UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

TRANSFER LEARNING FOR THE MULTILINGUAL AND MULTI-DOMAIN CLASSIFICATION OF MESSAGES RELATING TO CRISES

PROPUESTA DE TESIS PARA OPTAR AL TÍTULO DE MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

CINTHIA MABEL SÁNCHEZ MACÍAS

PROFESOR GUÍA:
BÁRBARA POBLETE LABRA

CO-GUÍA:
HERNÁN SARMIENTO ALBORNOZ

SANTIAGO DE CHILE
ABRIL 2020

1. Introducción

1.1. Contexto y Motivación

Con la aparición de las plataformas sociales surgieron nuevas posibilidades de alcance y comunicación (Halpern & Crichigno, 2018). De forma inmediata, los ciudadanos pueden publicar sus opiniones directamente en la web, crear y difundir contenido en diferentes formatos, e interactuar con otros usuarios sin importar las distancias geográficas.

De acuerdo a las estadísticas de Hootsuite & We Are Social, en enero del 2019 el número de usuarios de Internet fue 4.38 mil millones, lo que representaba más de la mitad de la población mundial (57%), con un crecimiento promedio superior al millón de nuevos usuarios por día. Según la misma fuente, las personas se conectan a Internet en promedio 6 horas y 42 minutos cada día. Aproximadamente un tercio de este tiempo es ocupado en plataformas sociales, entre ellas Facebook, Instagram y Twitter.

Las plataformas sociales permiten compartir y discutir temas que pueden estar relacionados con eventos del mundo real (Troudi et al., 2018), por ejemplo, deporte, política, accidentes, entre otros. Han llegado a convertirse en la principal fuente de información durante crisis, especialmente en lo que respecta a solicitudes de rescate y socorro (Khare et al., 2018).

De acuerdo a la Real Academia Española[1], una crisis es una situación mala o difícil. Se caracteriza por ser única, peligrosa, problemática, dañina, inesperada y si es descuidada o mal manejada, daría lugar a un desastre (Al-Dahash et al., 2016). Un desastre altera las condiciones normales y el funcionamiento de una comunidad o sociedad y causa pérdidas humanas, materiales, económicas o ambientales que superan las propias capacidades y recursos para afrontar la situación (Federación Internacional de Sociedades de la Cruz Roja y de la Media Luna Roja).

Los desastres tienen causas naturales (por ejemplo, terremotos, inundaciones y tsunamis), artificiales o inducidos por el ser humano (por ejemplo, derrames químicos, incendios y explosiones) y complejas o mixtas (combina cualquiera de las anteriores) (Palmer, 2017). Los desastres naturales afectan a los países independientemente de su ingreso; sin embargo, la gravedad del impacto está relacionada al nivel de ingreso y desarrollo (Centre for Research on the Epidemiology of Disasters & The UN Office for Disaster Risk Reduction, 2016). Los países más pobres son los más afectados por las catástrofes naturales, según las pérdidas humanas (Rentschler, 2013).

A las organizaciones encargadas de la gestión de desastres y asistencia humanitaria les resulta crucial conocer oportunamente cuál es la realidad durante un evento de emergencia.

[1] https://dle.rae.es/crisis

Este conocimiento les permite ser capaces de actuar y reducir el impacto en las personas afectadas (Graf et al., 2018). Las técnicas de recopilación de información (encuestas, entrevistas, observación, entre otras) han evolucionado, incorporando tecnología y técnicas de monitoreo remoto, permitiéndo superar algunas de las limitaciones de los métodos convencionales (Jansury et al., 2015).

En este punto, las redes sociales son una valiosa fuente de información que los evaluadores pueden usar para comprender la situación en el terreno. A diferencia de los medios tradicionales de comunicación, estas permiten la rápida difusión de información crítica, ya que los usuarios pueden aportar información desde el lugar de los hechos, de forma inmediata. No obstante, en este escenario la precisión y confiabilidad podrían verse afectadas por la presencia de información incompleta, incorrecta o falsa. En contraste de ambos aspectos, Castillo (2016) menciona que al inicio de un esfuerzo de respuesta de emergencia, existe un mayor riesgo de ignorar la información de las redes sociales, incluso si esta no se ha validado por completo.

