las apreciaciones del cambio antropogénico y el fortalecimiento de la capacidad de resiliencia del país existen en el discurso colectivo.

9. Marco Metodológico

En esta sección se expondrá la metodología utilizada en esta investigación, mostrando las tareas de Procesamiento del Lenguaje Natural, la selección de variables y la metodología para la extracción de datos de la red social.

9.1 Diseño de la investigación

Para la investigación se llevó a cabo la tarea TASBA, que consistió en detectar la mención en el texto de las entidades en el párrafo y la opinión expresada hacia dicha entidad. El modelo es lo suficientemente flexible como para incluir tantas categorías como se requieran, pero demasiadas categorías pueden socavar el rendimiento del modelo. Posteriormente, en la investigación se realizó un análisis cualitativo para mejorar la interpretación de los datos a través de la profundización en el sentido en que se vierte la opinión.

Para construir la base de datos se extrajo de la red social Twitter un conjunto de tuits que tratan el tema de la sequía.

En esta investigación, se utilizó la API (aplicación de desarrollador) de Twitter para extraer información de la plataforma. Esta plataforma solicita la creación de un usuario para acceder a esta información y un proceso de postulación para acceder a la información que la API ofrece. La cuenta gratuita utilizada para obtener información tiene algunas limitaciones, como restringir el número de tuits a los que se puede acceder cada 15 minutos (Twitter 2020), además de evitar la obtención de tuits que tienen más de una semana de antigüedad. En esta investigación, los tuits junto con toda la información que contienen, fueron descargadas diariamente desde el 11 de abril hasta el 11 de diciembre con estas limitaciones.

Con la intención de mejorar el rendimiento de clasificación se utilizó una versión en español del modelo BERT (Cañete et al. 2020) como modelo pre-entrenado, que posee un tamaño similar al modelo BERT base. Los pesos típicos para inicializar las nuevas últimas capas de los Transformers, en el proceso de fine tuning, son una distribución normal con una media de 0 y una desviación estándar de 0.02, con una semilla recomendada de 12, siendo este el número que genera números aleatorios en la nueva capa (Dodge et al. 2020). El artículo original de BERT recomienda usar tres o cuatro “epochs”, que refiere a cuantas veces el modelo ejecuta el algoritmo de aprendizaje (Devlin et al. 2018), recomendaciones que fueron seguidas en esta investigación.

9.2 Definición de las variables

Las categorías para la investigación fueron seleccionadas considerando tanto entidades como opiniones. Se determinó que las entidades adecuadas para atender la opinión pública de la sociedad chilena son 1) organismos públicos, 2) organismos privados y 3) localidades. Las críticas al régimen de propiedad del agua motivaron la incorporación de entidades públicas debido a que se esperan críticas sobre el proceso de formulación de políticas y cuestionando los vínculos entre políticos y propietarios de derechos de agua. En cuanto a las entidades privadas, se consideró necesaria su presencia por acusaciones de acaparamiento hacia industrias con uso intensivo de agua (minería y agricultura). Finalmente, se analizaron las localidades por ser un fenómeno que afecta a poblaciones específicas de manera territorializada, promoviendo conflictos territorializados.

Las opiniones seleccionadas para esta investigación fueron 1) positivas, 2) neutrales y 3) negativas. Estas son las opiniones que se utilizan habitualmente en los estudios de opinión pública. Además, tienen poca complejidad para el clasificador, por lo que fueron seleccionados. En la tarea de clasificación de TASBA, es común incluir opiniones inconsistentes, que es cuando se menciona opiniones dispares a una entidad, entregándose puntos de vista contradictorios, como hablar bien y mal de algo en un mismo texto, inicialmente se pensó en incluir esta categoría, pero para esta investigación se descartó esta idea debido a la brevedad de los textos que tienen los tuits, ya que este tipo de categoría es más apropiada cuando se utilizan textos con mayor extensión y es posible expresar opiniones más complejas.

Tanto para las instituciones públicas como para las privadas, la condición requerida para que exista su mención es que se mencione ya sea una institución o un miembro de dicha institución. Un ejemplo de mención de un miembro de la institución sería un comentario sobre un alcalde. Se interpretó como un comentario realizado al municipio, por tanto, una institución pública. En cuanto a los lugares, se entendió como positiva o negativa la mención en la que el apoyo o rechazo a la situación que están viviendo, haciendo referencia a la sequía, muestras de apoyo o empatía se consideró positiva. Por el contrario, el rechazo tanto de su situación como de sus acciones se consideró negativo. Para comentarios neutrales, solo se consideran menciones sin expresar opinión, como descripciones fácticas en las que no hay evaluación.

9.3 Población y muestra

Población: Todos los tuits que tratan el tema de la sequía desde abril a diciembre del 2020, emitidos desde Chile.

Muestra: Alrededor de 2.500 a 5.000 tuits por día, dando un total de 773.850 tuits descargados. Esperando ser filtrados por origen geográfico y temático, dando un total de 585 tweets clasificados manualmente.

Instrumentos: Se utilizará la herramienta de búsqueda proporcionada por Twitter para encontrar los tuits que tratan sobre la sequía en Chile. La herramienta de búsqueda que posee la aplicación de desarrolladores.

10. Análisis de resultados

En esta sección se detalla la implementación de varios modelos clasificadores que cumplen con la tarea TASBA. Como métodos de clasificación se aplican variantes de BERT para realizar las tareas TABSA y modelos LSTM para desarrollar un asistente de clasificación.

10.1 Construcción de la base de datos

Como esta investigación utiliza datos no estructurados de Internet, el mayor desafío de la minería de datos fue extraer los datos útiles para la investigación. La fuente de la base de datos fue la API de Twitter, a través de la cual el buscador de la plataforma fue utilizado, solicitando la búsqueda de datos a través del toquen "sequía".

Cada tuit recibido tiene una serie de información sobre el emisor, incluyendo el texto, los hashtags mencionados, los usuarios mencionados, el nombre del usuario, la descripción del usuario, la ubicación del usuario, entre otros. Para esta investigación, tanto el texto como la ubicación del usuario son los únicos datos utilizados. El único propósito de la ubicación fue garantizar que el usuario esté dentro de Chile.

La desventaja que posee este método de extracción de datos de Twitter sobre otros es la gran cantidad de casos no relacionados con la investigación que son recopilados, ya que los usuarios usan el término "sequia" para referirse a otros temas. Estos casos deben detectarse y eliminarse del conjunto de datos. Asimismo, se seleccionó este método de búsqueda por token porque la alternativa de usar hashtags para buscar tuits trae otros inconvenientes, como: hay temas, como sequía, que no tienen un hashtag tan potente que aborde el tema, y alternativas como #NoEsSaqueoEsSequía se consideraron no lo suficientemente fuertes para contener la discusión, es decir, la gente habla de la sequía sin usar el hashtag.

Además, un problema con el hashtag es que muchas veces no se refieren a un tema, sino que refieren a una posición sobre el tema, como asociar la sequía con el saqueo, lo que impediría obtener información de otras posiciones del tema. Esta utilización de hashtags no neutrales se produce, como se menciona en los antecedentes, porque la propia Twitter se convierte en un campo de disputa política.

La gran mayoría de los tuits recopilados consisten en retuits y muchos retuits manuales, los cuales, con el fin de entrenar al modelo, fueron excluidos del conjunto de datos.

Los retuits realizados por la plataforma fueron fáciles de extraer ya que usan RT como prefijo, por lo que fueron extraídos. Mientras que los retuits manuales fueron retirados en la medida en que fueran copias idénticas de un tuit anterior. Un defecto de este último método fue que hubo tuits idénticos que sufrieron modificaciones mínimas. No se utilizó ningún método en esta investigación para extraer estos casos de la base de datos.

El proceso de clasificación funciona mediante un bucle que recorre toda la base de datos. La clasificación humana se produce tras saltarse aquellas carpetas de días ya clasificados. El proceso de verificación funciona día a día porque la clasificación de la base de datos ocurre durante varias jornadas, debido a la gran cantidad de datos.

Para el etiquetado manual de datos se implementó una metodología de etiquetado asistido, donde un modelo LSTM determina, en primera instancia, si el usuario es residente en Chile o no y después determina si el tuit trata el tema de la sequía o no. Posteriormente se resuelve humanamente si se cumplen ambas clasificaciones de los modelos LSTM. Las ventajas de este tipo de clasificación son reducir la cantidad de error humano al clasificar y posibilitar que la clasificación se haga de manera más rápida.

Cada vez que una carpeta que contiene un día de tuits se clasificó, se entrenó un nuevo modelo LSTM. Este se entrena con los datos recientemente etiquetados junto con los datos ya disponibles. Por lo tanto, después de terminar de clasificar un día de tuits, el asistente de clasificación mejora.

Este método de etiquetado asistido se utiliza, pues detectar y clasificar aspectos a nivel de oración es, recordemos, una tarea de NLP difícil, por lo que se requiere etiquetar la mayor cantidad de datos posible para hacer viable que los modelos entrenados obtengan un buen rendimiento.

El clasificador que determina si el tuit trata el tema de sequía o no, se emplea debido a que el sistema de búsqueda de Twitter no distingue cuando el concepto de sequía usado en otro contexto. Los casos típicos en los que se utiliza la palabra sequía en otros contextos dentro de los tuits son en el deporte, refiriéndose a largos períodos en que no se gana y por los

fanáticos del k-pop para referirse a largos períodos en los que los grupos que siguen no generaron nuevos contenidos.

El objetivo final de estos sistemas de asistencia fue que, una vez que los dos modelos terminan de estar bien entrenados, la clasificación humana se realizara únicamente sobre los datos útiles para seguir la investigación. Esto es detectar tanto las entidades, su categoría y polaridad de la opinión. Sin embargo, para lograr tal nivel de precisión, se requiere una gran cantidad de datos clasificados manualmente, lo que se simplifico ya que el modelo tuvo la capacidad de descartar una cantidad significativa de datos no útiles para la investigación, acelerando la construcción de la base de datos.

10.2 Etiquetado de datos

Los datos de ubicación fueron etiquetados de la siguiente manera, se colocó un 0 si el lugar descrito no corresponde a Chile y 1 si corresponde a Chile. También, se le dio una clasificación dicotómica a los tuits, el cual determina si el texto trata sobre sequía o no. Siendo 0 el que no se trata la sequía y 1 el que se trata la sequía.

