Taller de Tesis: Elevator Spiel v2

Profesora: Amelia Gibbons
Alumno: Alfonso Sundblad

La temática original de mi trabajo era Economic Uncertainty and Mental Health. Sin embargo rotó a “The Linguistic Style of Economic Uncertainty. Esto me permite analizar el lenguaje de las personas en un formato “continuo” en vez de un clasificador mas bien concreto, y de forma mas detallada. En resumen me centrare en estudiar el estilo lingüístico asociado a la incertidumbre económica en el idioma español. La pregunta de investigación principal que guía mi trabajo es: "¿Podemos identificar el estilo lingüístico de la incertidumbre económica en el idioma español?" Para abordar esta pregunta, he recopilado datos de seis fuentes: Reddit (incluyendo publicaciones y comentarios), YouTube (que abarca títulos de videos y transcripciones) y diversas fuentes de noticias (que incluyen titulares y artículos completos). Estos datos abarcan el período desde 2018 hasta 2022.

Es importante destacar que este tipo de estudio es relativamente poco explorado en el idioma español, lo que nos lleva a la novedad e importancia de esta investigación. Para investigar este tema a fondo, planeo emplear diversas herramientas de análisis lingüístico, como el Análisis Lingüístico y de Palabras (LIWC), distintas evaluaciones de legibilidad, análisis de partes del discurso y examen de características a nivel de caracteres.

El objetivo principal de este análisis es identificar patrones y variaciones en las expresiones lingüísticas en estas fuentes de datos. Al hacerlo, espero establecer conexiones entre cambios lingüísticos e instancias de incertidumbre económica, tal como se indica mediante

métricas como el índice de riesgo país. Además, mi objetivo es explorar un aspecto secundario de este estudio, que implica examinar la posible relación con un lenguaje similar a la depresión. Asi pudiendo rescatar esta característica de la idea original de investigación pero de una forma mas profunda.

Algunos de los papers que utilizare no son específicamente de economía sino de Computer Sciences, Political Sciences, etc. debido a que esta aplicación en economía no es todavía usada habitualmente.

En primer lugar, el estudio de (Aromí, 2022) si bien es un working paper me incentivo a utilizar estas tecnologías en problemas de economía. Nos da aparte la posibilidad de armar un índice de incertidumbre para utilizar además del riesgo país. *"Este trabajo propone un índice que describe las opiniones económicas transmitidas por usuarios argentinos en la red social Twitter. Luego de identificar mensajes económicos, éstos son clasificadas según la frecuencia con la que se utilizan palabras asociadas a incertidumbre. La evaluación cualitativa del índice sugiere un fuerte vínculo con eventos económicos y políticos de relevancia. Estimaciones de modelos estadísticos indican que el índice contiene información sobre el ciclo económico, la confianza del consumidor y la evolución del mercado cambiario. Análisis complementarios demuestran que el foco en el concepto de incertidumbre y el uso de técnicas de procesamiento de lenguaje natural constituyen elementos clave para el desempeño satisfactorio de este indicador de opiniones."*

Luego, los hallazgos de (Vandoros et al., 2019) y de (Vandoros & Kawachi, 2021) ofrecen una visión fundamental en relación con suicidio/salud mental y periodos de incertidumbre económica. Si bien no utilizan metodología similar a la de mi tesis, aprotan un justificativo a esta investigación. Y los considero literatura base de la cual parto con nuevas metodologías a investigar una temática similar.

En este paper, publicado recientemente (Charemza et al., 2022), se profundiza en la utilización de NLP para medir incertidumbre económica y política. Resulta útil ver la forma en la que procesan y scrappean los datos de artículos de diario rusos. Si bien el objetivo final no es tan similar, las herramientas y métodos utilizados sirven de guía para varios elementos ya mensionados de la tesis, como el uso LIWC y tanto el filtrado, limado y scrape de diarios

virtuales. Ademas muestra la potencia que puede tener estas herramientas para el análisis de países con datos no siempre disponibles o confiables. *"El artículo propone un método para construir medidas específicas por país basadas en texto para la incertidumbre de políticas económicas. Para evitar problemas de traducción y costos de validación humana, aplicamos procesamiento de lenguaje natural y análisis de sentimiento para construir dichas medidas para Rusia. Comparamos nuestra medida con la desarrollada previamente mediante traducciones directas del inglés y validación humana. En esta comparación, nuestra medida funciona igual de bien al evaluar la incertidumbre relacionada con eventos clave que afectaron a Rusia entre 1994 y 2018 y tiene un mejor desempeño en la detección de los efectos de la incertidumbre en la producción industrial de Rusia."*