Con el objetivo de facilitar y hacer más eficiente la recopilación de información durante estos eventos, han surgido nuevas investigaciones y herramientas basadas en la información publicada en redes sociales. Algunas, además de enfocarse en la detección de mensajes relacionados a un evento, buscan filtrar aquellos que aporten información útil. Gran parte de estos enfoques consideran un caso de estudio particular, un tipo de evento o un lenguaje específico. En lo que respecta al lenguaje, la mayoría se enfocan en publicaciones en inglés, lo cual restringe su uso en zonas geográficas con otra lengua.

Un análisis de un período de 20 años (1996-2015) reveló que, en promedio, los países de bajos ingresos sufren casi cinco veces más muertes por evento de desastre que los países de altos ingresos (Centre for Research on the Epidemiology of Disasters & The UN Office for Disaster Risk Reduction, 2016). Dentro de este período, los cinco países con mayor número de muertes por desastre por cada 100,000 habitantes son Haití, Birmania, Somalia, Honduras y Sri Lanka, mismos que corresponden a países de ingreso medio bajo y bajo, todos con un idioma oficial distinto al inglés.

Considerando que las herramientas existentes de apoyo en la gestión de desastres y asistencia humanitaria no son adaptables a muchas de las zonas afectadas, resulta muy importante abordar este problema desde un enfoque multilingüe.

1.2. Problema

Las publicaciones realizadas en redes sociales además de ser cortas e imprecisas (Troudi et al., 2018) suelen contener un lenguaje informal, abreviaturas, errores ortográficos, construcciones multilingües, etc. (Ghosh et al., 2018), lo cual agrega ruido a los datos dificultando la extracción de información útil. Estas son algunas de las características de las publicaciones en redes sociales que vuelven desafiante su efectiva utilización en la gestión de desastres y ayuda humanitaria.

Entrenar modelos individuales para todos los posibles tipos de eventos de crisis y lenguajes no es factible por el tiempo y gastos que esto demanda (Khare et al., 2018). Uno de los principales inconvenientes al entrenar estos modelos por separado es la dificultad de adquirir suficientes datos, lo cual dificulta el aprendizaje supervisado ya que se requiere de muchas instancias (ejemplos).

A pesar del potencial que tienen las redes sociales para aportar información útil durante eventos de crisis, existe una evidente necesidad de hacerlo de forma eficiente. Para esto, es necesario eliminar el ruido y priorizar la información a monitorear. Existen investigaciones que abordan este problema; sin embargo, gran parte de ellas se enfocan en un lenguaje específico (generalmente inglés) o en un tipo de evento particular (por ejemplo, terremotos o inundaciones), lo cual limita su aplicabilidad en un escenario real.

Dado este contexto, surge la necesidad de estudiar si los mensajes publicados en redes sociales, de distintos idiomas y tipos de evento, comparten características o patrones que permitan la transferencia de conocimiento, en la tarea de detección automática de publicaciones relacionadas a crisis.

2. Estado del arte

Palen & Anderson (2016) definen *Crisis Informatics* como un campo interdisciplinario que combina computación y conocimiento de la ciencia social de desastres. Además, Palen (2008) la visualiza como un área de investigación que examina los aspectos técnicos, sociales y de información relacionada a desastres y crisis.

En este contexto, Twitter ha sido ampliamente utilizado como objeto de estudio en diversas tareas. Una de estas tareas es la clasificación binaria de tweets como informativos y no informativos, o relacionados y no relacionados a un evento de emergencia. También en la tarea de clasificación multiclase como el tipo de desastre, tipo de información humanitaria (por ejemplo, información sobre personas heridas o muertas, precaución, donaciones), entre otras.

Existen repositorios de datos de crisis conformados principalmente por publicaciones en Twitter, disponibles para realizar investigación, por ejemplo CrisisNLP[2] y CrisisLex.org[3]. Estos abarcan diversos eventos de emergencia de distintas zonas geográficas, con una notable predominancia de publicaciones en inglés frente a una poca presencia de publicaciones en lenguajes como español, italiano, francés y portugués. Dichos tweets suelen ser recuperados mediante palabras clave relacionadas al evento y posteriormente etiquetados por humanos para diversas tareas de clasificación.