Posteriormente, los datos de las entidades y su opinión se almacenaron dentro de una cadena de caracteres, donde el primer carácter brinda la información sobre el tipo de aspecto (pública (p), privada (e) o lugar(l)). Luego se agregó la polaridad de la opinión (positiva (p), neutral (e) y negativa (n)) y finalmente el token en el que se menciona el aspecto.

10.3 Número de datos extraídos

A continuación, se muestra una tabla con los datos extraídos. En esta tabla se detalla la cantidad de tuits procesados y cuántos de ellos terminaron siendo útiles.

Conjunto de datos	Cantidad
Cantidad de tuits recopilados	773.850
Tuits revisados, incluyendo los retuits, casos donde no se trata la sequía y casos donde se trata el tema de la sequía, pero que no fueron emitidos desde Chile. También están los tuits que eran útiles para la investigación, pero no fueron detectados por el modelo	231.708
Tuits no repetidos que tratan la temática de la sequía en cualquier ubicación (etiquetado manualmente)	3.247
Tuits no repetidos que tratan la temática de la sequía con Chile como ubicación declarada (etiquetado manualmente)	1.004
Tuits no repetidos que tratan la temática de la sequía con Chile como ubicación declarada y que cuenta con mención de entidades (etiquetado manualmente)	585

Fuente: Elaboración propia.

La base de datos se redujo debido a los retuits, usos del término investigado en otros contextos y usuarios que hablaban del tema fuera de Chile o que no mencionan ninguna

entidad. De esta forma, se terminó con una base de utilizables de 585 casos, estos textos cuentan con una o varias menciones a entidades, en los cuales se emiten o no opiniones respecto a su actuar ante los episodios de sequía que atraviesa Chile.

Una de las limitaciones de la investigación fue la limitada capacidad para construir datos de entrenamiento, por lo que fue imposible procesar toda la base de datos extraída de la página. También hay que considerar que la gran mayoría de los datos extraídos (entre 2.500 y 5.000 por día) acaban siendo retuits.

En cuanto a la posibilidad de la existencia de bots que publiquen tuits en la plataforma para alterar el debate sobre la sequía hay que aclarar que, por un lado, esta investigación no tiene ningún método para detectar bots, por lo que no se descartó ninguna cuenta en el conjunto de datos. Sin embargo, durante el proceso de clasificación de datos, fue sorprendente la cantidad de retuits manuales que recibieron tuits específicos a lo largo del tiempo, aunque da la impresión de que la mayoría de estos fueron robos de contenido y no bots. Por lo tanto, para futuras investigaciones, no se recomienda la implementación de metodologías para detectar cuentas de bots, ya que parece que este no es un tema lo suficientemente controvertido como para un uso descarado de estos que altere las conclusiones del estudio, como ocurre con otros temas como lo son, por ejemplo, las opiniones políticas durante las elecciones.

10.4 Preprocesamiento de texto

El preprocesamiento de los datos es un paso en el entrenamiento de modelos de inteligencia artificial en el que los datos que recibe el modelo reciben modificaciones para permitir y facilitar el aprendizaje de los mismos modelos. Los modelos BERT (tanto su versión en español como su versión multilingüe) que utilizados en esta investigación, solo funcionan con caracteres de texto en formato ASCII. Este formato es un conjunto de caracteres que se caracteriza por ser muy limitado, ya que solo tiene 127 caracteres únicos. Su desarrollo originalmente estaba pensando en lidiar con la memoria limitada de las computadoras antiguas, ya que permite almacenar el alfabeto del inglés y algunos otros caracteres usando

un mínimo de memoria. Por su parte, Twitter, como toda la internet contemporánea, utiliza UTF-8 como sistema de codificación de caracteres, esto incluye tanto los caracteres específicos del español (á, é, í, ó, ú, ñ), caracteres de otros idiomas y emojis.

Esta condición de utilizar caracteres ASCII existe porque se ha demostrado que reducir el universo de caracteres de los modelos aumenta el rendimiento de estos. Los caracteres no admisibles se eliminan de las palabras del texto para cumplir con este requisito del modelo. Además, en los casos que sea posible, estos caracteres se sustituyen por su versión ASCII. Este cambio hace que los caracteres no posean acentos, los caracteres especiales son removidos del texto junto con los emojis, mientras que las eñes son reemplazadas por enes.

Otro cambio en el texto realizado para aumentar el rendimiento fue eliminar los stop words, ya que estas solo agregan ruido al modelo, ejemplos de estas palabras son: y, ya, yo, él, otros, para, pero, entre otras.

Además, para reducir el ruido que debe tratar el modelo, se tomaron las siguientes decisiones:

1. Las menciones a los usuarios realizadas en la plataforma fueron reemplazadas. Estas fueron detectadas por el carácter que usa la plataforma para identificar las cuentas "@". El reemplazo fue el token de "usuario".
2. Para reemplazar los enlaces que los usuarios incorporan a los textos se seleccionó el token "link". Los usuarios incorporan estos enlaces con las herramientas para compartir que poseen diferentes páginas web para difundir su contenido en Twitter, por lo que tienen el mismo prefijo "https: //".
3. Ambos fueron reemplazados usando expresiones regulares como metodología. Este método permite crear un set de caracteres a partir de la creación de reglas. Para este caso, fue necesario utilizar el prefijo, poniendo el carácter de espacio como final de la secuencia capturada. Este método permite reemplazar los tokens que generan ruido. La expresión regular recupera el índice de la porción de texto que debe ser reemplazado, y con esta información el código ejecuta el reemplazo.

La siguiente tabla muestra cómo estos cambios afectan el texto para permitir que los modelos puedan procesarlos y de manera eficiente.

Versión	Texto tras codificación
Texto original	@jkhxgucci Las jikukas estamos en sequía ⁉️ https://t.co/EjdVanYO9l
Texto en formato ASCII	@jkhxgucci Las jikukas estamos en sequia https://t.co/EjdVanYO9l
Texto con los stop words removidos	@jkhxgucci Las jikukas sequía ⁉️ https://t.co/EjdVanYO9l
Texto en formato ASCII y con los stop words removidos	@jkhxgucci Las jikukas sequia https://t.co/EjdVanYO9l
Versión final, reemplazando aquellas palabras que generan ruido al modelo	user Las jikukas sequia link

Fuente: Elaboración propia.

También se requiere ejecutar el preprocesamiento sobre los nuevos casos que sean clasificados por la versión entrenada del modelo.

El set de palabras utilizado para entrenar cada modelo fue limitado por vocabulario manejado por cada versión de BERT. Todas las palabras que no se incluyan en este conjunto de palabras reciben el token de palabra desconocida. Esta limitación existe porque

la evidencia muestra que limitar el vocabulario ayuda al modelo a dar significado a las palabras que quedaron dentro del vocabulario, dando mejores resultados en la clasificación.

En la versión en español del modelo BERT, el conjunto de palabras se limita al español y algunas palabras de otros idiomas, mientras que, en el BERT multilingüe, este conjunto de palabras incluye tokens de varios idiomas.

Existen otras metodologías para limitar el vocabulario, como reemplazar conjuntos de palabras con sinónimos. Sin embargo, estos métodos se encuentran en desuso, ya que su efecto ha disminuido en los últimos modelos de NLP desarrollados, su efecto es especialmente notorio en los modelos de tipo probabilístico, los cuales se encuentran actualmente en desuso.

Un problema al trabajar con datos de Twitter es que la gente usa trucos para superar el límite de caracteres establecido por esta plataforma. Como usar emojis como palabras, usar abreviaciones o poner varias palabras sin espacios para utilizar menos caracteres. El modelo no procesa esta información debido a que el preprocesamiento del texto no incluyó métodos para arreglar estos mensajes. Se puede ver un ejemplo de compresión de palabras en la siguiente frase.

"reportajest13 muestran sequia... muestran dicha sequia? **robodeagua** petorca"

Fuente: Usuario Twitter

10.5 Entrenamiento de los modelos

Se implementó el modelo BERT, utilizando diferentes tareas para realizar la clasificación. Todas las aproximaciones utilizadas tienen en común que construyen una frase para realizar una next sentence prediction. La implementación de los modelos se realizó a partir del repositorio realizado por Sun, Huang, y Qiu (2019). Estos métodos se seleccionan porque son los que tienen mejor rendimiento para resolver la tarea TASBA.

Los métodos QA son aquellos que hacen la predicción sobre una pregunta. Mientras tanto, los métodos NLI hace la predicción a través de una pseudo fase. Estos métodos tienen dos variantes, el multiclase (con terminación M) mide la polaridad del sentimiento en sí a través de una única predicción que entrega todas las categorías que el modelo estima aparecen en el texto clasificado. Los métodos binarios (con terminación B) incluyen la categoría y polaridad en la primera frase del next sentence prediction. Esto provoca que su output sea binario (sí y no), siendo las técnicas binarias las más efectivas.

Estas primeras frases deben respetar el formato ASCII para que el modelo las entienda. Ejemplos de frases de preposición son:

Modelo	Ejemplo del primer texto que recibe el modelo para realizar el next sentence prediction
QA_B	la polaridad del aspecto publico es positive.
NLI_B	positive – publico
NLI_M	Publico
QA_M	que opinas de este organismo publico ?

Fuente: Elaboración propia.

Para cada una de las metodologías de clasificación fue necesario construir una base de datos, con la oración anterior correspondiente a cada método. Los modelos se ejecutaron en jupyter notebook a través de ordenadores remotos utilizando el servicio en la nube de Colab.

A continuación, se presenta la tabla con los resultados del accuracy de los modelos. Los modelos utilizan el 80% de los datos como base de entrenamiento. Este porcentaje se extrae proporcionalmente a cada categoría. El resto de los datos es utilizado como base de datos de evaluación, la cual sirvió para evaluar el accuracy de los modelos sobre un conjunto de datos no previamente visto mientras entrenaban.

Modelo pre-entrenado	Método de clasificación	Accuracy test dataset
BETO uncased	QA_B	93.65%
BETO uncased	NLI_B	93.33%
BETO uncased Multilingual	QA_B	91.64%
BETO uncased Multilingual	NLI_B	91.29%
BETO uncased	NLI_M	83.51%
BETO uncased	QA_M	82.61%
BERT uncased Multilingual	QA_M	81.00%
BERT uncased Multilingual	NLI_M	80.28%

Rendimiento de modelos en la clasificación TASBA, fuente: Elaboración propia.

