UNIVERSIDAD DEL BÍO-BÍO
FACULTAD DE CIENCIAS EMPRESARIALES

Análisis de comentarios de usuarios de la red social Twitter en temáticas de contingencia nacional en territorios específicos de Chile

Proyecto de Título presentado en conformidad a los requisitos para obtener el título de Ingeniero Civil en Informática

Nombre: Francisco Fernando Muñoz Coletti

Profesor guía: Pedro Gerónimo Campos Soto

Profesora co-guía: Nathalie Risso Sepúlveda

Fecha: xx/07/2022

Indice

1.- Introducción.

1.1 Motivación.

Las redes sociales en la actualidad poseen un muy alto flujo de información, sea formal o informal. Alrededor del 60.1% de la población mundial actualmente utiliza internet (4.7 billones de personas). De estos usuarios, el 92.12% utiliza plataformas de redes sociales activamente. (*Kemp, 2021*). Si se observa exclusivamente en Chile, un 83.5% de su población (16 Millones) son usuarios activos en redes sociales (*Alvino, 2021*).

El uso de dichas redes va desde el ocio, pasando por columnas de opinión hasta incluso denuncias ciudadanas sobre problemáticas que se viven día a día en nuestro país. Estas plataformas se han convertido actualmente en la voz del pueblo chileno y su uso se intensificó aún más post 18 de Octubre de 2019 (en comparación a meses anteriores) (*Ferreira, 2020*).

Sin embargo esta información queda registrada, mas no organizada para el posterior estudio de la misma por las dificultades que se presentan a la hora de clasificarla: Cuentas fraudulentas, información falsa, el alto costo que significaría tener un equipo encargado de recolectar y analizar los sentimientos expresados en dicha información de forma análoga, entre otras.

1.2 Problemática.

Este Proyecto de Título se enfoca en obtener opiniones emitidas respecto a las pensiones y en temáticas de salud por usuarios de Twitter en territorios específicos de Chile y obtener los sentimientos presentes en cada una de ellas.

En Chile, la problemática de las AFP se viene extendiendo hace mucho tiempo pues presenta varias condiciones desfavorables para los jubilados por este sistema como el valor de la pensión básica solidaria

de entre $164.356 a $176.096 pesos chilenos (*Instituto de Previsión Social de Chile, 2021*) o la tasa de mortalidad de 110 años (*Superintendencia de Pensiones, 2016*). Por otro lado la salud no queda libre de quejas, sobre todo por las "filas de espera" principalmente presentes en el sistema de salud pública las cuales generan frustración, impotencia y un sentimiento de abandono de promesas incumplidas y segregación(*Alister, 2018*).

Resta preguntar: ¿Qué herramientas se tienen al alcance para obtener las opiniones de la gente en respecto a lo mencionado en el párrafo anterior? Respuesta: Las encuestas de opinión realizadas por las empresas de investigación de opinión. Sin embargo, y poniendo el caso de la encuesta CADEM, una investigación periodística concluyó que de una muestra de 3,285 llamados a teléfonos prepago y postpago, solo cuenta con una efectividad del 21.76% siendo 715 llamados contestados. Adicionalmente la ficha técnica de dichas entrevistas no dice nada acerca de la no respuesta a la encuesta.(*Cumsille, 2018*). Adicionalmente, estas encuestas son encargadas y financiadas por el grupo que requiera obtener "titulares favorables" para dichos negocios, logrando la proeza de "mentir utilizando estadística"(*Matamala, 2021*)

Una revisión rápida por la red social Twitter revela hashtags que están al tanto con el acontecer nacional, desde denuncias ciudadanas hasta eventos. Siendo Twitter una forma fácil y rápida de generar opinión con respecto a estas temáticas y también cuenta con herramientas para obtener de manera efectiva estas opiniones.

Dicho todo lo anterior: ¿Existe una forma de mejorar la actual forma de recolección de opinión?

