Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 59-68
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Ingeniería de software
Recibido: 05/10/2023 | Aceptado: 22/12/2023 | Publicado: 30/03/2024

Identificadores persistentes:
DOI: 10.48168/innosoft.s15.a123
ARK: ark:/42411/s15/a123
PURL: 42411/s15/a123

Clasificación de comentarios suicidas en Reddit

Ranking Suicidal Comments on Reddit

Aron Josue Hurtado Cruz [1] [0009-0002-0423-1730]*, **Isabel Karina Ttito Campos** [2] [0009-0008-9159-7581]

[1] Universidad La Salle, Arequipa, Perú. ahurtadoc@ulasalle.edu.pe
[2] Universidad La Salle, Arequipa, Perú. ittitoc@ulasalle.edu.pe

* Autor para correspondencia: ahurtadoc@ulasalle.edu.pe

Resumen

El proyecto se enfoca en el desarrollo de un algoritmo de Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) diseñado para detectar comentarios suicidas en la plataforma Reddit y posteriormente realizar un análisis de sentimientos negativos con el propósito de brindar apoyo a los usuarios que puedan encontrarse en riesgo de suicidio. Para lograr este objetivo, el proyecto combina conceptos y técnicas de inteligencia artificial, procesamiento de lenguaje natural y psicología/psiquiatría.
Para evaluar la eficiencia del proyecto aplicamos la métrica F1 obteniendo un resultado bastante aceptable respecto a una clasificación textual.

Palabras clave: BERT, clasificación, NLP, depresión, suicidio.

Abstract

The project focuses on the development of a Natural Language Processing (NLP) algorithm designed to detect suicidal comments on the Reddit platform and subsequently perform a negative sentiment analysis for the purpose of providing support to users who may be at risk of suicide. To achieve this goal, the project combines concepts and techniques from artificial intelligence, natural language processing and psychology/psychiatry.
To evaluate the efficiency of the project we applied the F1 metric obtaining a fairly acceptable result with respect to a textual classification.

Keywords: *BERT, classification, depression, NLP, suicide.*

Introducción

El Procesamiento de Lenguaje Natural es esencial para analizar el contenido de los comentarios, identificando signos de advertencia relacionados con la ideación suicida. Además, los aportes de la psicología y la experiencia clínica contribuyen significativamente a una comprensión matizada de estas señales de advertencia, asegurando una identificación más precisa de los usuarios en riesgo.

En este contexto, es relevante señalar que investigaciones anteriores han contribuido significativamente al desarrollo de algoritmos que abordan la detección de la ideación suicida. Yeskuatov et al. (2022) aprovecharon Reddit para la detección de la ideación suicida, proporcionando una revisión exhaustiva de las técnicas de aprendizaje automático y procesamiento de lenguaje natural [1]. Además, estudios realizados por Aldhyani et al. (2022) [2], Tadesse et al. (2019) [3], Awatramani et al. (2021) [4], Pal et al. (2018) [5] y Rahat et al. (2019) [6] han explorado diversas metodologías, desde el aprendizaje profundo hasta el análisis de sentimientos, enriqueciendo el panorama de herramientas disponibles para tareas críticas como esta. Estas referencias constituyen una base sólida para el desarrollo y la evaluación de nuestro algoritmo.

Trabajos relacionados

En esta sección, revisaremos algunos enfoques y estudios relacionados que se han centrado en problemas similares al propuesto en este proyecto.

Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques.

Este artículo proporciona una revisión de la literatura de los artículos recientes que detallan los métodos de aprendizaje automático y procesamiento del lenguaje natural (NLP) utilizados para detectar la ideación suicida en Reddit. Los autores identifican tres enfoques principales:

- **Enfoques basados en reglas:** Estos enfoques utilizan un conjunto de reglas predefinidas, basadas en palabras y frases clave asociadas a la ideación suicida, para identificar publicaciones y comentarios que pueden indicar un riesgo de suicidio.
- **Enfoques basados en el aprendizaje automático:** Estos enfoques utilizan algoritmos de aprendizaje automático para entrenar un modelo que pueda predecir la probabilidad de que una publicación o comentario

indique ideación suicida. Los modelos de aprendizaje automático se entrenan en un conjunto de datos de publicaciones y comentarios etiquetados como suicidas o no suicidas.

- **Enfoques híbridos:** Estos enfoques combinan elementos de los enfoques basados en reglas y en el aprendizaje automático.

Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models

Propone un enfoque para detectar y analizar la ideación suicida en las redes sociales utilizando modelos de aprendizaje profundo y aprendizaje automático.

