Metodologías de investigación para un Análisis de Sentimiento de Migrantes por su paso por México

Introducción

El análisis de datos es un proceso fundamental en la era de la información, que implica la recopilación, transformación y organización de datos para extraer información útil y tomar decisiones informadas. En el contexto del análisis de sentimientos, este proceso se enfoca en interpretar las emociones y opiniones expresadas en textos, como publicaciones en redes sociales, comentarios, reseñas y otros tipos de contenido generado por usuarios.

El análisis de sentimientos, también conocido como minería de opiniones, es una técnica de procesamiento del lenguaje natural (PLN) que se utiliza para identificar y extraer subjetividad en los textos. Este análisis permite clasificar las opiniones expresadas en el texto como positivas, negativas o neutrales, proporcionando una comprensión más profunda de las percepciones y actitudes de las personas sobre diversos temas.

En el caso específico de los migrantes que atraviesan México, el análisis de sentimientos puede proporcionar una visión valiosa sobre sus experiencias y percepciones. Al analizar los tweets y otras publicaciones en redes sociales, se puede obtener información sobre los desafíos, esperanzas, miedos y experiencias de los migrantes durante su travesía. Este tipo de análisis es crucial para entender mejor las necesidades y problemas que enfrentan, y puede informar políticas y programas destinados a mejorar sus condiciones.

Varios estudios han demostrado la utilidad del análisis de sentimientos en diferentes contextos. Por ejemplo:

1. **"Twitter as a Tool for the Management and Analysis of Emergency Situations: A Systematic Literature Review"** - Este estudio revisa cómo se ha utilizado Twitter para gestionar y analizar situaciones de emergencia, destacando el papel del análisis de sentimientos para comprender las reacciones y necesidades del público en tiempos de crisis .

2. **"Analyzing Public Sentiments Online: The Role of Data Science in Sentiment Analysis"** - Este artículo aborda cómo la ciencia de datos y las técnicas de análisis de sentimientos se han aplicado para evaluar las opiniones públicas en diversas plataformas en línea, demostrando la eficacia de estas herramientas en la recolección y análisis de grandes volúmenes de datos .
3. **"Sentiment Analysis of Twitter Data during Hurricane Sandy"** - En este estudio, los investigadores utilizaron el análisis de sentimientos para evaluar las respuestas del público durante el huracán Sandy, proporcionando una comprensión más profunda de las emociones y necesidades de las personas afectadas por el desastre natural .

Estos estudios ejemplifican cómo el análisis de sentimientos puede ofrecer insights valiosos en contextos específicos, como el de los migrantes que atraviesan México. A través de la recopilación y análisis de datos textuales, es posible obtener una visión detallada de sus percepciones y experiencias, contribuyendo a una mejor comprensión y apoyo para estas comunidades vulnerables.

Metodología

Parte Técnica

La base técnica de nuestra metodología se centra en el uso del lenguaje de programación Python y varias librerías especializadas.

1. **Extracción de Datos:** Utilizamos la librería `twint` para la recolección de tuits desde la red social X (anteriormente Twitter). La elección de `twint` se basó en su capacidad para extraer tuits sin necesidad de autenticación API, lo que facilita la recopilación de grandes volúmenes de datos (Comet).

2. **Manipulación de Datos:** Para la manipulación y limpieza de los datos, utilizamos `pandas`. Esta librería es fundamental para estructurar los datos

en DataFrames, lo que nos permite realizar operaciones complejas de filtrado y transformación de manera eficiente (Svitla Systems).

3. **Análisis de Datos:** Empleamos `seaborn`, `matplotlib`, `numpy`, y `scipy` para el análisis y visualización de datos. Estas herramientas nos permiten generar gráficos y realizar análisis estadísticos que son cruciales para interpretar los resultados, (Built In) (datagy) (Stack Abuse) (datagy) (Finxter).

Para más información sobre las herramientas y técnicas utilizadas, puede consultar los siguientes enlaces:

- **Kaggle - Recursos para el Análisis de Sentimientos**

Desafíos Técnicos

Durante la extracción de tuits, nos enfrentamos a varios desafíos significativos**:**

- **Ruido en los Datos:** Muchos tuits recolectados no provenían de migrantes, sino de entidades gubernamentales o individuos comentando sobre la migración. Esto se debió a que las palabras clave iniciales (como "xenofobia", "albergue", "asilo") eran demasiado amplias y generaban mucho ruido.