1.3 Trabajos relacionados

Con anterioridad se realizó un trabajo relacionado al cambio de actividad en la red social Twitter post 18 de Octubre de 2019 de Miguel Ferreira. En dicho trabajo, se realizó una fase inicial de marco teórico donde especificaba todas las herramientas y técnicas a utilizar en las siguientes fases.

Posteriormente se llevó a cabo la fase de recolección de datos, la cual identifica en una primera instancia los usuarios chilenos mediante elemento GeoJSON. Luego utiliza los diferentes hashtags importantes de esa fecha para realizar un segundo filtro aparte del primero. Finalmente junta ambos filtros para obtener la cantidad de usuarios Chilenos.

Llegado a este punto, recolecta los datos de los perfiles de los usuarios filtrados previamente y realiza filtros a usuarios tipo bot, entidades, posibles cuentas fraudulentas o que no cumplan con ciertas reglas previamente establecidas por el mismo. Una vez realizado este filtro, se guardan todos los datos recolectados en una base de datos para el posterior análisis.

En el análisis revisa el horario de publicación, que actividades ha tenido la cuenta y con unas formulas estipuladas se obtienen porcentajes de participación en la red social. Luego graficó los resultados obtenidos y concluyó que la actividad post 18 de Octubre del 2019 aumentó en un 97.46%.

Otro trabajo relacionado fue realizado por Hidroy, Ekram, Islam, Ahmed y Rahman relacionado a la minería de opinión localizada en Twitter usando análisis de sentimientos relacionado a opiniones sobre el Iphone 6. En primera instancia se realiza la consulta a la API de Twitter relacionada a Iphone y se entregan parámetros de localización como latitud, longitud, radio de cobertura, entre otros. Con esto se obtienen todos los Tweets de esa localización dada y se proceden a almacenar en una base de datos MySQL.

Para la fase del análisis de sentimientos se utilizó SentiWordNet, una herramienta para procesamiento del lenguaje natural. El problema de esta herramienta es que no reconoce la oración por completo, sino que por palabra, lo cual puede traer ciertos errores de interpretación por parte de la misma. Para reducir estos errores, se utiliza un etiquetado gramatical (POS) dividiendo la oración en Pronombres, Adjetivos, Verbos y conectores. El proceso de etiquetado queda representado por la Figura 1.

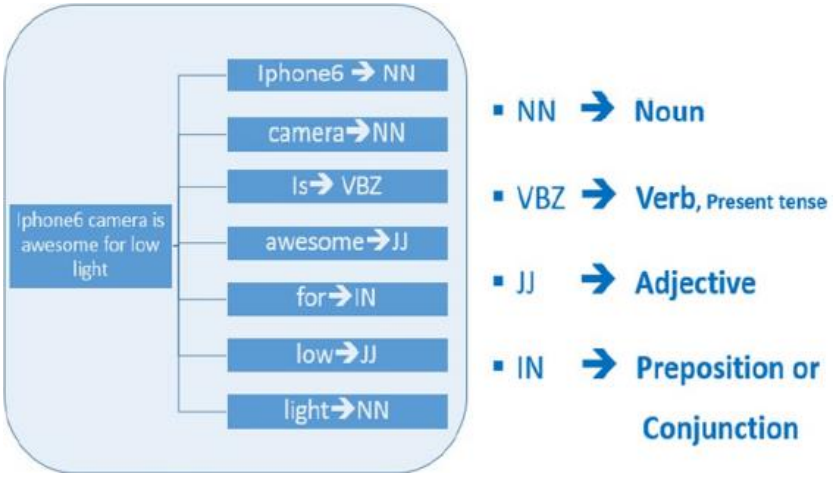

Figura 1: Ejemplo de Etiquetado Gramatical

Ahora para el puntaje total entregado por SentiWordNet está dado por la siguiente ecuación:

$$Score(Location_j) = \frac{\sum_{i=1}^{n} SentiScore_i}{n}$$

Donde n es el total de Tweets obtenidos, $SentiScore_i$ es el puntaje de cada palabra por separado, y $Location_j$ es la ubicación del Tweet.