BETO, como modelo, muestra más eficiencia en las tareas de TASBA que BERT multilingüe. Los modelos binarios obtuvieron un mejor desempeño que las variantes

multiclase. Este desempeño es consistente con los antecedentes bibliográficos. Aunque QA_B muestra ser el modelo con mejor desempeño, NLI_B tiene un desempeño muy similar. Esta pequeña diferencia no permite concluir que uno sea un modelo mejor que el otro.

El número reducido de variables manejadas en esta investigación podría ser la causa del mejor desempeño que tienen los modelos binarios. Una gran cantidad de variables pueden reducir el desempeño de este tipo de modelo pues la ejecución repetida de clasificaciones (una por categoría) genera más oportunidades para cometer error, situación que no ocurre en los modelos multiclase, que siempre hacen una evaluación para cada caso independientemente del número de categorías que se maneje.

La siguiente gráfica proporciona en su eje Y el f1-score por categoría del modelo en el conjunto de datos de validación, estando en el eje X de la gráfica el número de casos de cada categoría en el conjunto de datos de entrenamiento.

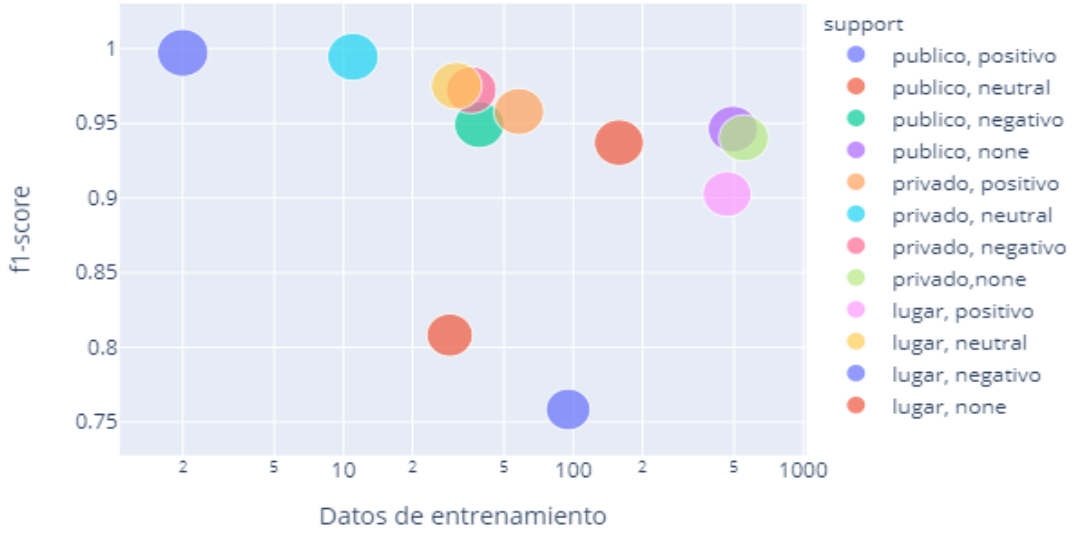

Fig. 19. F1-score del modelo QA_B dividido por categorías, elaboración propia.

Las categorías "público, positivo" y "público, neutral" son las que tienen el peor f1-score. Este bajo desempeño puede atribuirse a la similitud que poseen ambas categorías, lo cual dificulta su distinción.

El modelo muestra una tendencia a disminuir su desempeño en las categorías con más apariciones, lo que es síntoma de sobreajuste. En este tipo de escenarios conviene implementar una metodología para disminuir el sobreajuste. Sin embargo, en este conjunto de datos, de extensión corta y en donde las categorías menos comunes cuentan con muy pocos casos, no se pueden implementar. Los pocos casos que tienen las categorías menos frecuentes no poseen suficiente información para permitir que el modelo aprenda los patrones necesarios para ser reconocidos.

Llama la atención la puntuación alcanzada por las categorías "lugar, negativo" y "privado, neutral" debido al pequeño número de apariciones en el conjunto de datos de entrenamiento. Este desempeño se explica a partir de la capacidad del modelo para comprender cuándo no aparecen dichas categorías, lo cual se explica por el pequeño número de apariciones es fácil de aprender.

Los valores “none” representan cuando el valor de esa entidad no aparece. Una predicción acertada en estos casos sería detectar que no se menciona la entidad. Además, como se muestra en el gráfico, estas categorías no tienen un comportamiento diferente al resto de categorías.

A continuación, se detalla la actuación del modelo BERT QA_B, que fue el modelo que obtuvo el mejor desempeño. Viendo los resultados de forma dicotómica, si la categoría con el sentimiento aparece en el texto clasificado es etiquetado con un (1) mientras que si no aparece es etiquetado con un (0).

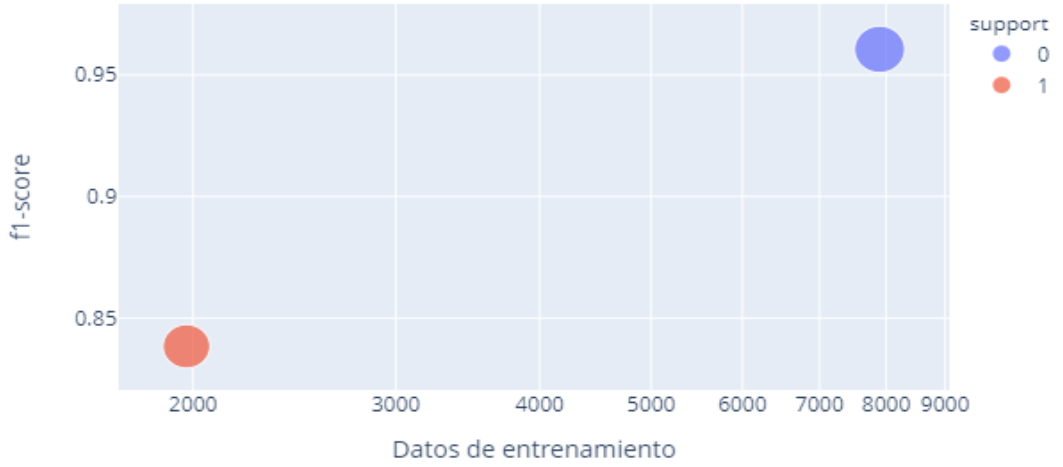

Fig. 20. F1-Score del modelo QA_B, elaboración propia.

El eje X presenta un alto número de casos, ya que cada categoría existente hace que el modelo necesite hacer más clasificaciones. Las 12 categorías se multiplican con 585 casos dando el total visto. Por tanto, el modelo necesita hacer 12 clasificaciones para cada ejemplo.

En la base de datos, la no ocurrencia de las categorías en el texto fue más común que la ocurrencia de una mención. Esta sobrerrepresentación permitió que el modelo desarrollara una mejor capacidad para detectar la no ocurrencia. La mayor prevalencia de casos en los que el usuario menciona pocas entidades en sus tuits se debe a la corta duración de cada tuit, lo que impide que los usuarios desarrollen sus ideas al escribir.

El modelo BERT con mejor desempeño ha mostrado una gran efectividad ya que, con los pocos casos que se contara para que aprendiera, ha logrado un buen desempeño. Además, la variante bilingüe ha logrado un aumento significativo en la efectividad del modelo sobre las alternativas multiclase.

10.6 Análisis discursivo

En esta sección se realizará un análisis del discurso sobre los datos utilizados en la investigación, permitiendo ampliar la interpretación de los resultados y comprender el

contexto de las menciones realizadas a las entidades. Además, otorga una instancia para poder detectar errores y encontrar nuevas oportunidades para modificar el estudio.

La mayoría de las localidades mencionadas provienen de las regiones de Coquimbo, Valparaíso y la Región Metropolitana. Este protagonismo es consistente con la población de estas regiones y la sequía que sufre esta zona durante los últimos diez años. Fueron pocas las menciones del gran norte del país, como se indica en los antecedentes, debido a que la sequía en esta zona del país tiene una larga historia y es asumida con mayor facilidad. Aunque hay conflictos mineros en curso, los tweets tienen pocas menciones. Además, esta parte del país está mucho menos habitada que la zona central, por lo que los pocos comentarios, en términos absolutos, son esperables.

Los comentarios alusivos al sur del país fueron menos frecuentes. Estos tuits critican las acciones de los forestales y también entregan argumentos con el propósito de que se reconozca a esta zona también como afectada por la sequía. En estos mensajes también se mencionaron los incendios forestales, relacionándolos con los fenómenos de sequía. Además, en los tuits mencionados aparecieron las malas prácticas de los forestales, los delitos y el acaparamiento de los recursos hídricos. Estos comentarios fueron menores que los dedicados a la zona central.

Muchos comentarios tienen contenido alusivo a la existencia de conflicto, entre los que destaca la mención de múltiples comunidades afectadas. Al mismo tiempo, hubo menos menciones respecto al Antropoceno.

En cuanto al contenido de los tuits, estos varían en torno a tres grupos: el contenido realizado por usuarios individuales de la plataforma, los medios de comunicación y las declaraciones oficiales hechos por instituciones.

Los usuarios individuales de la plataforma fueron quienes produjeron la mayor cantidad de comentarios negativos a empresas y organismos públicos. Estos tuits comúnmente atribuyen la sequía a las actividades económicas llevadas a cabo en las zonas afectadas, particularmente por la minería y la agricultura (palteras).

Además, desde los tuits generados por usuarios regulares de la plataforma había un permanente posicionamiento dentro de la narrativa del saqueo del agua. Estos comentarios

polarizados, por lo general, atribuyen el fenómeno de sequía hacía una mala gestión del recurso, emitiendo críticas para las autoridades y comentarios tipo denuncia para las grandes empresas privadas.

> "putaendo severa crisis agua sequia gobierno autoriza instalación minera zona... toma nota en: link"

Fuente: Usuario twitter

También se criticó largamente a las autoridades, el relato típico fue que éstas daban rienda suelta al actuar de las empresas. Hubo también menciones a la incapacidad técnica, como falta de capacidad para proveer agua debido a la no incorporación de tecnología, sin embargo, éstas fueron menores.