Considerando la escasez de datos etiquetados en ciertos lenguajes y tipos de desastres, Imran et al. (2016) mostraron que la adaptación de dominio (entrenando y evaluando el

[2] https://crisisnlp.qcri.org/
[3] https://crisislex.org/data-collections.html

modelo en diferentes tipos de crisis) puede ser útil cuando los dos lenguajes son similares (por ejemplo, español e italiano) pero no resulta útil cuando son distintos (por ejemplo, italiano e inglés). Para esto utilizaron caraterísticas textuales (unigramas y bigramas de palabras) y el clasificador Random Forest.

Traducir el contenido a un lenguaje común y luego aplicar técnicas tradicionales de Procesamiento de Lenguaje Natural (Natural Language Processing - NLP) es una alternativa para trabajar con contenido multilingüe. Este enfoque ha sido utilizado en distintas tareas, por ejemplo, detección de polaridad (Denecke, 2008; Demirtas & Pechenizkiy, 2013; Demirtas, 2013), reconocimiento de entidades nombradas (Dandapat & Way, 2016) y clasificación de tweets relacionados y no relacionados a crisis (Khare et al., 2018). No obstante, esta alternativa depende mucho de la precisión de la traducción; es decir, depende de una tarea que aún tiene ciertas deficiencias.

Con el propósito de encontrar patrones distintivos entre tweets informativos y no informativos, Graf et al. (2018) presentan un análisis del impacto de las características predictivas a lo largo de cuatro dimensiones, como son temporal, espacial, lingüística y origen. Utilizando SVM con kernel RBF y un conjunto de entrenamiento pequeño (aproximadamente 1,000 instancias) la clasificación de tweets pertenecientes al mismo dominio de entrenamiento obtuvo en promedio 75% de accuracy mientras que la clasificación multi dominio (seleccionando aleatoriamente las muestras y usando la estrategia de evaluación *Leave One Out*) obtuvo en promedio 79% de accuracy, sugiriendo que el clasificador entrenado con datos de diferentes tipos de desastres presenta un mejor desempeño que uno de dominio específico. Luego, para analizar el impacto del tamaño del conjunto de entrenamiento, se evaluó la clasificación multi dominio, esta vez utilizando un conjunto más grande (aproximadamente 28,000 instancias) obteniendo en promedio un 80% de accuracy, con lo cual se deduce que el aumento del número de instancias mejora ligeramente los resultados, pero no de forma considerable.

Por otra parte, Khare et al. (2018) expandieron la semántica del texto usando BabelNet y DBpedia en la clasificación de nuevos eventos o eventos en diferentes idiomas (inglés, italiano y español). En la clasificación monolingüe, las características semánticas no mejoraron los resultados, ya que el mejor modelo con características semánticas obtuvo igual F1-score (77.4%) que el modelo base, el cual usa caracteríticas estadísticas del texto. En escenarios donde el modelo fue entrenado en un idioma y testeado en otro, lograron incrementar el F1-score de 55.7% (modelo base) a 59.9% (características estadísticas + BabelNet + DBpedia) en promedio, lo cual, aún es inferior al modelo que evalúa tweets traducidos al mismo idioma de entrenamiento (63.3% con características estadísticas + BabelNet). Acorde a los autores, este enfoque sigue careciendo de adaptabilidad a nuevos tipos de datos y aún no es muy claro si la ganancia lograda es estadísticamente significativa.

Algunos enfoques utilizan *word embeddings* (Ghosh et al., 2018; Nalluru et al., 2019; Li et al., 2018), que según la definición utilizada por Yang et al. (2017), son representaciones vectoriales de palabras, aprendidas desde un corpus de texto. Estas representaciones capturan la relación entre palabras y pueden ser agrupadas semánticamente usando métricas de similitud.