Los autores del artículo utilizaron un conjunto de datos de publicaciones de Reddit que habían sido etiquetadas como suicidas o no suicidas. El conjunto de datos también incluía información sobre el historial de publicaciones de los usuarios.

Los autores utilizaron dos modelos para predecir la probabilidad de que una publicación fuera suicida:

- Un modelo de aprendizaje profundo basado en una red neuronal recurrente (RNN).
- Un modelo de aprendizaje automático basado en el algoritmo XGBoost.

Los resultados de la evaluación mostraron que el modelo de aprendizaje profundo pudo predecir la probabilidad de que una publicación fuera suicida con una precisión del 95 por ciento.

Detection of Depression-Related Posts in Reddit Social Media Forum

Este Artículo investiga el uso del lenguaje en Reddit para identificar mensajes que podrían indicar que un usuario está deprimido. Para ello utilizaron una combinación de técnicas de procesamiento del lenguaje natural (NLP) y aprendizaje automático para entrenar un modelo que pudiera clasificar los mensajes como relacionados o no relacionados con la depresión.

Los investigadores encontraron que los mensajes relacionados con la depresión tendían a utilizar más palabras relacionadas con la tristeza, la ansiedad, la culpa y la desesperanza. También encontraron que los mensajes relacionados con la depresión tendían a tener una mayor proporción de palabras negativas y una menor proporción de palabras positivas.

El modelo entrenado por los investigadores pudo clasificar correctamente los mensajes relacionados con la depresión con una precisión del 80 por ciento.

Detection of Suicidality among Opioid Users on Reddit: A Machine Learning Based Approach

El modelo entrenado pudo clasificar correctamente los mensajes suicidas con una precisión del 82 por ciento. Los investigadores identificaron varias características lingüísticas que estaban asociadas con el suicidio, como el uso de palabras relacionadas con la muerte, la desesperanza y la ideación suicida. También encontraron que los mensajes suicidas tendían a ser más cortos y menos coherentes que los mensajes no suicidas.

Sentiment Analysis of Mixed-Case Language using Natural Language Processing

El artículo propone un enfoque basado en el aprendizaje automático para el análisis de sentimientos de textos en lenguaje mixto. El enfoque utiliza una combinación de técnicas de NLP, como la detección de idiomas y la traducción automática, para preprocesar los textos en lenguaje mixto antes de aplicar un modelo de aprendizaje automático para clasificar los textos en categorías de sentimiento (positivo, negativo o neutral).

Materiales y métodos o Metodología computacional

En esta sección, se proporciona una descripción detallada del diseño de la investigación y la implementación de la metodología computacional para llevar a cabo el proyecto de clasificación de comentarios suicidas en Reddit. Se explica cómo se recopilaron los datos, se preprocesaron, se diseñaron los modelos y se evaluaron.

Recopilación de datos

Para llevar a cabo el proyecto, se utilizaron dos conjuntos de datos preexistentes:

- **Conjunto de datos de detección de suicidio y depresión en Reddit:** Este conjunto de datos se obtuvo de la plataforma Kaggle y contiene comentarios etiquetados como suicidas y no suicidas. Los comentarios fueron recopilados de la plataforma Reddit y se proporcionan con etiquetas que indican la presencia de ideación suicida.

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 59-68
https://revistas.ulasalle.edu.pe/innosoft

- **Conjunto de datos de sentimientos negativos:** Este conjunto de datos se obtuvo de Hugging Face a través de su API. Contiene comentarios etiquetados con sentimientos negativos, como tristeza, miedo y enojo. Este conjunto de datos se utilizó para realizar un análisis de sentimientos en combinación con la detección de ideación suicida.

Ambos conjuntos de datos se utilizaron en la fase de entrenamiento del modelo y en la evaluación de su rendimiento.

Preprocesamiento de datos

Antes de utilizar los datos en la fase de entrenamiento, se aplicaron diversas técnicas de preprocesamiento para limpiar y preparar los textos. Estas técnicas incluyeron:

- **Tokenización:** División de los comentarios en palabras o tokens.
- **Eliminación de stopwords:** Eliminación de palabras comunes que no aportan significado.
- **Lematización:** Reducción de palabras a sus formas base.
- **Eliminación de caracteres especiales:** Eliminación de caracteres no alfabéticos y símbolos.

El preprocesamiento aseguró que los textos fueran representativos y estuvieran listos para la fase de entrenamiento.

