Proyecto Final: Análisis de Texto con Reseñas.

Alumno: Irving Daniel Estrada López
Matrícula: 1739907

Paso 1: Buscar una problemática que vean en su día a día que les gustaría resolver.

Cada día se genera una cantidad enorme de información de la cual en su mayoría son datos no estructurados debido a las redes sociales, el texto y las imágenes son información valiosa que tanto el sector público como el privado deberían aprovechar para obtener una perspectiva de sus intereses. El análisis de interacción en texto es de suma importancia en la actualidad, con él podemos detectar áreas de oportunidad para el producto o servicio que se provee, así como intereses del cliente para futuras decisiones. Esto con el objetivo de mejorarlo o corregirlo. Estoy interesado en el análisis de interacción con texto ya que es común hoy en día el redactar reseñas en sitios como e-commerces, redes sociales, paginas oficiales de productos o servicios, entre otros sitios. El poder clasificar dichas reseñas nos puede ayudar a comprender el comportamiento que está teniendo nuestro producto o servicio en el mercado y basado en eso tomar decisiones.
Algunos de los beneficios que puede tener el análisis de texto son:

- Reducción de Costos y Velocidad
- Consistencia
- Escalabilidad
- Simplicidad

El conjunto de datos de interés para este proyecto es acerca de reseñas de las aplicaciones de 4 redes sociales: Facebook, Instagram, Twitter, TikTok. En la actualidad estas aplicaciones son comúnmente utilizadas. El conjunto de datos contiene un identificador del usuario que hizo la reseña, la reseña, su calificación y a que red social pertenece la reseña.

Referencia: SANSKAR HASIJA. (2022). Top 20 Play Store App Reviews (Daily Update). 14 de Julio 2022, de Kaggle Sitio web: https://www.kaggle.com/datasets/odins0n/top-20-play-store-app-reviews-daily-update?select=all_combined.csv

Paso 2: Describan cuales son las habilidades que Uds. tienen que pueden apoyar a resolver este problema.

He trabajado con texto desde el tetramestre pasado en la clase de Datos Masivos donde estudiamos el procesamiento de archivos de texto de forma paralela en Python, con la librería de multiprocesos. En ese proyecto me familiaricé con la terminología y los procesos que hay detrás de un preprocesado de texto, así como la extracción de características utilizando la técnica de TF-IDF. En dicha clase llegamos hasta esa fase, sin embargo, estoy interesado en la clasificación de texto por medio de modelos de Machine Learning y Deep Learning. En esta clase de Procesamiento y Clasificación de datos he aprendido más acerca de las posibles técnicas que se pueden llevar a cabo al trabajar con texto. Algo nuevo que aprendí en esta clase relacionado con texto, es el análisis de sentimiento basado en reglas con librerías como VADER, esto complementó lo que anteriormente había estudiado. Dichas librerías nos benefician al momento de no tener una cantidad de datos considerable, ya que esto es el punto débil del Machine Learning y el Deep Learning.

Paso 3: Hallazgos

El análisis de texto es algo que se estudia desde hace tiempo, la intención siempre ha sido poder extraer información valiosa de un conjunto de texto. A continuación, se enlistan los artículos de interés encontrados acerca del análisis de texto junto a una descripción del mismo.

1- Text Classification Algorithms: A Survey

En este artículo abordan una metodología que se debe de seguir al momento de trabajar con texto, que entre sus pasos está: Preprocesado de texto, Extracción de características, Reducción de dimensiones, Técnicas de Clasificación.

Desglosa de una manera muy detallada cada uno de los puntos mencionados, por dar un ejemplo en la parte de preprocesado nos enlista las técnicas más utilizadas para adecuar nuestros textos para poder ser clasificados. También nos habla del TF-IDF, entre distintas técnicas. También cabe destacar que abarca matemáticamente los modelos propuestos como: KNN, Naive Bayes, entre otros.