Desde los tuits los usuarios regulares expresaron una gran preocupación respecto de las localidades afectadas por la sequía. En los mensajes existió un énfasis en relatar cómo afectan a nivel económico los episodios de sequía a rubros específicos, como los apicultores y ganaderos que no pueden alimentar sus animales.
En términos generales, hubo consentimiento en los usuarios individuales de que el saqueo de agua es un fenómeno real. También hay una parte importante de usuarios que hacen la distinción y hablaban de saqueo y sequía al mismo tiempo. En otras palabras, es un fenómeno que a pesar de que haya mejoras en el uso del agua, seguirá ocurriendo.

> "si paran robo agua empresarios tiempo despues volveran misma sequia link"

Fuente: Usuario Twitter

Hubo muchos tuits que utilizaron lenguaje periodístico, los cuales provenían de cuentas de medios de comunicación. Estos tuits se caracterizaron por:

1. Comunicar la sequía y sus consecuencias en diferentes áreas sin aludir a ninguna entidad como responsable de estos fenómenos.
2. Comúnmente realizan una enumeración de las industrias y localidades afectadas.

3. Fue frecuente la mención de instituciones públicas que realizan acciones para enfrentar la sequía, como la entrega de beneficios o construcción de obras públicas, esta información era entregada con un tono informativo neutral.
4. No se mencionan alternativas o propuestas para prevenir el fenómeno de la sequía, pero sí apela al uso consciente del recurso.
5. No se pronuncian sobre la existencia de saqueo del agua o no. Este neutralismo diferencia a los medios de comunicación de los usuarios normales de Twitter que, en general, siempre toman posición sobre las acusaciones de saqueo.
6. Fueron los que hacían mayor mención del desarrollo tecnológico para enfrentar la sequía, como la modificación genética de especies y mejoras en los mecanismos de captación de agua.

"trabajando junto mesa nacional agua, integrada mop_chile, cnrchile, dga, parlamentarios, academicos expertos desafios mejorar gestion institucionalidad agua, frente sequia afecta pais. link"

Fuente: Usuario Twitter

"indap entrego aporte economico 700 pequenos ganaderos afectados sequia link link"

Fuente: Usuario Twitter

Los comentarios institucionales fueron los menos comunes y los realizaron tanto a instituciones públicas como privadas. Ambos tipos de instituciones en sus comunicaciones mencionan las tareas que realizan estas instituciones para enfrentar la sequía o la negación de responsabilidad por los fenómenos de sequía. También mencionan declaraciones de sus directivos, lanzamiento de campañas de financiamiento de proyectos para enfrentar la sequía y realizan mención de proyectos en los que participan.

En lo referente a eventos que marcaron la discusión, la COP 19 no fue un evento relevante en las comunicaciones referentes a la sequía en Chile. Quizás esta no presencia del evento en los comentarios se produjo por los meses en los que este estudio recopiló información, varios meses después que dicho evento terminase.

El evento que sí tuvo relevancia fue la lluvia invernal, que dejó un excedente de agua no visto en la zona central del país desde hace diez años. Una de las hipótesis de la investigación es que tomaría relevancia al discutir la sequía dentro de Twitter el efecto ambiental positivo que tuvo la pandemia producida por el covid-19, lo cual ocurrió, pero solo al comentar el aumento de las lluvias durante 2020 en Chile.

El superávit de agua durante el año 2020 no generó comentarios alusivos a un fin al problema de la sequía como se esperaba, por el contrario, los comentarios fueron por lo general muy cautelosos. Existió un gran consenso tanto en los comunicados institucionales, la prensa y los usuarios individuales de la plataforma de que el superávit de agua debe entenderse como un hecho circunstancial.

Este llamado de cautela ante el superávit de agua caracterizó los mensajes que se emitieron a la plataforma durante los meses de invierno. Los que más reiteraron el mensaje de cautela fueron los usuarios individuales, quienes plantearon que la gran cantidad de lluvias que hubo no implicaban un cambio de tendencia en la sequía que lleva esta zona del país por tantos años. Además, estos mismos usuarios individuales también entregaron comentarios alusivos a una disputa, referente a las interpretaciones del excedente de lluvia que hubo en Chile, comentarios defensivos respecto a la interpretación más cautelosa del superávit, pese a que los comentarios negando lo excepcional del fenómeno fueron mínimos.

Un número significativo de los tuits no se refirió a ninguna entidad estudiada, en los cuales se hacían mensajes del tipo “no es saqueo, es sequía” sin agregar más contenido.

10.6.1 Análisis de instituciones públicas

Los mensajes que mencionan instituciones públicas tendían a ser, por un lado, personalizados, de modo que muchas veces para referirse a un ministerio o alcaldía, se refieren al máximo responsable de esas instituciones. Por otro lado, instituciones más específicas de la sequía y entendidas públicamente como instituciones técnicas, como la Dirección General de Aguas (DGA) o el Instituto de Desarrollo Agropecuario (INDAP), no mencionaron a las personas. En su lugar, mencionan la institución.

Los comentarios positivos de las autoridades aludieron al trabajo realizado para combatir las necesidades derivadas de la sequía. La prensa fue la principal emisora de estos comentarios, generalmente refiriéndose a instituciones públicas y alcaldes.

Los tweets que contenían comentarios negativos aluden a lo permisivas que son las instituciones con el uso excesivo del agua, siendo catalogadas como cómplices de la generación de sequía producida por la industria. El código de aguas recibió críticas, responsabilizando a las autoridades de su existencia.

Hubo varias críticas a la actuación del gobierno de Sebastián Piñera, declaraciones que apuntaban a conflictos de intereses de los ministros para enfrentar la sequía en el país. Estos integrantes del gobierno terminan caracterizados como grandes propietarios de derechos de agua o propietarios de plantaciones de paltos. Sin embargo, no hubo menciones a proyectos gubernamentales para abordar el tema de la sequía, destacando la poca mención que tuvo la carretera hídrica, junto con pocos tuits que reflexionan sobre políticas específicas a mejorar o nuevas propuestas a implementar. La falta de apreciación de la sequía como fenómeno provocado por cambios antropogénicos puede tender a que los usuarios se centren en la denuncia de apropiación del recurso, descartando comentarios que tratan de la adaptación del país en la nueva etapa de escasez hídrica que se ha presentado.

> "fexhernandez rmunda codigo aguas modelo privado agua danado acuiferos gravemente. rios secan solo sequia sino porque, especialmente vi region norte, agua alcanza llegar mar. queda pegada termoelectricas, mineria, monocultivo, paltos."

Fuente: Usuario Twitter

Los comentarios neutrales fueron generalmente menores, en su mayoría citas de las declaraciones de las autoridades.

En resumen, el tono de denuncia de los hechos concretos por las políticas de despojo ha capturado el discurso de los usuarios. Mientras que las acciones de las autoridades locales en defensa del uso de los recursos por parte de los habitantes del espacio, tanto para las formas de producción agrícola familiar como sanitaria, recibieron el apoyo más significativo en las menciones. Los comentarios muestran un mantenimiento de la persecución del agua como recurso-mercancía, porque el uso del agua no se refiere a la destrucción del medio ambiente a partir de un relato personificado de la naturaleza.

10.6.2 Análisis de empresas privadas

En términos generales, hubo una imagen negativa hacia las empresas privadas cuando se trató el tema de la sequía. Esta crítica se dirige principalmente hacia las empresas mineras y la industria frutícola, en particular los cultivos de palta. Los cuestionamientos se centraron en el uso extensivo del agua que tienen sus actividades. Estas críticas generalmente vinculan su comentario con localidades específicas y cómo se ven afectadas.

La generación de animadversión de los usuarios hacia el sector privado proviene de las críticas al despojo atribuidas a estas instituciones, sumadas con las malas prácticas que han realizado estas empresas en los territorios donde operan. En conjunto, esta percepción negativa se debe a la percepción mayoritaria de que el problema de la sequía se debe al aumento de la demanda de recursos. Por ello, las empresas del sector minero y exportador de frutas, a través de los comentarios, terminan responsabilizadas por los efectos adversos que las actividades tuvieron en las comunidades afectadas. Estas críticas van desde las prácticas de extracción de agua ilegales y legales pero excesivas, que conducen a la imposibilidad de sustento de las comunidades afectadas, y a la represión y amenazas a los actores que presentan formas de resistencia.

Entre los comentarios a favor de las empresas se destacan los comunicados corporativos, en los que se presenta en parte evidencia científica para negar responsabilidades que tienen ante la presencia de sequía en las zonas donde operan, junto con comentarios que resaltan la activación económica que viven las zonas en es estas empresas se encuentran instaladas.

En menor medida, algunos tuits defendieron las actividades realizadas por privados. Además, la gran mayoría de estos se refieren a pequeñas empresas o autónomos. Hubo otro tipo de apoyo a las grandes empresas, que iba desde declaraciones institucionales defendiendo su labor para el desarrollo económico de las zonas afectadas hasta comentarios que informaban sobre los efectos positivos de las medidas de mitigación que estas empresas realizan.

El hecho de que la categoría “privados” no distingan entre pequeñas y grandes empresas terminó siendo problemático, ya que dentro de los comentarios la gente se preocupa por distinguir entre ellas. Es por esto que para investigaciones futuras se recomienda separar ambas categorías, generando una categoría específica para actores privados pequeños y grandes.

Las pequeñas empresas privadas suelen aparecer cuando sus actividades se ven afectadas por la sequía y por la difusión de compensaciones monetarias entregadas por el Estado ante la situación de sequía. También es en estos casos donde los internautas elogian la actuación de las autoridades.

"Seremi Agriculture entrega suplemento nutricional apicultores afectados sequía"

Fuente: Usuario Twitter

La desatención de los megaproyectos en los comentarios puede deberse a lo poco avanzados que están los planes, mientras que los proyectos actuales no están recibiendo atención, probablemente por la menor capacidad de los movimientos de resistencia para tener notoriedad pública. Es de suponer que habrá un aumento en los comentarios sobre estos proyectos si se anuncia un proyecto grande.

10.6.3 Análisis de localidades

La gran cantidad de comentarios alusivos a las localidades son de carácter positivo, mostrando tanto apoyo y preocupación por la situación que están viviendo. Estos comentarios, por lo general, refuerzan la cosmovisión de que la sequía que viven los territorios es producto del saqueo y son emitidos por usuarios individuales de la plataforma. También van acompañados de relatos que exponen la situación vivida por las localidades afectadas, sobre todo los rubros que se ven afectados por la sequía.