Además, (Nanath et al., 2022) resulto de ayuda y ejemplo para explorar una posible forma de relacionar *depression-like language* con los resultados de la primera mitad del modelo. Proveen un framework sumamente útil para este tipo de análisis, la única problemática es que es en Ingles. "*Los gobiernos de todo el mundo han implementado restricciones rigurosas para frenar la propagación de la pandemia de COVID-19. Aunque beneficiosas para la salud física, estas medidas preventivas podrían tener un profundo efecto perjudicial en la salud mental de la población. Este estudio se centra en el impacto de los confinamientos y las restricciones de movilidad en la salud mental durante la pandemia de COVID-19. En primer lugar, desarrollamos un novedoso índice de salud mental basado en el análisis de datos de más de tres millones de tweets a nivel global, utilizando el enfoque de aprendizaje automático de Microsoft Azure. Luego, los puntajes del índice de salud mental calculados se regresionan con el índice de rigidez del confinamiento y el índice de movilidad de Google utilizando una regresión de mínimos cuadrados ordinarios (OLS) con efectos fijos. Los resultados revelan que la reducción de la movilidad laboral, la reducción de la movilidad en tiendas y actividades recreativas, y el aumento de la movilidad residencial (confinamiento en la residencia) han afectado negativamente la salud mental. Sin embargo, se encontró que las restricciones de movilidad en parques, tiendas de comestibles y farmacias no tienen un impacto significativo. El índice de salud mental propuesto proporciona un camino para estudios teóricos y empíricos de salud mental utilizando las redes sociales.*"

Las decisiones en (Penczynski, 2019) para clasificar comunicaciones resulta de gran ayuda para tomar decisiones sobre el modelo a desarrollar. El *fine tunning* puede ser a veces arbitrario y un proceso extremadamente tedioso, lo cual este paper resulta de utilidad para ahorrarnos y justificar decisiones de *fine tunning*

Sabemos que el uso de texto procesado como datos puede ser riesgo. Como toda herrmienta tiene sus fuerzas y debilidades, como posible fallas en su utilización. (Benoit et al., 2009) escribe un paper en detalle de los posibles problemas, biases y errores de estas herramientas, lo cual nos permite reforzar o aceptarlas.

Por otro lado, el trabajo de (Yang & Srinivasan, 2016) es un excelente ejemplo en metodología y uso de redes sociales, en su caso twitter para el análisis del comportamiento de las personas en su vida diaria. "*La satisfacción con la vida se refiere a una evaluación cognitiva relativamente estable de la propia vida. La satisfacción con la vida es un componente importante del bienestar subjetivo, el término científico para la felicidad. El otro componente es el afecto: el equilibrio entre la presencia de emociones positivas y negativas en la vida diaria. Mientras que el afecto ha sido estudiado utilizando conjuntos de datos de redes sociales (particularmente de Twitter), la satisfacción con la vida ha recibido poca o ninguna atención. En este estudio, examinamos las tendencias en las publicaciones sobre satisfacción con la vida a partir de una muestra de datos de Twitter de dos años. Aplicamos una metodología de vigilancia para extraer expresiones tanto de satisfacción como de insatisfacción con la vida. Un resultado destacado es que, de acuerdo con sus definiciones, las tendencias en las publicaciones sobre satisfacción con la vida son inmunes a eventos externos (políticos, estacionales, etc.), a diferencia de las tendencias en el afecto informadas por investigadores anteriores. Al comparar usuarios, encontramos diferencias entre usuarios satisfechos e insatisfechos en varios aspectos lingüísticos, psicosociales y otros. Por ejemplo, estos últimos publican más tweets expresando enojo, ansiedad, depresión, tristeza y sobre la muerte. También estudiamos a usuarios que cambian su estado con el tiempo, pasando de estar satisfechos con la vida a estar insatisfechos o viceversa. Es destacable que las características de los tweets psicosociales de los usuarios que cambian de satisfechos a insatisfechos son bastante diferentes de aquellos que se mantienen satisfechos con el tiempo. En general, las observaciones que realizamos son coherentes con la intuición y con las observaciones en la investigación de las ciencias sociales. Esta investigación contribuye al estudio del bienestar subjetivo de las personas a través de las redes sociales."*

(Ramírez-Esparza et al., 2007) abordan una vez mas este tema pero desde una perspectiva que nos es extremadamente útil: en español. No es un paper demasiado con un objetivo de análisis en particular, sino que desarrollan y explican herramientas de NLP para el análisis psicológico en español. "*El Buscador Lingüístico y Contador de Palabras (LIWC, por sus siglas en inglés, Pennebaker, Francis, & Booth, 2001) es un programa de computadora que analiza textos. Este programa calcula el porcentaje de palabras dentro de un texto de acuerdo a varias docenas de categorías. La fiabilidad de este programa ha sido demostrada ampliamente en el ámbito de la lengua inglesa. En esta investigación dos estudios se llevaron a cabo para analizar la equivalencia del programa en español al programa en inglés. En el Estudio 1 se presenta el procedimiento de traducción del LIWC del inglés al español, y se demuestra la equivalencia entre las categorías del LIWC en inglés y sus correspondientes categorías en la versión en español. En el Estudio 2 se muestra el uso del LIWC en inglés y en español al comparar el lenguaje utilizado por mujeres en foros de discusión de depresión y de cáncer de mama en Internet. Los resultados mostraron que las versiones correlacionan en la mayoría de las categorías. Asimismo, se encontró que las mujeres en foros de depresión utilizan distintas categorías de palabras que las mujeres en foros de cáncer de mama y estas diferencias son similares en foros de discusión en español e inglés. Se discuten las implicaciones de usar este programa dentro de la lengua española."*