- **Relevancia del Léxico:** El léxico que construimos inicialmente no capturaba adecuadamente las experiencias de los migrantes. Aunque incluimos términos específicos, el algoritmo seguía recuperando tuits irrelevantes. Por ejemplo, la palabra "miedo" podría aparecer en contextos no relacionados con la migración.

Para mitigar estos problemas, realizamos varias iteraciones de prueba y error para ajustar el léxico. Aumentamos la especificidad del léxico agregando frases comunes utilizadas por migrantes. Sin embargo, el problema persistió debido a la ambigüedad de las palabras clave.

Parte Analítica

Durante nuestra investigación sobre el análisis de datos centrado en migrantes que han expresado sus opiniones en plataformas como X (anteriormente Twitter), hemos abordado diversos métodos y enfrentado múltiples desafíos.

Inicialmente, empleamos una metodología manual que consistía en leer y buscar tweets de migrantes en la aplicación. Este enfoque nos permitió identificar palabras clave relevantes para nuestro léxico. Sin embargo, descubrimos que al usar solo palabras sueltas, muchos de los tweets obtenidos no eran pertinentes. Gran parte de estos tweets contenían palabras clave que, aunque relacionadas contextualmente, no tenían ninguna conexión con opiniones de migrantes (SpringerLink).

En un intento por mejorar la precisión, filtramos las palabras más efectivas y relevantes. A pesar de este esfuerzo, encontramos que los tweets más alineados con nuestra investigación eran sobre migración, pero no específicamente sobre México, lo cual era un requisito esencial para nuestro estudio.

Decidimos entonces probar una nueva estrategia: agregar el hashtag "México" a cada palabra clave para contextualizar los tweets dentro del ámbito geográfico relevante. Aunque esta técnica mejoró la precisión en cierto grado, la mayoría de los tweets recolectados eran de cuentas gubernamentales o personas opinando sobre temas políticos y migratorios sin ser migrantes ellos mismos.

Al constatar que nuestras metodologías automatizadas no estaban produciendo los resultados esperados, regresamos a un enfoque manual. Exploramos otras plataformas sociales como Reddit y Facebook, buscando grupos y comentarios de migrantes. Sin embargo, tampoco logramos obtener un volumen significativo de datos relevantes. Esto nos llevó a formular varias hipótesis sobre las posibles razones de nuestra baja tasa de éxito.

Una hipótesis plausible es que los migrantes pueden no tener la capacidad económica para acceder a dispositivos móviles con aplicaciones de redes sociales, o prefieren utilizar plataformas más privadas como WhatsApp o Messenger para comunicarse y expresar sus opiniones. Estas plataformas no permiten el acceso a través de sus APIs de la misma manera que Twitter, lo que limita nuestra capacidad de extracción y análisis de datos.

En conclusión, nuestra investigación ha revelado que, aunque existen herramientas y métodos potentes para el análisis de datos en redes sociales, hay desafíos significativos cuando se trata de obtener datos específicos de migrantes. La necesidad de utilizar enfoques mixtos y considerar limitaciones tecnológicas y sociales es fundamental para avanzar en esta área de estudio (PLOS).

1. **Filtrado de Datos:** Una vez recolectados los tuits, procedimos a filtrar aquellos que no eran relevantes. Utilizamos hashtags y frases clave adicionales para eliminar tuits irrelevantes. Por ejemplo, eliminamos todos los tuits que contenían hashtags no relacionados con la migración, como "#Peligro", "#Miedo" #"Refugio", #"asilo".

2. **Revisión Manual:** A pesar de nuestros esfuerzos de filtrado automatizado, la revisión manual fue necesaria. Leímos cada tuit para identificar aquellos que realmente provenían de migrantes. Esta tarea fue extremadamente laboriosa pero esencial para asegurar la calidad de los datos.

3. **Exploración de otras Plataformas:** Intentamos recolectar datos de otras plataformas como Facebook y Reddit. En Facebook, nos unimos a grupos relacionados con la migración, pero descubrimos que la mayoría de las publicaciones eran de entidades gubernamentales o personas ofreciendo ayuda, no de migrantes compartiendo sus experiencias. En Reddit, los tuits relevantes eran escasos y principalmente de migrantes europeos, no de aquellos que pasaban por México.