Prácticamente no hubo comentarios negativos dirigidos a las localidades, hubo muy pocos cuestionamientos sobre el uso del agua por parte de las personas en las áreas afectadas.

Los comentarios neutrales que se refieren a aquellos casos en los que se mencionan las localidades sin reconocer que en ellos exista episodios de sequía. Generalmente, la mención neutra de los lugares permite hacer comparaciones con otro lugar.

No hubo menciones dentro de los comentarios de conflictos locales entre vecinos, mientras que algunos mencionan el cambio del paisaje. Fueron pocas las menciones a la tierra, en las que la tierra se relaciona con otros valores, como la libertad obtenida por el propietario del campo. Si bien los usuarios de los tuits comentaron sobre las comunidades indígenas, se refieren principalmente a los comentarios de las comunidades campesinas.

11. Recomendaciones para implementar la metodología

El modelo posee un buen desempeño, puesto que logra un buen nivel de performance en una tarea de NLP complicada. Este desempeño tiene el potencial de crecer, a través de un aumento en la cantidad de datos clasificados.

Este tipo de modelo es ideal para trabajar con comentarios realizados en entornos controlados, como una encuesta, libros de quejas, situaciones en las que se establece claramente la pregunta formulada.

La implementación de este modelo en entornos no estructurados como Internet posee la ventaja de la constante generación de contenido, pero a su vez provoca una dependencia sobre otros modelos que hagan el filtrado de la información, generando varios desafíos que deben ser confrontados. Estos modelos extra deben determinar (a) la ubicación y (b) si es apropiado cómo es utilizado el concepto para los fines de la investigación. Los errores de estos modelos terminaron acumulados. El modelo BERT funciona bien en tareas individuales, pero la acumulación de modelos aumenta el desafío de crear un sistema que registre la opinión en tiempo real.

Es fundamental ser realista en este tipo de investigación. Los buenos resultados son posibles cuando se poseen suficientes ejemplos de entrenamiento, especialmente en temas poco populares, como la sequía, pues si bien los temas que reciben menos atención por parte de la sociedad tienen menos variabilidad en su discusión, lo que los convierte en temas simples de ser clasificados, se cuenta con menos matices y diversidad en la discusión imposibilitando tratar el tema con una mayor profundidad.

Los temas tratados con mayor frecuencia en las redes sociales, como las discusiones producidas por los procesos electorales, tienen más datos disponibles y poseen un debate por sí mismo más profundo. Estos datos permiten la construcción de categorías más complejas para ahondar en el tema de estudio. Mas esto también conlleva un mayor costo en producir la base de datos y mantener los modelos en el tiempo.

La creciente profundización que se debe hacer para encontrar aquellos casos útiles produce un mayor costo para la investigación. Esta profundización es necesaria para obtener suficientes casos clasificados sobre todas las categorías y encontrar suficientes casos en las categorías más específicas para ser correctamente clasificados por el modelo y que no ocurran problemas de sobreajuste.

Internet está lleno de datos de texto no estructurados, lo que significa que es difícil encontrar datos útiles. En esta investigación, por ejemplo, categorías como "lugar, negativo", "privado, neutral" y "público, neutral" poseen menos de 50 ocurrencias. Por lo tanto, la estandarización lograda en la construcción de las encuestas, donde las categorías que utilizadas para investigar una temática se encuentran estandarizadas e implementadas a lo largo de múltiples regiones del mundo es difícil de replicar para este tipo de investigaciones.

Por lo tanto, es más desafiante implementar una investigación impulsada por construcciones teóricas complejas a partir de datos obtenidos en Internet, más en un estudio longitudinal, debido al mantenimiento necesario para mantener la efectividad del modelo. Una mala implementación de las técnicas de aprendizaje automático puede generar un modelo incapaz de clasificar correctamente la mayoría de las categorías e impedir al equipo de investigación generar conclusiones valiosas. Por tanto, una alternativa que debe ser considerada es simplificar las categorías propuestas por los antecedentes, para facilitar la implementación.

A estos problemas para llevar a cabo este tipo de metodología sobre datos de internet, es necesario agregar que la evolución de la discusión genera nuevos temas abordados en el debate mientras hay temas que pierden protagonismo. Los modelos no fueron entrenados originalmente sobre estos nuevos temas, por lo que necesitan ser entrenados nuevamente con nuevos datos clasificados. Además, otro problema es la necesidad de incluir nuevas categorías en medio de la investigación. Este problema puede ocurrir debido a las oportunidades encontradas en medio de la investigación, también por adaptaciones que necesitan ser hechas sobre las categorías que maneja la investigación, debido a deficiencias

en la elección original de las categorías para abordar la base de datos. Finalmente, la evolución del lenguaje de los usuarios puede generar que la discusión adopte nuevos lemas, jerga o memes para expresar sus opiniones, situaciones en las que los modelos deben ser entrenados sobre nueva data clasificada.

Una consideración que hay que mantener sobre los datos generados en Twitter es que tienen una presencia distorsionada de determinadas categorías producida por la forma en que escriben los usuarios. Los comentarios hechos por usuarios individuales tienden a mencionar menos entidades por tuit. Son comentarios que se refieren de manera extendida a la realidad de las diferentes localidades mencionadas y cómo se enfrenta a la sequía. Este estilo de escritura es muy diferente a los tuits que hacen los medios.

Los medios de comunicación suelen mencionar muchas entidades en un mismo tuit y, por tanto, no desarrollan mucho su opinión en sus comentarios, generando una gran cantidad de menciones neutrales de las entidades que menciona. Esta característica de escritura existe debido a la dependencia de los medios de capturar tráfico, por lo que la escritura captura más entidades y sus respectivos hashtags. Por tanto, puede ser recomendable descartar aquellos comentarios que superen un máximo de menciones a entidades para mantener aquellos tuits con opiniones más sustanciales y con mayor probabilidad de ser emitidos por usuarios individuales.

Estos son los retos más críticos que se deben tener en cuenta para realizar este tipo de investigación. Además, es necesario agregar plazos de tiempo al planificar la investigación reservados para abordar los problemas no previstos que surgen en medio del estudio.

Si estos desafíos no se pueden resolver, entonces no sería recomendable implementar una investigación que utilice la metodología TASBA. Otras tareas, como la clasificación de texto, generan menor cantidad de desafíos y pueden generar resultados útiles.

12. Conclusión

Durante el período estudiado, hubo una gran cantidad de comentarios negativos hacia las instituciones públicas y privadas, junto con mensajes de apoyo a las localidades, esto en medio de una narrativa de conflicto y denuncia que es síntoma del malestar existente hacia la elite política y económica del país.

Así, se encontró que, con el aumento de las prácticas de acaparamiento de agua por parte de las empresas, hay un rechazo a la política y a los grupos empresariales pertenecientes a estas áreas. Esta menor disponibilidad de agua, que implicó un severo impacto de la producción agrícola, deterioro del medio ambiente y aumento de la conflictividad en las áreas afectadas, ha llevado a los usuarios de Twitter a enfocar sus mensajes en denunciar estas condiciones, impidiendo que traten la sequía como un fenómeno antropogénico.

Otro aspecto de los comentarios analizados es el lugar de procedencia. Además de los usuarios habituales, se detectan comentarios oficiales de instituciones y publicaciones de los medios de comunicación, es importante estar pendiente de este tipo de contenido, ya que pueden poseer una cantidad particularmente grande de menciones de entidades, que en caso de ser excesivas pueden distorsionar los resultados, o contener un lenguaje tan artificial que genere ruido al modelo. Por otro lado, las redes sociales en sí mismas se han convertido en un campo de interés político, desde el cual es posible modificar las opiniones de personas susceptibles, el uso correcto de las redes ha logrado en muchos casos ampliar la convocatoria a los movimientos sociales. Los territorios afectados por la sequía han demostrado despertar el interés y la empatía de los usuarios, por lo que existe una gran posibilidad de que aumente la relevancia de las redes sociales digitales para los movimientos en defensa del agua de Chile a lo largo del tiempo.

En la investigación se logró desarrollar un modelo clasificador que permite medir la opinión expresada sobre la sequía hacia las entidades en Twitter, extrayendo una cantidad significativa de tuits. Aun así, existe gran cantidad de información por ser extraída desde la base de datos, tweets útiles sin procesar que permitirían seguir profundizando, tanto a través de la clasificación de datos a mayor escala como con categorías más específicas. Una posibilidad de utilizar categorías más específicas se observó durante la revisión cualitativa

de los tuits, y se determinó que es importante distinguir entre grandes entidades privadas y pequeñas entidades privadas, ya que en la narrativa que la sociedad posee actualmente de la sequía se entienden como miembros de lados opuestos en el conflicto y, ambos grupos tienen una significativa cantidad de menciones en tweets, por lo que su estudio es factible.

En otro ámbito, el territorio chileno seguirá enfrentando episodios de sequía, por lo que se generarán diferentes estrategias para enfrentarla, desde la innovación en tecnologías hasta nuevos regímenes de distribución de agua. La proyección, a partir de esta investigación, es que a medida que desaparezca el despojo del agua de los territorios con escasos recursos, la sociedad se centrará menos en el aspecto conflictivo del fenómeno y más en el cambio climático global que afecta a Chile.

Las opiniones son muy cautelosas con respecto al aumento de las lluvias en el invierno de 2020, a diferencia de cómo se estimó en la hipótesis. Esto puede deberse a que la población se centra en el despojo que ha existido hacia los territorios por parte de las empresas usuarias del agua, situación que no ha tenido grandes cambios. Los cambios que ocurren a nivel legislativo y los eventos de interés político pueden hacer que las personas cambien la forma en que entienden la sequía.

Respecto a posibles investigaciones futuras, es recomendado desarrollar 1) un clasificador de ubicaciones y 2) detector de la temática de sequía que funcione con la suficiente fidelidad para clasificar los tweets en tiempo real, permitiendo ver cómo evoluciona la opinión en diferentes épocas del año, y enfrentar eventos de interés. Asimismo, se invita a los investigadores sociales a implementar esta y otras técnicas de extracción de características en sus respectivas temáticas de interés. Cada uno, desde su experiencia, podrá ver la oportunidad de explorar en profundidad diferentes temas y plataformas, las cuales no han sido suficientemente explorados desde las Ciencias Sociales por falta de información primaria. La realidad es que hay mucha información en documentos digitales para explorar y construir conocimiento que no es aprovechada en la actualidad que nos permitirá profundizar nuestro entendimiento de los fenómenos sociales a niveles que son insospechados. El aumento en el rendimiento de los modelos de NLP abre la posibilidad de explorar campos sociales digitales en volúmenes nunca vistos a través de la automatización del clasificado.