Al igual que el paper previo, (Pérez et al., 2021) presenta una herramienta de clasificación de sentimientos en español, pero la particularidad es que es en español argentino. "*La extracción de opiniones de textos ha generado un gran interés en los últimos años, ya que estamos experimentando un volumen sin precedentes de contenido generado por usuarios en redes sociales y otros lugares. Un problema que los investigadores sociales encuentran al utilizar herramientas de minería de opiniones es que suelen estar detrás de API comerciales y no están disponibles para otros idiomas que no sean el inglés. Para abordar estos problemas, presentamos pysentimiento, una herramienta de Python multilingüe para Análisis de Sentimientos y otras tareas de Procesamiento de Lenguaje Natural (NLP) en entornos sociales. Esta biblioteca de código abierto ofrece modelos de vanguardia para el español y el inglés de manera transparente, lo que permite a los investigadores acceder fácilmente a estas técnicas."*

Finalmente, (Khalid & Srinivasan, 2020) ofrecen una solida aplicación y variedad de herramientas para el análisis de estilos lingüísticos en: redes sociales como 4chan y reddit;

noticias acumuladas con voat. "*Históricamente, el contenido ha sido el enfoque principal para estudiar el lenguaje en las comunidades en línea. En cambio, este artículo se centra en el estilo lingüístico de las comunidades. Si bien sabemos que las personas tienen estilos distinguibles, aquí nos preguntamos si las comunidades tienen estilos distinguibles. Además, mientras que el trabajo previo se ha basado en una definición estrecha de estilo, empleamos una definición amplia que involucra 262 características para analizar el estilo lingüístico de 9 comunidades en línea de 3 plataformas de redes sociales que discuten sobre política, televisión y viajes. Descubrimos que las comunidades efectivamente tienen estilos distintivos. Además, el estilo es un excelente predictor de la pertenencia a un grupo (puntuación F de 0.952 y precisión del 96.09%). Aunque en promedio es estadísticamente equivalente a las predicciones utilizando solo el contenido, es más resistente a las reducciones en los datos de entrenamiento.*"

Referencias

Aromí, J. D. (2022). Medición de Incertidumbre Económica en Redes Sociales en Base a Modelos de Procesamiento de Lenguaje Natural. *Working Papers*. https://ideas.repec.org/p/aoz/wpaper/179.html

Benoit, K., Laver, M., & Mikhaylov, S. (2009). Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science*, *53*(2). https://doi.org/10.1111/j.1540-5907.2009.00383.x

Charemza, W., Makarova, S., & Rybiński, K. (2022). Economic uncertainty and natural language processing; The case of Russia. *Economic Analysis and Policy*, *73*, 546–562. https://doi.org/10.1016/J.EAP.2021.11.011

Khalid, O., & Srinivasan, P. (2020). Style Matters! Investigating Linguistic Style in Online Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 360–369. https://doi.org/10.1609/ICWSM.V14I1.7306

Nanath, K., Balasubramanian, S., Shukla, V., Islam, N., & Kaitheri, S. (2022). Developing a mental health index using a machine learning approach: Assessing the impact of mobility and lockdown during the COVID-19 pandemic. *Technological Forecasting and Social Change*, *178*. https://doi.org/10.1016/j.techfore.2022.121560

Penczynski, S. P. (2019). Using machine learning for communication classification. *Experimental Economics*, *22*(4). https://doi.org/10.1007/s10683-018-09600-z

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. https://arxiv.org/abs/2106.09462v1

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá, R. (2007). The psychology of word use: A computer program that analyzes texts in Spanish. *Revista Mexicana de Psicología*, *24*(1), 85–99. https://psycnet.apa.org/record/2007-08364-008

Tham, W. W., Sojli, E., Bryant, R., & McAleer, M. (2021). Common Mental Disorders and Economic Uncertainty: Evidence from the COVID-19 Pandemic in the U.S. *PLOS ONE*, *16*(12), e0260726. https://doi.org/10.1371/JOURNAL.PONE.0260726

Vandoros, S., Avendano, M., & Kawachi, I. (2019). The association between economic uncertainty and suicide in the short-run. *Social Science & Medicine*, *220*, 403–410. https://doi.org/10.1016/J.SOCSCIMED.2018.11.035

Vandoros, S., & Kawachi, I. (2021). Economic uncertainty and suicide in the United States. *European Journal of Epidemiology*, *36*(6), 641–647. https://doi.org/10.1007/S10654-021-00770-4/TABLES/3

Yang, C., & Srinivasan, P. (2016). Life Satisfaction and the Pursuit of Happiness on Twitter. *PLOS ONE*, *11*(3), e0150881. https://doi.org/10.1371/journal.pone.0150881