Para más información sobre las metodologías utilizadas, puede consultar los siguientes enlaces:

- **NCBI - Aspectos Técnicos del Análisis de Sentimientos**

Desafíos Técnicos

Durante la extracción de tuits, nos enfrentamos a varios desafíos significativos que afectaron la calidad y relevancia de los datos recopilados:

Ruido en los Datos

Uno de los principales problemas fue la presencia de ruido en los datos recolectados. Esto se refiere a tuits que no eran relevantes para nuestro estudio de migrantes, ya que provenían de fuentes no deseadas.

1. **Entidades Gubernamentales y Organizaciones:** Muchos tuits eran publicados por cuentas oficiales de entidades gubernamentales o por organizaciones que trabajan con migrantes. Estos tuits a menudo discutían políticas migratorias, programas de ayuda, estadísticas y eventos relacionados con la migración, pero no reflejaban las experiencias personales de los migrantes. Por ejemplo, un tuit de una agencia gubernamental podría hablar sobre la apertura de un nuevo centro de refugio, lo cual es relevante para el contexto general de la migración, pero no proporciona información sobre las experiencias o sentimientos de los migrantes individuales.

2. **Individuos Comentando sobre Migración:** Otro grupo significativo de tuits provenía de personas que discutían la migración desde una perspectiva externa, ofreciendo opiniones personales, comentarios políticos o discutiendo noticias relacionadas con la migración. Estos tuits, aunque relacionados con el tema de la migración, no provenían de migrantes y, por lo tanto, no proporcionaban la perspectiva directa que estábamos buscando. Por ejemplo, un tuit que comenta sobre una nueva política migratoria desde una perspectiva política no nos da información sobre cómo esa política afecta directamente a los migrantes.

Relevancia del Léxico

El segundo gran desafío fue la relevancia del léxico utilizado para filtrar y analizar los tuits. El léxico inicial incluía una serie de palabras clave

diseñadas para captar tuits relevantes, pero no siempre lograba su objetivo debido a varios factores**:**

1. **Ambigüedad de Palabras Clave:** Muchas de las palabras clave utilizadas, como "miedo", "refugio", y "asilo", son ambiguas y pueden aparecer en una variedad de contextos que no están relacionados con la migración. Por ejemplo, la palabra "miedo" podría ser utilizada en un contexto personal que no tiene nada que ver con la experiencia de migrar, como "Tengo miedo de los exámenes". Esta ambigüedad llevó a la recolección de muchos tuits irrelevantes.
2. **Inadecuación del Léxico Inicial:** El léxico inicial no capturaba adecuadamente las experiencias específicas de los migrantes. Aunque incluimos términos que pensábamos que eran específicos, como "xenofobia" y "asilo", el algoritmo continuaba recuperando tuits que no eran directamente relevantes para la experiencia migratoria. Por ejemplo, la palabra "refugio" podría referirse a un albergue para personas sin hogar en un contexto urbano, lo cual no se relaciona directamente con la experiencia de un migrante.

Mitigación de Problemas

Para mitigar estos problemas, llevamos a cabo varias iteraciones de prueba y error para ajustar y mejorar el léxico utilizado.

1. **Aumento de Especificidad:** Incrementamos la especificidad del léxico agregando frases comunes utilizadas por migrantes, como "cruzar la frontera", "buscando asilo", y "huir de la violencia". Estas frases fueron seleccionadas basándonos en estudios previos y en el análisis de tuits que ya habíamos recolectado. Por ejemplo, frases como "buscando una vida mejor" y "huyendo de la violencia" fueron identificadas como más específicas y directamente relevantes a la experiencia de migrar.

2. **Filtrado Adicional:** Implementamos técnicas de filtrado adicional para eliminar tuits irrelevantes. Esto incluía el uso de hashtags específicos y el análisis contextual de los tuits para determinar su relevancia. Sin embargo, debido a la naturaleza ambigua del lenguaje y a la variedad de contextos

en los que se utilizan ciertas palabras, el problema de ruido en los datos persistió.

A pesar de nuestros esfuerzos por mejorar la precisión del léxico y el filtrado de datos, la ambigüedad inherente de las palabras clave y la diversidad de contextos en los que se utilizan hicieron que fuera difícil eliminar completamente el ruido de los datos. Esta experiencia subraya la importancia de desarrollar métodos más avanzados y específicos para la recolección y análisis de datos en estudios futuros.