Diseño del Modelo

1. Pipeline

El diseño de la propuesta puede verse reflejado en el siguiente pipeline:

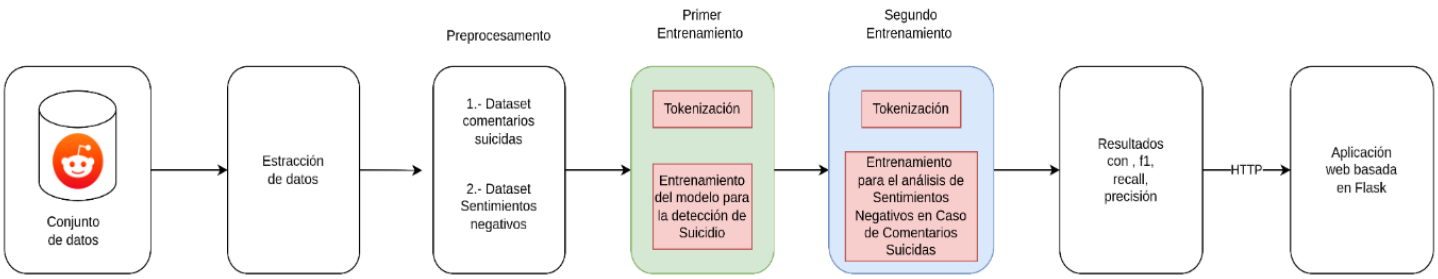

Figura 1. Pipeline descriptivo de la propuesta.

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 59-68
https://revistas.ulasalle.edu.pe/innosoft

2. Entrenamiento del Modelo

Algorithm 1: Entrenamiento del Modelo

Data: Datos de entrenamiento y prueba
Result: Modelo entrenado y resultados del entrenamiento

```
while épocas do
    Inicializar métricas de entrenamiento;
    Inicializar métricas de validación;
    for cada época do
        Iniciar temporizador de época;
        // Entrenamiento
        for cada lote en datos de entrenamiento do
            Procesar lote;
            Calcular pérdida y actualizar pesos;
            Actualizar métricas de entrenamiento;
        end
        Calcular métricas promedio de entrenamiento;
        // Validación
        for cada lote en datos de validación do
            Procesar lote;
            Calcular pérdida y métricas de validación;
        end
        Calcular métricas promedio de validación;
        Registrar resultados y métricas en archivo;
    end
    Guardar modelo y tokenizer;
end
```

Figura 2. Entrenamiento del Modelo

3. Ventajas

- **Contexto Bidireccional:** BERT es capaz de capturar el contexto bidireccional en el que se encuentra una palabra en una oración, lo que le permite comprender mejor el significado de las palabras en relación con su contexto.
- **Transferencia de Conocimiento:** BERT se entrena en grandes cantidades de datos no etiquetados, lo que le permite capturar patrones lingüísticos generales. Esto hace que sea efectivo para tareas específicas con conjuntos de datos más pequeños, ya que el modelo ya tiene un conocimiento lingüístico general.

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 59-68
https://revistas.ulasalle.edu.pe/innosoft

- **Rendimiento de Estado del Arte:** BERT ha demostrado superar a muchos modelos previos en una variedad de tareas de procesamiento del lenguaje natural, incluyendo traducción automática, respuesta a preguntas y análisis de sentimientos.
- **Facilidad de Uso:** BERT está disponible preentrenado y se puede ajustar fácilmente para tareas específicas mediante el entrenamiento con datos adicionales.
- **Modelo de Atención:** BERT utiliza un mecanismo de atención que le permite asignar pesos diferentes a diferentes partes de la entrada, lo que ayuda a capturar relaciones complejas.

4. Desventajas

- **Requisitos Computacionales:** BERT es un modelo grande y complejo que requiere recursos computacionales significativos. Su implementación puede ser costosa en términos de tiempo y potencia de cálculo.
- **Memoria Necesaria:** Debido a su tamaño, BERT puede requerir una cantidad considerable de memoria, lo que puede limitar su uso en dispositivos con recursos limitados.
- **Interpretabilidad:** Dada la complejidad del modelo, entender y explicar las decisiones de BERT puede ser un desafío. La interpretabilidad del modelo es un área en la que se está trabajando activamente.
- **No Considera el Orden de las Palabras:** Aunque BERT es capaz de capturar el contexto bidireccional, no tiene en cuenta el orden exacto de las palabras en una oración, ya que su estructura de atención no es posicional.