Finalmente los resultados son clasificados en primera instancia según el sexo y luego según el puntaje obtenido del proceso anterior y como conclusión obtuvieron que la metodología que utilizaron es viable para realizar un estudio de opinión utilizando herramientas de análisis de sentimientos a usuarios de Twitter de distintas localidades.

1.4 Objetivos.

1.4.1 Objetivo general:

Categorizar opiniones de usuarios de Twitter en territorios específicos de Chile sobre temas de interés nacional mediante Análisis de Sentimientos.

1.4.2 Objetivos específicos:

- Revisar literatura relacionada a análisis sentimental de textos en redes sociales, uso de la API de Twitter y geolocalización en formato GeoJSON.

- Identificar y filtrar usuarios y sus respectivos Tweets pertenecientes a territorios específicos dentro de Chile mediante las funcionalidades de herramientas como Tweepy y GeoJSON.

- Seleccionar y entrenar un modelo de análisis de sentimientos para aplicar a temas de interés nacional sobre Tweets obtenidos.

- Implementar una interfaz gráfica sencilla para visualizar los resultados obtenidos.

1.5 (Posible) Estructura del informe.

Capítulo 1. Resumen del documento, introducción al tema en cuestión, problemática y objetivos.

Capítulo 2. Marco teórico del documento, abarcando las diferentes herramientas, software y conocimientos requeridos para realizar la investigación.

Capítulo 3. Recolección de datos y filtros. Todo proceso necesario para obtener un resultado fiable: filtros de usuarios bot, spam, instituciones, entre otras. Almacenamiento de los Tweets obtenidos según zonas definidas y desarrollo de scripts en Python para este proceso.

Capítulo 4. Procesamiento de lenguaje natural. Mediante machine learning se realiza el procesamiento de los comentarios para clasificarlos según el estado emocional que estos presentan.

Capítulo 5. Resultados. Toda información obtenida será graficada para realizar el estudio y las conclusiones posteriores.

Capítulo 6. Conclusiones y Trabajo futuro. En base a los resultados obtenidos se realizan conclusiones y se proponen posibles mejoras para futuro.

2. Marco teórico.

2.1 Twitter.

Es una plataforma social abierta de comunicación bidireccional con el que se puede compartir información de diverso tipo de forma rápida y sencilla (*Twitter Inc., 2021*). Esta opera mediante ciertas acciones propias de la plataforma:

- **Tweet:** Publicaciones realizadas en la plataforma.
- **Likes:** Cantidad de aprobaciones que tiene dicho Tweet.
- **ReTweet(RT):** Compartir la publicación de otra persona y/o las veces que se ha compartido la publicación de esa persona.
- **Followers:** Personas que siguen a una cuenta en específico.
- **Hashtag:** Representado con el símbolo # se utiliza para marcar tendencias dentro de la plataforma.
- **Trending Topic(TT):** Tendencias más habladas/publicadas en la plataforma en un tiempo en específico.
- **Comentarios:** Opiniones de usuarios sobre alguna publicación.

- **Timeline:** Lugar donde aparecen las publicaciones propias y de los usuarios a quien se sigue.

- **Mensaje Directo (DM o MD):** La plataforma te da la posibilidad de tener una comunicación uno a uno de manera privada, este mensaje no se verá publicado en la Timeline.

Para esta investigación, se utilizará Twitter (más específicamente la API de Twitter) para obtener información acerca de temas de contingencia nacional, opiniones de dichos usuarios para posteriormente realizar un análisis de sentimientos con respecto a esas opiniones.

2.2 Twitter API.

Es un set de utilidades que se pueden aplicar desde aprender hasta interactuar con Twitter. Esta API te permite encontrar, obtener, interactuar o crear una variedad de recursos incluidos los siguientes:

- Tweets
- Usuarios
- Mensajes Directos
- Listas
- Tendencias
- Archivos Multimedia
- Lugares

La API de Twitter cuenta con: API Rest y API Streaming. Siendo la primera para obtener datos históricos de la red social (contenido publicado) y la segunda para obtener datos en tiempo real. La figura 2 representa las distintas interacciones que se realizan al hacer uso de Twitter API mediante Tweepy.