Desafíos Analíticos

1. **Filtrado de Datos**
   Una vez recolectados los tuits, procedimos a filtrar aquellos que no eran relevantes. Utilizamos hashtags y frases clave adicionales para eliminar tuits irrelevantes. Por ejemplo, eliminamos todos los tuits que contenían hashtags no relacionados con la migración, como "#vacaciones" o "#celebración". Este proceso fue crucial para reducir la cantidad de datos no relevantes, pero no fue suficiente para eliminar completamente el ruido.

2. **Revisión Manual**
   A pesar de nuestros esfuerzos de filtrado automatizado, la revisión manual fue necesaria. Leímos cada tuit para identificar aquellos que realmente provenían de migrantes. Esta tarea fue extremadamente laboriosa pero esencial para asegurar la calidad de los datos. La revisión manual permitió identificar contextos y matices que los filtros automáticos no podían detectar, pero también representó un enorme consumo de tiempo y recursos humanos.

3. **Exploración de Otras Plataformas**
   Intentamos recolectar datos de otras plataformas como Facebook y Reddit. En Facebook, nos unimos a grupos relacionados con la migración, pero descubrimos que la mayoría de las publicaciones eran de entidades gubernamentales o personas ofreciendo ayuda, no de migrantes compartiendo sus experiencias. En Reddit, las publicaciones relevantes

eran escasas y principalmente de migrantes europeos, no de aquellos que pasaban por México. Esta exploración fue fundamental para entender las limitaciones de cada plataforma y la dificultad de encontrar contenido relevante en diferentes contextos sociales.

4. **Ambigüedad Contextual**
   Las palabras clave a menudo se usaban en contextos no relacionados con la migración. Por ejemplo, "refugio" podría referirse a una casa, y "miedo" podría describir cualquier situación personal. Esta ambigüedad dificultó la tarea de filtrar automáticamente los tuits relevantes, ya que muchas palabras comunes tienen múltiples significados dependiendo del contexto.

5. **Relevancia de Hashtags**
   Muchos tuits relevantes no contenían hashtags específicos de migración, lo que dificultó el filtrado automático. Aunque los hashtags pueden ser útiles para categorizar contenido, su ausencia en tuits relevantes significó que dependíamos más del contenido textual para identificar la relevancia, aumentando así el trabajo manual requerido para asegurar la precisión.

6. **Revisión Manual Intensiva**
   La necesidad de revisar manualmente miles de tuits para asegurar su relevancia fue un proceso intensivo en tiempo y recursos. Aunque esta revisión fue crucial para garantizar la calidad de los datos, también representó un gran desafío logístico y de gestión de tiempo. Este proceso subraya la importancia de desarrollar mejores herramientas de filtrado automatizado para futuros estudios.

Resultados

Después de un proceso exhaustivo de recolección y filtrado, obtuvimos un conjunto de tuits limitados pero relevantes de migrantes. Los resultados indican varias tendencias y desafíos clave:

1. **Acceso Limitado a Tecnología:** Una de nuestras hipótesis es que muchos migrantes no tienen acceso a dispositivos tecnológicos o redes sociales debido a sus circunstancias. Esto limita su capacidad para compartir sus experiencias en línea.
2. **Prioridades Diferentes:** Los migrantes a menudo priorizan su seguridad y necesidades inmediatas sobre la expresión de sus experiencias en redes sociales. Esto reduce la cantidad de datos disponibles para el análisis.

3. **Privacidad y Seguridad:** Los migrantes pueden evitar compartir públicamente sus experiencias para proteger su seguridad, especialmente si están en una situación vulnerable.

4. **Contexto de Publicaciones:** En plataformas como Facebook, las publicaciones relevantes eran escasas y a menudo enfocadas en logística o solicitudes de información, no en compartir experiencias personales.

Conclusiones

Nuestra investigación sugiere que las dificultades para obtener datos directamente de migrantes en redes sociales se deben a varias razones clave:

1. **Falta de Recursos Tecnológicos:** Los migrantes a menudo no tienen acceso a dispositivos o redes sociales debido a sus circunstancias.

2. **Prioridades Diferentes:** La prioridad de los migrantes es su traslado y seguridad, no la expresión de sus experiencias en redes sociales.

3. **Privacidad y Seguridad:** Los migrantes pueden preferir no compartir públicamente sus experiencias para proteger su seguridad.