Referencia: Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

2- A Survey Of Text Classification Algorithms

Este artículo se trata de técnicas más avanzadas de Machine Learning y clasificación basada en reglas, las cuales se emplean o son ideales para la clasificación de texto. Habla acerca de la selección de características para poder entrenar nuestro modelo de forma adecuada para que tenga un buen desempeño. Se estudia a profundidad las técnicas de clasificación como: Naive Bayes, SVM, ANN, entre otras. A diferencia del artículo anterior que nos mostraba el preprocesado ideal y las demás secciones de un buen análisis de texto, este se enfoca en su mayoría en los modelos de clasificación.

Referencia: Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer, Boston, MA.

3- Automatic Text Classification: A Technical Review

Este artículo es más corto que los anteriores y se centra en la metodología genérica que se debería usar al momento de trabajar con texto. Dentro de la

metodología que se propone se cuenta con la obtención de los datos, después la fase de preprocesado donde nos mencionan alguna técnica como remover las stop-words, hacer el stemming, entre otras. Después tenemos la parte de extracción de características las cuales pueden ser extraídas con el TF-IDF, LSI o la Multiword. La siguiente fase consta de elegir el modelo de clasificación, utilizando Machine Learning, entrenar el modelo y por último evaluarlo.

Referencia: Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: a technical review. International Journal of Computer Applications, 28(2), 37-40.

4- Text Classification Using Machine Learning Techniques

Aquí nos proponen otra metodología la cual seguir para la clasificación de texto. Dentro de las etapas de este artículo se encuentra: leer el documento, tokenizar el texto, hacer el stemming, eliminar las stopwords, hacer la representación del texto de forma vectorial, la extracción de características y por último la fase de aprendizaje donde se debe elegir el modelo que se usará.

Referencia: Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.

Después de leer cada uno de los artículos encontrados, nos da un norte por donde ir. Con lo que hemos visto en clase y lo leído en estos artículos, el preprocesado es una fase importante en la clasificación de texto. Los modelos propuestos en estos artículos serán probados y me quedaré con los mejores, todavía no decido la cantidad de modelos a comparar, pero sin duda uno de los que utilizaré será el Random Forest debido a que es un modelo de aprendizaje en conjunto.

Paso 4: Esqueleto de metodología

A continuación, se desarrolla la estructura del documento final, omitiendo la introducción y conclusión. Como en las instrucciones del desarrollo de este

documento lo indica la metodología puede cambiar según el desarrollo del proyecto. Esta metodología está basada en los artículos anteriormente mencionados.
La metodología propuesta para llevar a cabo el análisis de texto es la siguiente:

Análisis descriptivo

Para el proyecto es importante comenzar con un análisis exploratorio y descriptivo para poder detectar áreas de oportunidad en el preprocesado, algunos ejemplos de esto son: patrones de ruido en las reseñas, conjunto de datos no balanceado, entre otras cosas.

Preprocesado

Dentro de la etapa del preprocesado está contemplado la limpieza de patrones de ruido en caso de que los haya, remover emojis ya que es común al momento de escribir reseñas en un dispositivo móvil el agregarlos, remover URL, quitar signos de puntuación y números, convertir en minúsculas, revisión de ortografía, tokenizar, la eliminación de las stopwords, tenemos que lematizar o hacer el stemming, entre otras cosas. Hay que hacer distintas pruebas para que el preprocesado se adecuado, ya que no en todas las ocasiones podemos utilizar el mismo preprocesado, una situación de lo anteriormente mencionado es que en algunas ocasiones se utilizan ciertas abreviaturas y expresiones al momento de escribir las cuales la revisión de ortografía termina confundiéndolas con otras palabras.

Modelos

En la sección de modelos se compararán distintos de estos, todavía no defino cuantos modelos se compararán, sin embargo, se harán pruebas con los que fueron utilizados en los artículos mencionados, como: Random Forest, Decision Tree, ANN, Naive Bayes, SVM, entre otros. Permanecerán los que obtengan mejores resultados para la clasificación de texto.

Resultados

Para evaluarlos utilizaremos una matriz de confusión, así como un reporte de clasificación. Lo ideal sería guardar un subconjunto de reseñar para al final evaluarlas, evitando que el modelo sea entrenado con ellas, simulando entradas nuevas. Dentro de esta sección se tendrá una tabla comparativa del desempeño de los modelos en la clasificación de texto.