En este contexto, Li et al. (2018) evaluaron diferentes técnicas de vectorización de palabras (GloVe, FastText y Word2Vec) usando modelos pre-entrenados en grandes corpus independi-

entes del dominio y entrenando modelos en un corpus de tweets de crisis. Además evaluaron el impacto de diferentes maneras de combinar los vectores de palabras para representar tweets (Mean, MinMaxMean y Tf-idf-Mean), así como diferentes algoritmos de clasificación supervisada (Gaussian Naive Bayes - GNB, Random Forest - RF, K Nearest Neighbors - KNN y Support Vector Machines - SVM). Los datos utilizados corresponden a eventos en inglés de diferentes tipos de desastres (CrisisLexT6, CrisisLexT26 y 2CTweets). La combinación que obtuvo los mejores resultados fue SVM (seguido de RF), con GloVe (pre-entrenado con tweets) y la agregación MinMaxMean.

Según el estado del arte, dentro de los algoritmos tradicionales de clasificación destacan Support Vector Machines (SVM) y Random Forest (RF), mientras que dentro del aprendizaje profundo destacan las Redes Neuronales Convolucionales (CNN). Los desarrolladores de CrisisDPS[4], una plataforma que ofrece servicios automáticos de procesamiento de datos de crisis (Firoj et al., 2019), realizaron un estudio comparativo entre los algoritmos de aprendizaje más populares, en el cual obtuvieron que SVM y RF proporcionan resultados muy competitivos, mientras que las CNN los superan ligeramente.

3. Preguntas de investigación

Partiendo del supuesto que existe suficiente información dentro de un dominio e idioma para adaptar a nuevos eventos la detección de publicaciones relevantes a crisis, se plantea:

- P1: ¿Es posible transferir conocimiento desde un dominio (o tipo de desastre, por ejemplo, terremoto) para luego clasificar datos de otro dominio (por ejemplo, inundaciones)?
- P2: ¿Es posible transferir conocimiento desde un idioma (por ejemplo, inglés) para luego clasificar datos de otro idioma (por ejemplo, español)?
- P3: ¿Cuál es la manera más eficiente de realizar la transferencia de conocimiento?
- P4: De ser posible la transferencia de conocimiento: ¿En qué idiomas y dominios funciona mejor?

4. Hipótesis

- H1: Existen patrones transversales (o independientes) al dominio, lo cual posibilita la generalización de un dominio a otro.
- H2: Existen patrones transversales (o independientes) al idioma, lo cual posibilita la generalización de un idioma a otro.

[4] https://crisisdps.qcri.org/

5. Objetivos

5.1. Objetivo general

Estudiar si es posible realizar transferencia de conocimiento a partir de datos de redes sociales de un dominio/idioma para identificar otros datos, aplicado en la detección de publicaciones relacionadas a crisis.

5.2. Objetivos específicos

1. Crear un conjunto de datos que contenga publicaciones relacionadas a crisis de diferentes tipos e idiomas, así como publicaciones no relacionadas a crisis.
2. Realizar una caracterización de las publicaciones relacionadas a crisis, con la finalidad de encontrar patrones independientes al dominio.
3. Crear diferentes representaciones vectoriales de las publicaciones relacionadas a crisis, independientes al idioma.
4. Evaluar la efectividad de las diferentes representaciones vectoriales y demás características de las publicaciones relacionadas a crisis, en la tarea de detección automática de publicaciones relacionadas a crisis.

6. Metodología

De forma general, se seguirán las siguientes actividades principales para conseguir el objetivo de investigación planteado.

6.1. Selección de los datos

Se ocuparán datos de Twitter de dos importantes repositorios de acceso público llamados CrisisNLP y CrisisLex.org. Considerando la disponibilidad de datos, se pretende trabajar con los idiomas inglés, español e italiano, y validar las representaciones con eventos de diferentes tipos (por ejemplo, terremoto e inundación).

Estas colecciones contienen tweets publicados durante una serie de eventos de crisis de tipo natural, inducido por el ser humano y mixto, producidos en diferentes países. Contiene una serie de atributos como la categoría y tipo de peligro, etiqueta que indica si el tweet está o no relacionado al evento, entre otros.

6.2. Extracción de características independientes al dominio

Con la finalidad de encontrar patrones comunes y distintivos para la clase objetivo, se tomará como referencia el estudio realizado en Graf et al. (2018), agregando nuevas características a la dimensión lingüística. Este análisis se realizará para todos los eventos e idiomas a estudiar.