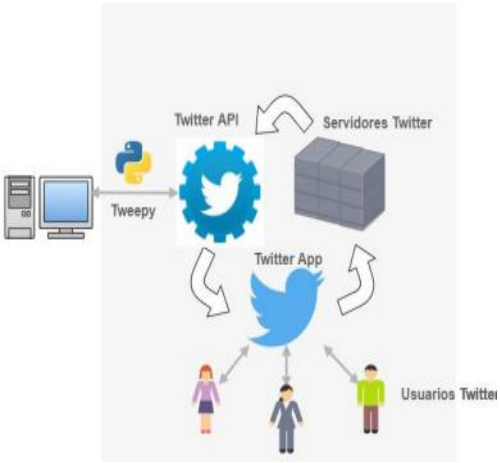

Figura 2: Interacción entre Usuarios, Twitter y Tweepy

2.2.1 Limitaciones de API de Twitter.

La API de Twitter solo puede procesar 180 requerimientos cada 15 minutos. Esto se traduce en:

- 450 Tweets por requerimiento (total 81.000)
- 900 usuarios por requerimiento (total 162.000)
- 1500 Tweets por timeline. Solo los últimos 3200 Tweets, nada anterior a ello (incluyendo retweets) (total 270.000)

2.2.2 Autenticación.

Para extraer información desde Twitter se debe autenticar. En primera instancia se debe contar con una cuenta asociada a Twitter para solicitar permiso de desarrollador, con esto se puede crear

aplicaciones de Twitter (apps.twitter.com). El protocolo de seguridad que utiliza para la autenticación Twitter es OAuth 2.0.

Una vez creada la aplicación en Twitter, se hará entrega de cuatro claves: API Key, API Secret, Access Token y Access Token Secret.

2.2.3 GeoJSON

Información estructurada sobre ubicaciones geográficas y sus propiedades. Utiliza JavaScript Object Notation (JSON) . Descubierto en Freenode: #openlayers and #gdal, 2006 y las especificaciones de su formato fue publicada en 2008.(*Butler, Daly, Doyle, Gillies, Hagen, Schaub & Wilde, 2014)*.

Un objeto GeoJSON representa una figura geometría, una característica o una colección de características. Considerar que GeoJSON es un objeto JSON. Cuenta con un miembro de nombre "type". Y el valor del miembro DEBE ser uno de tipo GeoJSON.

2.3 Tweepy

El lenguaje a utilizar en este proyecto será Python 3.9. En Python podemos trabajar con la API de Twitter mediante la librería/dependencia Tweepy. Esta librería contiene todas las funciones a requerir para la recolección de datos desde la API Twitter.

2.4 Machine Learning.

Es una ramificación del manejo de Inteligencia Artificial. Se basa en algoritmos para imitar el aprendizaje humano de manera gradual y bajo constante entrenamiento.

El termino fue nombrado por Arthur L. Samuel, ex trabajador de IBM quien además fue pionero en el campo de los videojuegos.

2.5 Transfer Learning.

Metodología de machine learning donde un modelo ya desarrollado para alguna tarea en específico es reutilizado para realizar otra tarea. Estos modelos usualmente ya vienen pre-entrenados y son ampliamente utilizados en el Procesamiento de Lenguaje Natural.

2.6 Deep Learning.

El Deep Learning se ve englobado tanto por las la inteligencia artificial como por los modelos de Machine Learning. Mientras que en los modelos más tradicionales de Machine Learning se le entregan las características de los datos de entrada, el Deep Learning se encarga de extraerlos por si solo, además de determinar la transformación a aplicar para dichas características.(*Borrero & Arias, 2021*).

2.7 Contexto de Estudio.

- Pandemia Covid19.
- Avance de vacunación.
- Elección presidencial/parlamentaria.
- Grupos anti vacuna.
- Apertura del país.
- Fin del toque de queda.
- Clases Presenciales.
- Crisis social en Chile.
- Polarización política.
- Conflictos en la Lista del Pueblo.