En este sentido, se recomienda estar constantemente actualizado sobre las innovaciones que existan dentro del campo de la NLP. La revolución que involucró a los mecanismos de atención se produjo en 2017; fue un fenómeno reciente. Esta tendencia generó la creación de modelos pre-entrenados, siendo estos cada vez más grandes en tamaño, permitiendo que los modelos crecieran en rendimiento constantemente. Esta mejora de desempeño puede generar a futuro cambios positivos, como lo es requerir menos datos de entrenamiento que los que se necesitan en la actualidad.

Aun así, la evolución de los modelos de NLP también puede generar tendencias negativas, como lo es la creación de modelos cerrados, cuyo acceso depende de la contratación de un servicio, como es el caso del modelo GPT-3, o la difusión de contenidos falsos, generados por los mismos modelos, que pueden alcanzar un alto nivel de fidelidad, haciéndolos indistinguibles del contenido generado por humanos, atacando tanto 1) la fuente de información de la construcción de los modelos, que el contenido generado por humanos en Internet, como también 2) desencadenar la producción de contenidos difamatorios y manipuladores, provocando que los documentos digitales pierdan su condición de fuente válida de información para estudiar la sociedad, así como produciendo contenidos sesgados que manipulan la opinión pública.

13. Referencias bibliográficas

Alammar, Jay. 2018. “Visualizing a Neural Machine Translation Model (Mechanics of Seq2seq Models with Attention).” In.

———. 2020. “Jekyll Now.”

Aldunce, Paulina, Dámare Araya, Rodolfo Sapiain, Issa Ramos, Gloria Lillo, Anahí Urquiza, and René Garreaud. 2017. “Local Perception of Drought Impacts in a Changing Climate: The Mega-Drought in Central Chile.” *Sustainability* 9 (11): 2053.

Aldunce, Paulina, Roxana Bórquez, Katy Indvik, and Gloria Lillo. 2015. “Identificación de Actores Relacionados a La Sequía En Chile.”

ALOMAR, JORGE MOLINA. 2019. “Ley de Cambio Climático Priorizará Agua Para Consumo Humano, Fija Metas Sectoriales de Mitigación Y Crea Beneficios Tributarios Para Donaciones Privadas a Proyectos Verdes.” urlhttps://www.paiscircular.cl/agenda-2030/ley-de-cambio-climatico-priorizara-agua-para-consumo-humano-fija-metas-sectoriales-de-mitigacion-y-crea-beneficios-tributarios-para-donaciones-privadas-a-proyectos-verdes/.

Anderson, Chris. 2008. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *Wired Magazine* 16 (7): 16–07.

Aral, Sinan, and Dean Eckles. 2019. “Protecting Elections from Social Media Manipulation.” *Science* 365 (6456): 858–61.

Arellano, Alberto. 2017. urlhttps://ciperchile.cl/2017/03/21/el-negocio-de-la-sequia-el-punado-de-empresas-de-camiones-aljibe-que-se-reparte-92-mil-millones/.

Arroyo, Amaris. 2020. “Sesgos En Inteligencia Artificial.” In. Escuela de Gobierno.

Baillard, Dominique. 2020. “Ideología Del Algoritmo: Cómo El Modelo Económico de Las Redes Fomenta La Polarización.” *Microsoft News*.

Black, Erin. 2020. “Meet the Man Who ‘Invented’ the #Hashtag.” *Cnbc*, April. https://www.cnbc.com/2018/04/30/chris-messina-hashtag-inventor.html.

Bolados García, Paola, Fabiola Henríquez Olguín, Cristian Ceruti Mahn, and Alejandra Sánchez Cuevas. 2018. “La Eco-Geo-Política Del Agua: Una Propuesta Desde Los Territorios En Las Luchas Por La Recuperación Del Agua En La Provincia de Petorca (Zona Central de Chile).” *Revista Rupturas* 8 (1): 159–91.

Bosse, Stefan. 2018. “Data Mining with Machine Learning for the Social Sciences.” https://doi.org/10.13140/RG.2.2.12746.67526.

Bottaro, Lorena, Alex Latta, and Marian Sola. 2014. “La Politización Del Agua En Los Conflictos Por La Megaminería: Discursos Y Resistencias En Chile Y

Argentina." *European Review of Latin American and Caribbean Studies/Revista Europea de Estudios Latinoamericanos Y Del Caribe*, 97–115.

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. "Emotion Shapes the Diffusion of Moralized Content in Social Networks." *Proceedings of the National Academy of Sciences* 114 (28): 7313–8. https://doi.org/10.1073/pnas.1618923114.

BRITZ, DENNY. 2016. "Attention and Memory in Deep Learning and Nlp." *Wildml*. http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/.

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *arXiv Preprint arXiv:2005.14165*.

Bryden, John, and Eric Silverman. 2019. "Underlying Socio-Political Processes Behind the 2016 US Election." Edited by Haroldo V. Ribeiro. *PLOS ONE* 14 (4): e0214854. https://doi.org/10.1371/journal.pone.0214854.

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. "Spanish Pre-Trained Bert Model and Evaluation Data." In *To Appear in Pml4dc at Iclr 2020*.

Chen, Chi-Chung, Chia-Lin Yang, and Hsiang-Yun Cheng. 2018. "Efficient and Robust Parallel Dnn Training Through Model Parallelism on Multi-Gpu Platform." *arXiv Preprint arXiv:1809.02839*.

Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches." *arXiv Preprint arXiv:1409.1259*.

Claeser, Daniel, Samantha Kent, and Dennis Felske. 2018. "Multilingual Named Entity Recognition on Spanish-English Code-Switched Tweets Using Support Vector Machines." In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, 132–37.

CONAF. 2011. "Plantaciones Forestales."

Corbett-Davies, Sam, and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *arXiv Preprint arXiv:1808.00023*.

Correa-Parra, Juan, José Francisco Vergara-Perucich, and Carlos Aguirre-Nuñez. 2020. "Privatización Y Desigualdad Del Agua: Coeficiente de Gini Para Los Recursos Hídricos En Chile."

CR. 2015. "La Megasequía 2010-2015: Una Lección Para El Futuro." *Centro de Resiliencia*.

DeepAI. 2020. "Hidden Layer." https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-

learning#:~:text=In%20neural%20networks%2C%20a%20hidden,inputs%20entered%20into%20the%20network.

Deisenroth, Marc Peter, A Aldo Faisal, and Cheng Soon Ong. 2020. *Mathematics for Machine Learning*. Cambridge University Press.

desarrollo, Libertad y. 2020. "AMBICIOSA Ley Marco de Cambio Climático En Un Contexto de Recesión Mundial." urlhttps://lyd.org/wp-content/uploads/2020/05/tp-1447-emisiones.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv Preprint arXiv:1810.04805*.

De-Wit, Lee, CAMERON Brick, and SANDER Van Der Linden. 2019. "Are Social Media Driving Political Polarization." *Greatergood. Berkeley. Edu*.

DGA. 2020a. "Derechos de Aprovechamiento de Aguas Registrados En Dga." Dirección General de Aguas. https://dga.mop.gob.cl/productosyservicios/derechos_historicos/Paginas/default.aspx.

———. 2020b. "Derecho de Aprovechamiento de Aguas Superficiales." Dirección General de Aguas. https://www.chileatiende.gob.cl/fichas/7109-derecho-de-aprovechamiento-de-aguas-superficiales.

Dirac, Leo. 2019. "LSTM Is Dead. Long Live Transformers!"

Dodge, Jesse, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping." *arXiv Preprint arXiv:2002.06305*.

dontloo. 2019. "What Exactly Are Keys, Queries, and Values in Attention Mechanisms?" Cross Validated. https://stats.stackexchange.com/q/424127.

EstNltk. 2020. "Named Entity Recognition." In. urlhttps://estnltk.github.io/estnltk/1.4.1/tutorials/ner.html.

Foundation, Panoptykon. 2019. "Poland: Banks Obliged to Explain Their Credit Decisions." *EDRi*.

Garreaud, René D, Juan P Boisier, Roberto Rondanelli, Aldo Montecinos, Hector H Sepúlveda, and Daniel Veloso-Aguila. 2020. "The Central Chile Mega Drought (2010–2018): A Climate Dynamics Perspective." *International Journal of Climatology* 40 (1): 421–39.

Gibson, David. 2020. "The Potential for Big Data in Social Science?"

González, Mauro E, Susana Gómez-González, Antonio Lara, René Garreaud, and Ignacio Díaz-Hormazábal. 2018. "The 2010–2015 Megadrought and Its Influence on the Fire Regime in Central and South-Central Chile." *Ecosphere* 9 (8): e02300.

González, Tomás. 2020. “Comunidades Del ñuble Obligan a Dga a Declarar Desierto Remate de Derechos de Uso Sobre Agua.” Diario UChile. https://radio.uchile.cl/2020/01/13/comunidades-del-nuble-obligan-a-dga-a-declarar-desierto-remate-de-derechos-de-uso-sobre-agua/.

Grgic-Hlaca, Nina, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. “The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making.” In *NIPS Symposium on Machine Learning and the Law*, 1:2.

Herrera, Mauricio, Cristian Candia, Diego Rivera, Douglas Aitken, Daniel Brieba, Camila Boettiger, Guillermo Donoso, and Alex Godoy-Faúndez. 2019. “Understanding Water Disputes in Chile with Text and Data Mining Tools.” *Water International* 44 (3): 302–20.

Hodges, Heather E, and Galen Stocking. 2016. “A Pipeline of Tweets: Environmental Movements’ Use of Twitter in Response to the Keystone Xl Pipeline.” *Environmental Politics* 25 (2): 223–47.

Honey, Courtenay, and Susan C Herring. 2009. “Beyond Microblogging: Conversation and Collaboration via Twitter.” In *2009 42nd Hawaii International Conference on System Sciences*, 1–10. Ieee.

Huang, Jeff, Katherine M Thornton, and Efthimis N Efthimiadis. 2010. “Conversational Tagging in Twitter.” In *Proceedings of the 21st Acm Conference on Hypertext and Hypermedia*, 173–78.