Algunas de las características dentro de la dimensión lingüística (basada en el contenido del texto) podrían ser el tamaño de la publicación, número de hashtags y menciones, cantidad de sustantivos y nombres propios, etc.

Luego de haber seleccionado las características candidatas, se identificarán cuáles permiten separar los tweets relacionados versus los no relacionados a eventos de crisis. Se crearán gráficos para facilitar la interpretación de los patrones dentro de cada evento.

6.3. Modelamiento de mensajes, independiente del idioma

Trabajar con un enfoque multi lenguaje basado principalmente en el texto de las publicaciones es un gran desafío, pues hay idiomas que difieren más con unos que con otros. Incluso, publicaciones en un mismo idioma pueden diferir según el origen (por ejemplo, el país).

Para encontrar características que permitan transferir conocimiento de un lenguaje a otro, se aplicarán técnicas de vectorización populares en el área (GloVe, FastText y Word2Vec). Además, se pretende estudiar el desempeño de otro tipo de arquitecturas más sofisticadas como BERT (Bidirectional Encoder Representations from Transformers) y LASER (Language-Agnostic SEntence Representations) para modelar a nivel de oración las publicaciones en los distintos idiomas, debido a su gran desempeño en otras tareas de clasificación.

Se utilizarán modelos pre-entrenados en datos de dominio general y ajustándolos al dominio. Se ocuparán modelos universales (multi lenguaje), que en un mismo espacio vectorial permiten representar la semántica del texto de diferentes idiomas.

6.4. Diseño de experimentos

Para validar el desempeño de la clasificación de nuevos tweets utilizando el enfoque propuesto (multi lenguaje y multi dominio) frente a modelos entrenados en un dominio o lenguaje específico, se diseñarán varios experimentos, como los indicados a continuación:

1. **En dominio y en idioma:** Los datos de entrenamiento y validación corresponden al mismo tipo de desastre e idioma.
2. **Fuera del dominio y en idioma:** Los datos de entrenamiento y validación pertenecen a diferentes tipos de desastres y un mismo idioma.
3. **En dominio y fuera del idioma:** Los datos de entrenamiento y validación corresponden al mismo tipo de desastre y diferentes idiomas.
4. **Fuera del dominio e idioma:** Los datos de entrenamiento y validación corresponden a diferentes tipos de desastres y diferentes idiomas.

6.5. Evaluación

La colección de datos será dividida en datos de entrenamiento y validación, procurando respetar el orden cronológico de ocurrencia de los eventos. Se considerará una proporción 80:20 para el entrenamiento y validación, respectivamente.

Se realizarán diversos estudios de caso, donde se evaluará la calidad de las características empleando los algoritmos de clasificación más utilizados en este tipo de tareas, como son SVM, RF y CNN (con una arquitectura diseñada para trabajar con tweets, texto de longitud corta).

Se aplicará la técnica de validación cruzada (conservando la secuencia de los datos) para obtener métricas de evaluación más confiables. Las métricas a utilizar son: precisión, recall, F1-score y accuracy.

7. Aporte de la tesis

1. Estudio comparativo de las características (similares y distintivas) de las publicaciones relacionadas y las no relacionadas a una crisis, independientes al dominio.
2. Metodología para el modelado y clasificación de tweets relevantes a situaciones de crisis en diferentes idiomas.

Bibliografía

Al-Dahash, H., Thayaparan, M., & Kulatunga, U. (2016). Understanding the terminologies : disaster, crisis and emergency. In *Association of researchers in construction management (arcom)* (pp. 1191–1200).

Castillo, C. (2016). *Big crisis data*. Cambridge University Press.