- Conmemoración del Golpe de Estado

- Cuarto retiro del 10% de las AFP

- IFE Laboral

- Inflación

- Aumento en la Tasa de Política Monteria

- Acusación constitucional en contra de Sebastián Piñera

- Aumento de casos de Covid-19

- Recesión mundial.

- Aumento en la inflación.

2.8 Procesamiento del Lenguaje Natural.

2.8.1 ~~Definición de~~ Lenguaje.

Método utilizado para lograr la comunicación. Este puede ser escrito, visual o verbal. Es un conjunto de frases formadas a partir un alfabeto las cuales cumplen con cierta reglamentación (sintaxis y gramática), posee coherencia(semántica) y permite expresar de manera clara ideas y opiniones.

2.8.2 ~~Definición de~~ Lenguaje Natural.

Uso cotidiano del lenguaje. Este se ve afectado en ~~su~~ gran medida por la idiosincrasia del lugar en el que se esté, su cultura, su historia y el folklore, por lo cual varía mucho en el tiempo. Generando así modismos, acentos, entonaciones y expresiones distintas las cuales no necesariamente siguen un reglamento establecido, sino que se medirían según la aceptación social y comprensión entre pares de una misma comunidad.

2.8.3 Uso del Procesamiento del Lenguaje Natural(PLN).

Principalmente utilizado en facilitar la comunicación con una computadora, ya que esta, mediante el entrenamiento requerido, podrá ser capaz de entender y comprender las oraciones que le sean entregadas. Esta técnica es ampliamente utilizada en la extracción de Información (Opinion Mining), traducción automática, reconocimiento de voz, entre otros. (*Cortez, Vega, Pariona, Huayna, 2009*).

2.9 Análisis de sentimientos.

Tipo de procesamiento de lenguaje natural usado para obtener el efecto emotivo que genera algún tema, objeto u opinión en la población a estudiar. Se basa en determinar de forma automatizada el nivel de subjetividad de un texto, la polaridad de este (positivo, negativo o neutro) y la fuerza con la que se emite el texto.(*Brooke, Tofiloski & Taboada, 2009*).

Este proceso puede ser utilizado de muchas formas, por ejemplo: obtener opiniones sobre un producto para recomendar similares, aplicaciones en marketing (*Vinodhini, & Chandrasekaran, 2012)*, realizar un estudio de la población sobre temas relacionados a necesidades básicas, en estadística, en psicología, entre otras.

Algunos de los desafíos del análisis de sentimientos son la capacidad que debe tener la máquina para interpretar cuando una opinión es positiva en un contexto pero negativa en otra o el entender las diferentes formas de expresar una opinión.

Un ejemplo de lo mencionado anteriormente puede ser: “La película fue muy buena” tiene claramente una intención positiva y se clasificaría como más positiva que negativa. Pero si se escribe “La película no fue muy buena” sigue teniendo palabras similares y es probable que siga siendo considerada como opinión positiva.

Mediante Machine Learning se puede entrenar a una máquina para ir superando estos desafíos anteriormente planteados, con la finalidad de que esta pueda detectar de la manera más precisa posible la intención tras la opinión planteada sobre el tópico a analizar.

Bibliografía

Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., Schaub, T., & Wilde, E. (2014). GEOJSON. https://www.ietf.org/proceedings/92/slides/slides-92-dispatch-3.pdf

Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., Schaub, T., & Wilde, E. (2016). The GeoJSON Format. https://www.hjp.at/(de)/doc/rfc/rfc7946.html#sec1

Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment Analysis and Opinion Mining: A Survey (6.ª ed., Vol. 2). https://www.researchgate.net/profile/Vinodhini-G-2/publication/265163299_Sentiment_Analysis_and_Opinion_Mining_A_Survey/links/54018f330cf2bba34c1af133/Sentiment-Analysis-and-Opinion-Mining-A-Survey.pdf