Huenchumil, Paula. 2020. “El Mapa Que Muestra Cómo La Expansión Forestal Presiona a Los Mapuche En La Provincia de Arauco.” *Interferencia*.

Jost, John T., Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A. Tucker. 2018. “How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks.” *Political Psychology* 39 (February): 85–118. https://doi.org/10.1111/pops.12478.

Jurafsky, Daniel, and James H Martin. 2008. “Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing.” *Upper Saddle River, NJ: Prentice Hall*.

Kazemnejad, Amirhossein. 2019. “Transformer Architecture: The Positional Encoding.” In.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. “OpenNMT: Open-Source Toolkit for Neural Machine Translation.” In *Proc. ACL*. https://doi.org/10.18653/v1/P17-4012.

KNIGHT, WILL. 2019. “Facebook’s Head of Ai Says the Field Will Soon ‘Hit the Wall’.” *Wildml*. https://www.wired.com/story/facebooks-ai-says-field-hit-wall/.

Krishnamurthy, Prabhakar. 2019. “Understanding Data Bias.” *Towards Data Science*. https://towardsdatascience.com/survey-d4f168791e57.

Larraín, Sara. 2006. “El Agua En Chile: Entre Los Derechos Humanos Y Las Reglas Del Mercado.” *Polis. Revista Latinoamericana*, no. 14.

LeCun, Yann. 2020. “ML Systems Are Biased When Data Is Biased.” https://twitter.com/ylecun/status/1274782757907030016.

Li, Chuan. 2020. “OpenAI’s Gpt-3 Language Model: A Technical Overview.” *Lambdalabs*. https://lambdalabs.com/blog/demystifying-gpt-3/.

Liu, Jingyun, Jackie CK Cheung, and Annie Louis. 2019. “What Comes Next? Extractive Summarization by Next-Sentence Prediction.” *arXiv Preprint arXiv:1901.03859*.

Luis Carvajal, Matías Poch, and Rodrigo Osorio. 2013. “METODOLOGÍA Para La Identificación de Localidades En Condición de Aislamiento.”

Min, Peter. 2018. “Aspect-Based Opinion Mining (Nlp with Python).” *Medium*, June. https://medium.com/@pmin91/aspect-based-opinion-mining-nlp-with-python-a53eb4752800.

Montes, Carlos. 2020. “Década 2010-2019 Cierra Como La Más Seca En La Zona Central: ¿Qué Tendría Que Pasar Para Que Se Acabe La Megasequía?”

Nadeau, David, and Satoshi Sekine. 2007. “A Survey of Named Entity Recognition and Classification.” *Lingvisticae Investigationes* 30 (1): 3–26.

Nic, Newman, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. 2018. *Reuters Institute Digital News Report 2018*. Reuters Institute for the Study of Journalism.

Ognjanovski, Gavril. 2019. “Everything You Need to Know About Neural Networks and Backpropagation — Machine Learning Easy and Fun.” *Medium*, January. https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a.

Olah, Chris, and Shan Carter. 2016. “Attention and Augmented Recurrent Neural Networks.” *Distill*. https://doi.org/10.23915/distill.00001.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. “Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries.” *Frontiers in Big Data* 2: 13.

Palacio, Arturo Sánchez. 2020. “Finetuning Bert with Tensorflow Estimators in Only a Few Lines of Code.” In.

Panez Pinto, Alexander. 2018. “Agua-Territorio En América Latina: Contribuciones a Partir Del análisis de Estudios Sobre Conflictos Hídricos En Chile.” *Revista Rupturas* 8 (1): 193–217.

paperswithcode. 2020. “Language Modelling on Penn Treebank (Word Level).” https://paperswithcode.com/sota/language-modelling-on-penn-treebank-word.

Pearce, W, K Holmberg, I Hellsten, and B Nerlich. 2013. “Climate Change on Twitter: Topics.” *Communities and Conversations About the*.

Peña-Torres, Julio, Emiliano Vargas, and Guillermo Donoso. 2019. “Remate de Derechos de Agua En Chile:?‘ Qué Pasó Posreforma Del código de Aguas de 2005?” *Estudios Públicos*, no. 155.

Radford, Jason, and Kenneth Joseph. 2020. “Theory in, Theory Out: The Uses of Social Theory in Machine Learning for Social Science.” *Frontiers in Big Data* 3 (May): 18. https://doi.org/10.3389/fdata.2020.00018.

Ribeiro, Ricardo, Fernando Batista, and Laboratório de Sistemas de Língua Falada. 2013. “Natural Language Processing for the Social Sciences.”

Rivera, Diego, Alex Godoy-Faúndez, Mario Lillo, Amaya Alvez, Verónica Delgado, Consuelo Gonzalo-Martin, Ernestina Menasalvas, Roberto Costumero, and Angel Garcia-Pedrero. 2016. “Legal Disputes as a Proxy for Regional Conflicts over Water Rights in Chile.” *Journal of Hydrology* 535: 36–45.

Ruder, Sebastian. 2018. “NLP’s Imagenet Moment Has Arrived.” *The Gradient*. https://thegradient.pub/nlp-imagenet/.

Rudin, Cynthia. 2015. “Can Machine Learning Be Useful for Social Science?” urlhttp://citiespapers.ssrc.org/can-machine-learning-be-useful-for-social-science/.

Scherman, Andrés, Arturo Arriagada, and Sebastián Valenzuela. 2014. “Student and Environmental Protests in Chile: The Role of Social Media.” *Politics* 35 (2): 151–71. https://doi.org/10.1111/1467-9256.12072.

Senado. 2020. “Comisión de Medio Ambiente Aprueba Proyecto Que Fija La Ley Marco de Cambio Climático.” https://www.senado.cl/comision-de-medio-ambiente-aprueba-proyecto-que-fija-la-ley-marco-de/senado/2020-07-09/162241.html.

Sharma, Harshit. 2017. “Identifying Traffic Signs with Deep Learning.” *Medium*. https://towardsdatascience.com/identifying-traffic-signs-with-deep-learning-5151eece09cb.

Singh, Praphul. 2020. “Multi-Head Self-Attention in Nlp.” In. oracle.

Singhal, Amit, and others. 2001. “Modern Information Retrieval: A Brief Overview.” *IEEE Data Eng. Bull.* 24 (4): 35–43.

Sukhbaatar, Sainbayar, Jason Weston, Rob Fergus, and others. 2015. “End-to-End Memory Networks.” In *Advances in Neural Information Processing Systems*, 2440–8.

Sun, Chi, Luyao Huang, and Xipeng Qiu. 2019. “Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),*

380–85. Minneapolis, Minnesota: Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1035.

TensorFlow. 2020. "Neural Machine Translation with Attention." https://www.tensorflow.org/tutorials/text/nmt_with_attention.

Tucker, Patrick. 2016. "Refugee or Terrorist? IBM Thinks Its Software Has the Answer." *Defense One* 27.

Twitter. 2020. "Rate Limiting." https://developer.twitter.com/en/docs/basics/rate-limiting#:~:text=Per%20User%20or%20Per%20Application,per%20window%20per%20access%20token.

Urrutia-Jalabert, Rocío, Mauro E González, Álvaro González-Reyes, Antonio Lara, and René Garreaud. 2018. "Climate Variability and Forest Fires in Central and South-Central Chile." *Ecosphere* 9 (4): e02171.

Valencia, Felipe Tapia. 2019. "Regulación de La Sequía En Chile: Análisis Normativo de La Declaración de Escasez." *Revista de Derecho Administrativo Económico*, no. 29: 117–38.

Vargas, Cristian A, René Garreaud, Ricardo Barra, Felipe Vásquez-Lavín, Gonzalo S Saldías, and Óscar Parra. 2020. "Análisis: El Agua de Los Ríos No Se Pierde Cuando Llega Al Mar|(CR) 2."

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 5998–6008.

Vega, Francisca de la. 2021. "Sequía: Los Desafíos Para Chile de Un Futuro Con Menos Agua." In. https://portaluchile.uchile.cl/noticias/168766/sequia-los-desafios-para-chile-de-un-futuro-con-menos-agua.

Venkatachalam, Mahendran. 2019a. "An Introduction to Attention." In.

———. 2019b. "Self Attention and Transformers." In.

Wagler, Adam, and Karen J Cannon. 2015. "Exploring Ways Social Media Data Inform Public Issues Communication: An Analysis of Twitter Conversation During the 2012-2013 Drought in Nebraska."

Weng, Lilian. 2018. "Attention? Attention!" *Lilianweng.github.io/Lil-Log*. http://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html.

Yamamoto, Masahiro, Matthew J Kushin, and Francis Dalisay. 2013. "Social Media and Mobiles as Political Mobilization Forces for Young Adults: Examining the Moderating Role of Online Political Expression in Political Participation." *New Media & Society* 17 (6): 880–98. https://doi.org/10.1177/1461444813518390.

Yardi, Sarita, and Danah Boyd. 2010. “Dynamic Debates: An Analysis of Group Polarization over Time on Twitter.” *Bulletin of Science, Technology & Society* 30 (5): 316–27.

Zhang, Lei, and Bing Liu. 2014. “Aspect and Entity Extraction for Opinion Mining.” In *Data Mining and Knowledge Discovery for Big Data*, 1–40. Springer.

14. Anexos

La sección de anexos da instrucciones para implementar el modelo. las instrucciones incluyen todas las etapas, desde la extracción de datos de Twitter, la clasificación de los datos, el preprocesamiento de la base de datos hasta el entrenamiento de los modelos.

De la sección 14.1 a 14.3, se utiliza el código proporcionado por el siguiente enlace: https://drive.google.com/drive/folders/1X9cFUKH66VX7DI6LGqMDCzIG_NzfQy9l?usp=sharing

De la sección 14.4 a 14.5, se utiliza el código proporcionado por el siguiente enlace: https://drive.google.com/drive/folders/1AYvrpGHSz8yvTcc8qpesEQJ3fLBD9x4j?usp=sharing

14.1 Construcción de base de datos de tuits

La información proporcionada por la API de Twitter viene en formato JSON, que además de traer el contenido del tuit en sí, viene con una serie de datos, como la identificación del usuario, la fecha, si es un retuit o no, el lugar de residencia declarado por el usuario, entre otras cosas.

```
import tweepy
import os
import datetime

date = datetime.datetime.now()
date = str(date.strftime("%d")) + " " + str(date.strftime("%b"))
os.mkdir("tweets final/" + date)

consumer_key = "SECRETO"
consumer_secret = "SECRETO"

access_token = "SECRETO"
access_token_secret = "SECRETO"

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)

auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=Tr
ue)
```

Fig. 21. Inicializar la comunicación con la API.