Centre for Research on the Epidemiology of Disasters, & The UN Office for Disaster Risk Reduction. (2016, 10). *Poverty death: Disaster mortality, 1996-2015* (Tech. Rep.). Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/CRED_Disaster_Mortality.pdf

Dandapat, S., & Way, A. (2016, 09). Improved named entity recognition using machine translation-based cross-lingual information. *Computacion y Sistemas*, *20*, 495-504. doi: 10.13053/CyS-20-3-2468

Demirtas, E. (2013). *Cross-lingual sentiment analysis with machine translation.* Retrieved from https://pure.tue.nl/ws/portalfiles/portal/46951131/761617-1.pdf

Demirtas, E., & Pechenizkiy, M. (2013). Cross-lingual polarity detection with machine translation. In *Proceedings of the second international workshop on issues of sentiment discovery and opinion mining.* Association for Computing Machinery. doi: 10.1145/2502069.2502078

Denecke, K. (2008, April). Using sentiwordnet for multilingual sentiment analysis. In *2008 ieee 24th international conference on data engineering workshop* (p. 507-512). doi: 10.1109/ICDEW.2008.4498370

Federación Internacional de Sociedades de la Cruz Roja y de la Media Luna Roja. (n.d.). *¿qué es un desastre?* (https://www.ifrc.org/es/introduccion/disaster-management/sobre-desastres/que-es-un-desastre/ [Accessed: 07.03.2020])

Firoj, A., Imran, M., & Ofli, F. (2019). Crisisdps: Crisis data processing services. In *In proceedings of the 16th international conference on information systems for crisis response and management (iscram).*

Ghosh, S., Ghosh, K., Ganguly, D., Chakraborty, T., Jones, G., Moens, M.-F., & Imran, M. (2018). Exploitation of social media for emergency relief and preparedness: Recent research and trends. *Information Systems Frontiers*, *20*(5), 901–907.

Graf, D., Retschitzegger, W., Schwinger, W., Pröll, B., & Kapsammer, E. (2018). Cross-domain informativeness classification for disaster situations. In *Proceedings of the 10th international conference on management of digital ecosystems* (pp. 183–190).

Halpern, D., & Crichigno, B. (2018). *Guía de uso de rrss en desastres de origen natural.* Facultad de Comunicaciones Pontificia Universidad Católica de Chile, Alameda: TrenDigital.

Hootsuite, & We Are Social. (2019). Digital 2019 Global Digital Overview. (https://datareportal.com/reports/digital-2019-global-digital-overview [Accessed: 06.09.2019])

Imran, M., Mitra, P., & Srivastava, J. (2016). Cross-language domain adaptation for classifying crisis-related short messages. In *13th proceedings of the international conference on information systems for crisis response and management, rio de janeiro, brasil, may 22-25, 2016.*

Jansury, L., Moore, J., Peña, J., & Price, A. (2015, 05). *Findings in monitoring and evaluations practices during humanitarian emergencies* (Tech. Rep.). George Washington University. Retrieved from https://europa.eu/capacity4dev/file/93219/download?token=DKj_R6Ht

Khare, P., Burel, G., Maynard, D., & Alani, H. (2018). Cross-lingual classification of crisis data. In *The semantic web – iswc 2018* (pp. 617–633). Cham: Springer International Publishing.

Li, H., Caragea, D., Li, X., & Caragea, C. (2018). Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks. *en. In: New Zealand*, 13.

Nalluru, G., Pandey, R., & Purohit, H. (2019). Classifying relevant social media posts during disasters using ensemble of domain-agnostic and domain-specific word embeddings. *2019 AAAI Fall Symposium: AI for Social Good Program.*

Palen, L. (2008). Online social media in crisis events. *Educause quarterly*, *31*(3), 76–78.

Palen, L., & Anderson, K. (2016). Crisis informatics—new data for extraordinary times. *Science*, *353*(6296), 224–225. Retrieved from https://science.sciencemag.org/content/353/6296/224 doi: 10.1126/science.aag2579

Palmer, J. (2017). *What causes a disaster?* (https://gobgr.org/what-causes-a-disaster/ [Accessed: 07.03.2020])

Rentschler, J. (2013). *Why resilience matters-the poverty impacts of disasters* (Policy Research Working Paper Series No. 6699). The World Bank. Retrieved from https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-6699

Troudi, A., Zayani, C., Jamoussi, S., & Amor, I. (2018). A new mashup based method for event detection from social media. *Information Systems Frontiers*, *20*(5), 981–992.

Yang, X., McCreadie, R., Macdonald, C., & Ounis, I. (2017). Transfer learning for multi-language twitter election classification. In *Proceedings of the 2017 ieee/acm international conference on advances in social networks analysis and mining 2017* (pp. 341–348).