Superintendencia de Pensiones. (2016, junio). NUEVAS TABLAS DE MORTALIDAD DEL SISTEMA DE PENSIONES COMISIÓN DE FAMILIA Y ADULTO MAYOR. https://www.spensiones.cl/portal/institucional/594/articles-10993_recurso_1.pdf

Alister, J. (2018). Colas del sistema de salud en Chile. https://www.sistemaspublicos.cl/en_la_prensa/colas-del-sistema-de-salud-de-chile-columna-de-javier-alister-alumno-de-curso-introduccion-a-los-sistemas-publicos/

Borrero, I. P., & Arias, M. E. G. (2021). DEEP LEARNING. Universidad de Huelva. https://books.google.cl/books?hl=es&lr=&id=kzsvEAAAQBAJ&oi=fnd&pg=PA1&dq=que+es+deep+

learning&ots=Q1DVPW3LkE&sig=kGXDrMLylg7kxpCYuaGIQynFgGU&redir_esc=y#v=onepage&q&f=false

W. (2020, 6 enero). ¿Qué es twitter? ¿Cómo funciona? ¿Cómo puedo usarlo para mi organización? Webempresa. Recuperado el 8 de septiembre de 2021, de https://www.webempresa.com/blog/que-es-twitter-como-funciona-2.html

Twitter, Inc. (s. f.). Getting Started with Our Platform. Docs | Twitter Developer Platform. Recuperado 8 de septiembre de 2021, de https://developer.twitter.com/en/docs/getting-started

Education, I. C. (2021, 12 agosto). Machine Learning. IBM. Recuperado el 12 de septiembre de 2021, de https://www.ibm.com/cloud/learn/machine-learning

Brownlee, J. (2019, 16 septiembre). A Gentle Introduction to Transfer Learning for Deep Learning. Machine Learning Mastery. Recuperado el 12 de septiembre de 2021, de https://machinelearningmastery.com/transfer-learning-for-deep-learning/

J. (s. f.). GitHub - jlhonora/geo: GeoJSON data for Chile. GitHub. Recuperado 14 de septiembre de 2021, de https://github.com/jlhonora/geo

Amazon Reviews for Sentiment Analysis. (2019, 18 noviembre). Kaggle. Recuperado 21 de octubre de 2021, de https://www.kaggle.com/bittlingmayer/amazonreviews

Instituto de Previsión Social. (2021). Pensión Básica Solidaria de Vejez (PBSV). Recuperado 11 de noviembre de 2021, de https://www.ips.gob.cl/servlet/internet/content/1421810823272/pension-basica-solidaria-vejez

Cumsille, G. (2018, 3 septiembre). Las debilidades metodológicas de la encuesta Cadem. El Desconcierto - Prensa digital libre. Recuperado 18 de noviembre de 2021, de https://www.eldesconcierto.cl/opinion/2018/09/03/las-debilidades-metodologicas-de-la-encuesta-cadem.html

Matamala, D. (2021, 6 noviembre). Columna de Daniel Matamala: Cómo mentir con estadísticas. La Tercera. Recuperado 18 de noviembre de 2021, de https://www.latercera.com/la-tercera-domingo/noticia/columna-de-daniel-matamala-como-mentir-con-estadisticas/SIURJKU4TRATTF2H46JZ4IRKVE/

Cortez Vásquez, A., Vega huerta, H., Pariona Quispe, J., & Huayna, A. M. (2009). Procesamiento de lenguaje natural. Revista De investigación De Sistemas E Informática, 6(2), 45–54. Recuperado en 2021 a partir de https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923

Brooke, J., Tofiloski, M., & Taboada, M. (2009, September). Cross-linguistic sentiment analysis: From English to Spanish. In Proceedings of the international conference RANLP-2009 (pp. 50-54). https://aclanthology.org/R09-1010.pdf

Hridoy, S. A. A., Ekram, M. T., Islam, M. S., Ahmed, F., & Rahman, R. M. (2015). Localized twitter opinion mining using sentiment analysis. Decision Analytics, 2(1), 1-19.