La función "import" permite la importación de módulos dentro de Python. "Tweepy" es un módulo que facilita la comunicación con la API de Twitter. El módulo "os" permite la comunicación con el sistema operativo. En este caso, se utiliza el método "mkdir" para crear una nueva carpeta donde almacenar los tuits descargados.

El módulo "datetime" genera una cadena con la fecha actual y se usa para nombrar la carpeta mencionada anteriormente, organizando el JSON por día en el que se solicitó, ese no es el mismo día en que se creó, esa información está en el JSON.

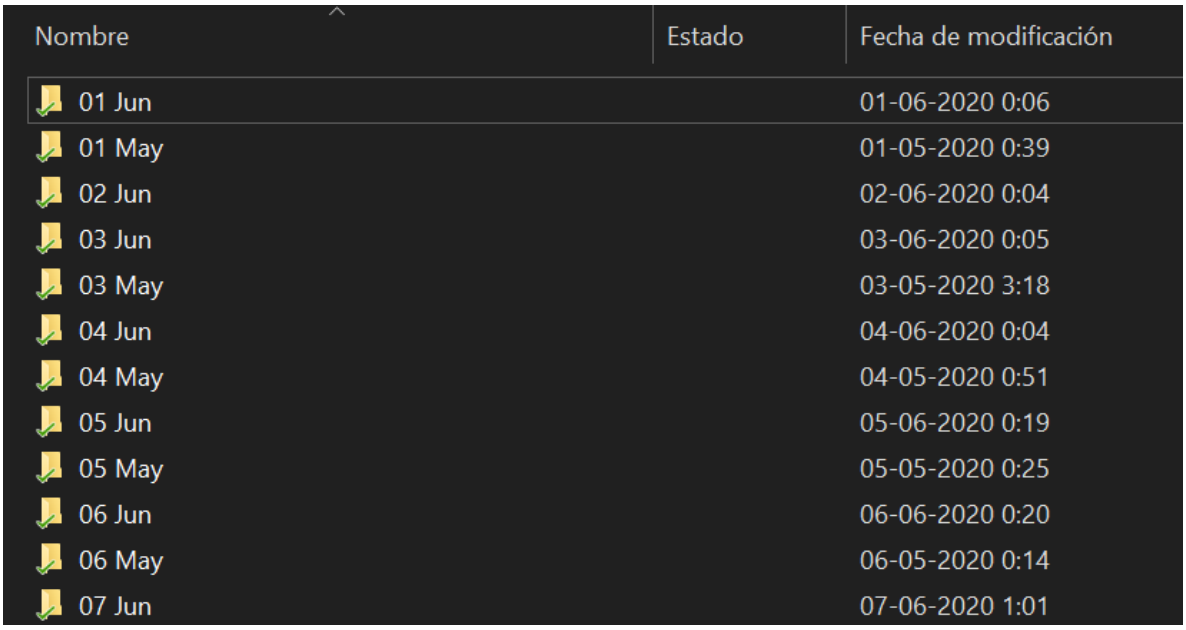

Fig. 22. Carpetas nombradas por el día en que se solicitó el JSON.

Después de esto es necesario crear la "autenticación" con la API, a través de crear una instancia de la clase "tweepy.AuthHandler class", esto ofrece autenticación "OAuth" con la API de Twitter, que tiene tanto el (1) access token como él (2) consumer information como protocolos de seguridad, ambas claves deberán ser ingresadas para ingresar a la aplicación de desarrolladores.

Primero se crea una instancia, con tanto el usuario de acceso y las claves, y a través del método "set_access_token" se incorporan los tokens que Twitter nos entrega al crear una cuenta para desarrolladores.

Finalmente, con el método "tweetpy.api" se genera autenticación. Los parámetros que comienzan con “wait” dentro de la función ayudan a mantener la conexión a la API una vez alcanzados el límite de tuit que permite obtener la API en lapsos de 15 minutos.

```
numero = 0
import json
for tweet in tweepy.Cursor(api.search , q = "sequía", tweet_mode = "extended
").items(5000):
    numero += 1
    numero_2 = str(numero)
    nombre = "tweets final/"+ date + "/" + "tweet numero "+ numero_2 + ".txt
"
    j = open(nombre, "w", encoding="utf-8")
    j.write(json.dumps(tweet._json, ensure_ascii=False))
    j.close()
```

Fig. 23. Extracción de tuits.

La herramienta principal utilizada para administrar los datos que entrega Twitter es el módulo "JSON". En esta situación, los datos deben cambiar para ser entendidos por el lenguaje python, y este módulo permite la conversión de JSON en un diccionario python.

El bucle "for" debe guarda cada tuit. Como los tuits se recibían a pedido con la función "tweetpy.Cursor", se almacenan en un archivo con un nombre. Para comenzar la búsqueda se usa el argumento "api.search", para especifica el término buscado con el argumento "q", a la vez que se indicando la versión extendida del tuit "280 caracteres" a través del argumento "tweet_mode".

La variable "nombre" ofrece la posibilidad de introducir la ruta y el nombre del archivo en el que se almacenará el JSON.

La penúltima línea escribe los datos en el documento, "json.dumps" convierte el objeto de Python al formato JSON y el "sure.ascii = False" hace que los caracteres en el formato UTF-8 estén intactos. Finalmente, el documento escrito es cerrado.

14.2 Clasificación manual de los tuits

Para realizar la clasificación manual de la base de datos, ejecute el archivo "revisión de JSON.py". Una vez funcionando, inicializará un bucle sobre las carpetas descartando las ya clasificadas. Si la carpeta no está clasificada todavía, el código abrirá los tuits, uno por uno.

Primero preguntando "¿trata la sequía? Si (s) o No (n)", luego, si la respuesta fue sí, preguntará por las entidades con las siguientes dos instrucciones: "Objetivos: Organismos públicos (p), Empresas privadas (e), Lugares (l) o ninguno más (x) " y " categoría: Positivo (p), Negativo (n), Neutral o Ecuánime (e) o Indeterminado (i) ". Estas dos categorías se ponen primero, y luego se pega la entidad, todo esto sin agregar espacios en blanco.

El código permite agregar las múltiples entidades mencionadas en los tuits. Las múltiples entidades entran en la línea de comando en diferentes líneas. Luego, cuando no queden más entidades para clasificar, se pone el carácter (x).

Finalmente se clasifica la ubicación declarada del usuario que realizó el tuit, con la pregunta en la línea de comando "La ubicación está dentro de Chile, si (s), no (n) o es indeterminable (i)", luego de agregar el personaje con la respuesta, haga clic en enter y aparecerá un nuevo case de tuit por clasificar.

14.3 Clasificación asistida

Muchos de los tuits vendrán de una ubicación que no es la investigada, también se puede mencionar el concepto analizado por la investigación, pero con otro significado. Cuando esto sucede, debemos clasificar estos casos como se expuso anteriormente, pero podemos obtener ayuda en este proceso a través de un sistema de clasificación asistida. Para eso se crean dos modelos, uno puede determinar si la ubicación es apropiada al rango de estudio de la investigación, dicho modelo se encuentra en un archivo llamado "clasificador location.py". Otro script llamado "clasificador trata sequia o no.py" determina el uso del concepto, para concluir si es relevante para la investigación. Solo es necesario ejecutar estos scripts para tener la asistencia, esto se debe hacer cada vez que aumenta el número de datos clasificados. De esa manera, los modelos tendrán los datos recientemente clasificados para entrenar, y harán mejor su asistencia de clasificación.

14.4 Preparar la base de datos

El primer paso para pre-procesar los datos está en el archivo "generar_xml.py". Para crear los conjuntos de datos de entrenamiento y testeo, el código divide los datos clasificados,

dando un 20% a la base de datos de prueba (un parámetro que se puede cambiar), también extrae los tokens de los stop words para evitar ruido en el proceso de entrenamiento del modelo. El código reemplaza los enlaces presentes en los tuits con un token llamado "link". Finalmente, todos los caracteres se convertirán a su versión ASCII, pues el modelo admite este tipo de caracteres. Este código crea una versión XML de los conjuntos de datos, uno llamado "Output_Train.xml" y el otro llamado "Output_Test.xml". Ahora necesitamos crear un conjunto de datos específico para lidiar con cada metodología probada en el documento, usando los scripts "generate/generate_semeval_QA_M.py" y "generate/generate_semeval_NLI_M.py", esto genera los conjuntos de datos, pero necesita una corrección hecha con el script. "reemplazar_categorias.py". Finalmente, se ejecuta el siguiente código "generate/generate_semeval_NLI_B_QA_B.py" y se crearán los cuatro conjuntos de datos utilizados.

14.5 Entrenamiento del modelo

Primero, necesitamos descargar el modelo en su versión de PyTorch. Para ejecutar el modelo, necesitamos usar el siguiente script usando los siguientes argumentos (CUDA_VISIBLE_DEVICES = 0,1,2,3 python3 /content/drive/MyDrive/beto/run_classifier_TABSA.py --task_name semeval_NLI_M --data_dir data / semeval2014 / bert-pair / --vocab_file uncased_L-12_H-768_A-12 / vocab.txt --bert_config_file uncased_L-12_H-768_A-12 / bert_config.json --init_checkpoint uncased_L-12_H-768_A-12model / eval_test --do_lower_case --max_seq_length 512 --train_batch_size 12 --learning_rate 2e-5 --num_train_epochs 4.0 --output_dir results / semeval / NLI_M --seed 42).

Se agrega la tarea seleccionada después del argumento task_name y se pone el nombre de la tarea al final del argumento output_dir. Ajuste el tamaño del lote si el modelo no se puede ejecutar. Un lote grande puede contener demasiada información para ser manejada por la computadora, siempre considerando que un tamaño de lote pequeño puede hacer que el proceso de entrenamiento se vuelva innecesariamente lento.

La información relacionada con el desempeño de cada época aparecerá impresa, describiendo el accuracy del modelo sobre el conjunto de datos de prueba.