Hipótesis de las Dificultades

Durante el desarrollo de esta investigación, hemos formulado varias hipótesis para explicar por qué no pudimos obtener datos directamente de migrantes:

1. **Acceso Limitado a Tecnología:** Los migrantes a menudo no tienen acceso a dispositivos tecnológicos como smartphones o computadoras, ni a conexiones de internet estables. Esto limita su capacidad para usar redes sociales y compartir sus experiencias.

2. **Prioridades Diferentes:** Los migrantes están en una situación de vulnerabilidad donde su principal preocupación es su seguridad y necesidades básicas como alimentación y refugio. La expresión de sus experiencias a través de redes sociales no es una prioridad en su situación.

3. **Privacidad y Seguridad:** Compartir públicamente sus experiencias puede poner en riesgo su seguridad, especialmente si están huyendo de situaciones de violencia o persecución. Esto los lleva a evitar publicar detalles sobre su migración en plataformas públicas.

Tipo de Personas Objetivo

El objetivo de nuestra investigación era obtener datos de migrantes que estuvieran pasando o hubieran pasado por México. Queríamos captar sus experiencias, emociones y desafíos durante este proceso. Sin embargo, la mayoría de los datos que obtuvimos provenían de:

1. **Entidades Gubernamentales:** Muchas publicaciones relacionadas con la migración provienen de cuentas oficiales de gobierno, que discuten políticas, programas y estadísticas migratorias, pero no reflejan las experiencias personales de los migrantes.

2. **Organizaciones de Ayuda:** Varias publicaciones eran de organizaciones que proporcionan ayuda a migrantes, como ONGs, que comparten información sobre sus actividades y servicios.

3. **Personas Opinando Sobre Migración:** Una gran cantidad de tuits eran de personas que expresan su opinión sobre la migración, pero no eran migrantes ellos mismos. Estos tuits incluyen opiniones políticas, comentarios sobre noticias y discusiones generales sobre el tema de la migración.

Recomendaciones

Para futuras investigaciones y apoyo a los migrantes, recomendamos:

1. **Creación de Plataformas Específicas:** Desarrollar plataformas seguras y anónimas donde los migrantes puedan compartir sus experiencias. Esto no solo ayudaría a obtener mejores datos para investigaciones futuras, sino también a brindar apoyo y recursos a los migrantes durante su paso por México.
2. **Colaboración con Organizaciones:** Trabajar con organizaciones que apoyan a migrantes para obtener datos más precisos y relevantes. Estas organizaciones pueden tener acceso a información que no está disponible públicamente.
3. **Educación y Sensibilización:** Promover la educación y sensibilización sobre la importancia de compartir experiencias para mejorar el apoyo a los migrantes. Esto puede incluir campañas que animen a los migrantes a compartir sus historias en plataformas seguras.

Que podemos sacar con la información ya obtenida?

Hipótesis y Análisis del DataFrame

Dado el DataFrame que contiene principalmente opiniones de entidades gubernamentales y medios de comunicación, aquí tienes algunas hipótesis que puedes explorar:

1. **Hipótesis 1: Percepción Pública de la Migración en México**
    - **Descripción**: Analizar cómo se percibe la migración en México según las publicaciones disponibles en el DataFrame.
    - **Análisis**: Realiza un análisis de sentimiento de los comentarios en el DataFrame para identificar el tono general (positivo, negativo, neutral) hacia la migración. Puedes usar técnicas de procesamiento de lenguaje natural (NLP) en Python, como `VADER` o `TextBlob`, para evaluar el sentimiento.

Hipótesis 2: Prevalencia de Comentarios Negativos sobre Migración

- **Descripción**: Investigar la cantidad de comentarios negativos en comparación con
- los comentarios positivos y neutrales.
- **Análisis**: Clasifica los comentarios como negativos, positivos o neutrales y calcula los porcentajes correspondientes.

Hipótesis 3: Opiniones sobre Políticas y Acciones Gubernamentales

- **Descripción**: Examinar cómo las entidades gubernamentales y medios de comunicación expresan sus opiniones sobre las políticas y acciones relacionadas con la migración.
- **Análisis**: Realiza un análisis de temas para identificar las políticas y acciones gubernamentales más mencionadas en el DataFrame.

Hipótesis 4: Diferencias Regionales en la Opinión sobre Migración

- **Descripción**: Analizar si hay diferencias en la percepción de la migración según la región geográfica mencionada en las publicaciones.
- **Análisis**: Extrae y clasifica los datos por región para observar las variaciones en el sentimiento o los temas discutidos.