Resultados y discusión

1. Comparativa

Figura 3. Detección de comentarios suicidas

Modelo	SVM	LSTM	Naive Bayes(Bag of words)	Naive Bayes (TF-IDF)	Propuesto 10° Época
Precisión	87.82	88.63	96.70	98.40	94.00
Accuracy	89.51	90.12	91.12	90.12	94.00
Recall	91.41	91.71	85.16	81.60	94.00
F1 score	89.58	90.14	90.56	89.20	94.00

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 59-68
https://revistas.ulasalle.edu.pe/innosoft

Figura 4. Análisis de sentimientos negativos

Modelo	Propuesto
Precisión	97.00
Accuracy	98.00
Recall	97.00
F1 score	98.00

(1)

- **SVM**: Es un algoritmo de aprendizaje automático supervisado que se utiliza para clasificación binaria y multiclase. Se basa en la idea de encontrar un hiperplano que separe los datos de los dos o más grupos.
- **LSTM**: Es una arquitectura de red neuronal recurrente que se utiliza para tareas de PLN, como la traducción automática, el reconocimiento de voz y la generación de texto. Se basa en la idea de utilizar una celda de memoria para almacenar información a lo largo del tiempo.
- **Naive Bayes**: Se basa en la idea de que la probabilidad de que una instancia pertenezca a un cierto grupo es proporcional a la probabilidad de que se produzcan los atributos de la instancia dado que pertenece a ese grupo.

2. Resultados

Para todos estos resultados se usaron data sets equilibrados de 30K de datos, entre los que más resaltan es el entrenamiento con BERT y Naive Bayes (TF-IDF), aunque el método de NB es más rápido al sacar resultados BERT puede superar estos resultados con el entrenamiento exacto sin sobre entrenarlo.

En cuanto a este trabajo que se realizó en BERT, se usó el modelo de "bert-base-uncase" ya que nuestros recursos para trabajar este proyecto fueron muy limitados, al no contar con una tarjeta de video usamos google colab (T4 GPU) el cual nos prestaba su GPU por unas 6 horas promedio, limitándonos a usar data sets de 30K de datos por 10 epocas, cuando quisimos trabajar con otros modelos como "bert-large-uncase" o "bert-large-uncase-enmascared" que son más eficientes no pudimos ya que el T4 GPU de colab nos limitaba a solo usar "bert-base-uncased" y así no pudimos obtener los resultados deseados.

Conclusiones

- **Potencial Impacto en la Salud Mental:** El proyecto presenta una aplicación práctica de la tecnología para mejorar la salud mental al prevenir posibles tragedias. La detección temprana de usuarios en riesgo, incluso si

no expresan abiertamente sus pensamientos suicidas, abre la puerta para la intervención oportuna de profesionales de la salud mental.

- **Desafíos Computacionales:** Aunque BERT ha demostrado ser una herramienta poderosa, su implementación puede ser computacionalmente costosa y requerir recursos significativos de hardware. Este desafío se debe tener en cuenta al considerar la escalabilidad y la implementación práctica del modelo.
- **Resultados Aceptables:** Los resultados obtenidos, especialmente en términos de precisión, recall y F1 score, indican que el modelo propuesto tiene un rendimiento aceptable. La aplicación de métricas como F1 score es crucial en tareas de clasificación de texto para evaluar el equilibrio entre precisión y recall.

Contribución de Autoría

Aron Josue Hurtado Cruz : Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Isabel Karina Ttito Campos:** Conceptualización, Investigación, Metodología, Análisis formal, Recursos, Visualización, Supervisión, Administración de proyectos, Adquisición de fondos, Curación de datos, Escritura, revisión y edición.

Referencias

[1] Yeskuatov E, Chua SL, Foo LK. Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques. Int J Environ Res Public Health. 2022 Aug 19;19(16):10347. doi: 10.3390/ijerph191610347. PMID: 36011981; PMCID: PMC9407719.

[2] Aldhyani, Theyazn H. H., Saleh Nagi Alsubari, Ali Saleh Alshebami, Hasan Alkahtani, and Zeyad A. T. Ahmed. 2022. "Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models" International Journal of Environmental Research and Public Health 19, no. 19: 12635. https://doi.org/10.3390/ijerph191912635.

[3] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," in IEEE Access, vol. 7, pp. 44883-44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[4] P. Awatramani, R. Daware, H. Chouhan, A. Vaswani and S. Khedkar, "Sentiment Analysis of Mixed-Case Language using Natural Language Processing," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 651-658, doi: 10.1109/ICIRCA51532.2021.9544554.

[5] S. Pal, S. Ghosh, and A. Nag, “Sentiment Analysis in the Light of LSTM Recurrent Neural Networks,” Int. J. Synth. Emot., vol. 9, pp. 33–39, 2018.

[6] A. M. Rahat, A. Kahir, and A. K. M. Masum, “Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset,” in 8th Int. Conf. Syst. Model. Adv. Res. Trends, 2019, pp. 266–270.