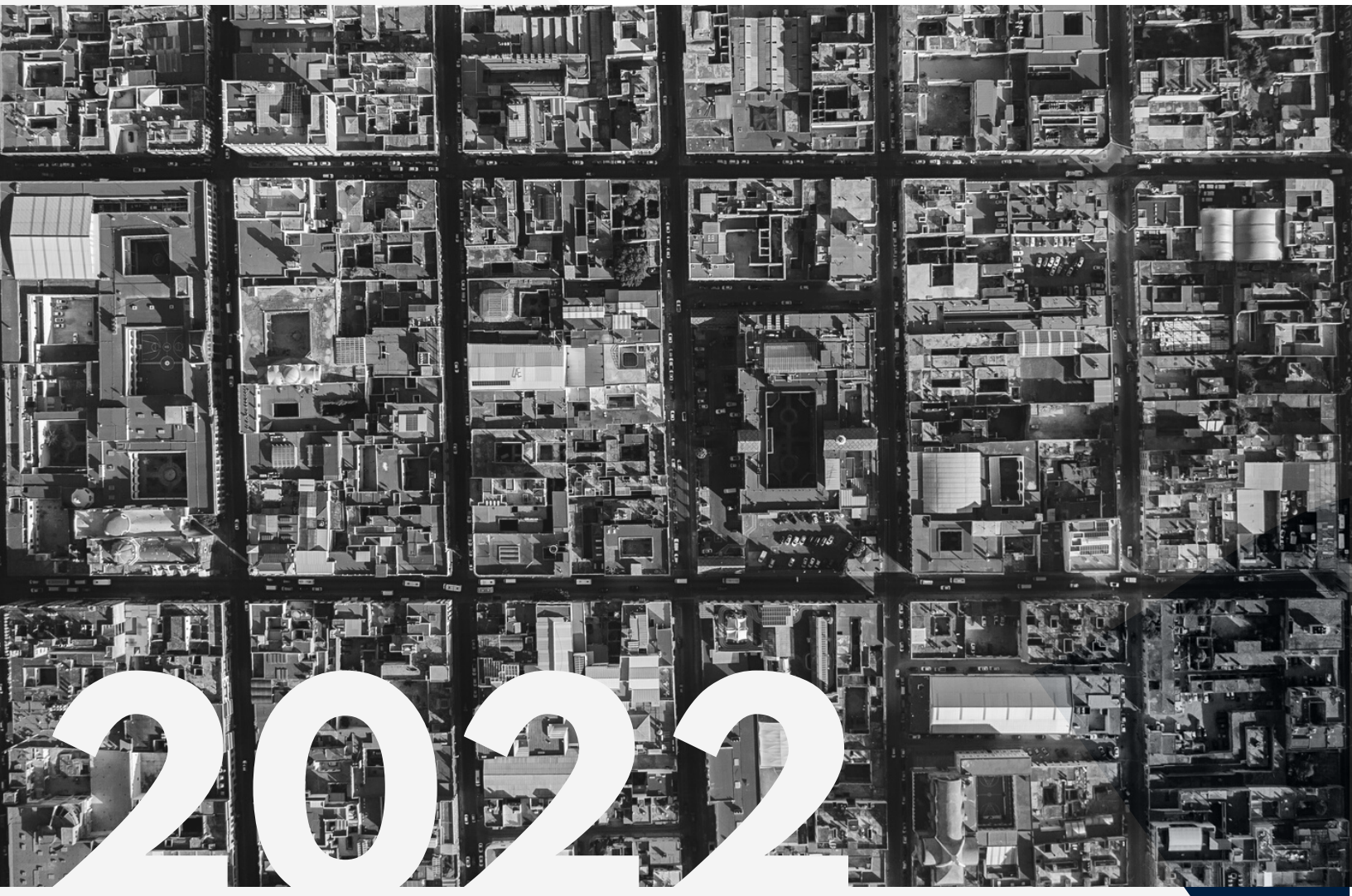
DOCUMENTO

Pysentimiento - Ficha Técnica

Jessica Viridiana Gómez Otero A01378414
Itzel Themis Vargas Tejada A01378973
Romina Rivas Escobar A01378229
Emilio Villacis Mora A01750148



2022

Índice

Pysentimiento - Ficha Técnica

¿Qué es?

- Un conjunto de herramientas multilingüe que proporciona modelos basados en transformaciones de vanguardia para el análisis de sentimientos y el análisis de emociones listos para usar, para idioma español e inglés.

Pysentimeinto tiene base en el repositorios de ***Hugging Face*** y ***transformers***, mismo que se explican a continuación:

Hugging Face

Hugging Face es una comunidad de código abierto con bibliotecas de Transformers que están respaldadas por PyTorch y TensorFlow. Estas bibliotecas proporcionan miles de modelos pre entrenados para ajustarlos de acuerdo con nuestros requisitos.

Transformers

Transformers son parte de los repositorios de Hugging Face (startup que genera diferentes paquetes y módulos en ciencia de datos de NLP). Son uno de los repositorios más utilizados de hugging, que proporciona miles de modelos y API pre-entrenados para descargar y usar rápidamente esos modelos para obtener mejores resultados usando nuestros conjuntos de datos.

Pysentimiento es una especie de modelo para la clasificación de textos proporcionado por transformers. Transformers se centra principalmente en el procesamiento del lenguaje natural. Algunos de los modelos proporcionados por los transformadores son muy fáciles y confiables para realizar tareas de NLP como clasificación, extracción de información, respuesta a preguntas, traducción en más de 100 idiomas.

Descargas

La librería dispone de más de 91,095 en el último mes y más de 200 mil a nivel general.

Información adicional

La librería fue creada en 2021 y se encuentra en GitHub, se puede importar a partir del siguiente link: https://github.com/pysentimiento/pysentimiento

Autores

- Juan Manuel Pérez bio: https://finiteautomata.github.io/
- Juan Carlos Giudici
- Franco Luque

Tipos de Análisis

Sentiment Analysis (Spanish, English)

Otorga una clasificación del texto entre: Positivo, Negativo y Neutral con sus respectivas probabilidades.

Ejemplo:

returns AnalyzerOutput(output=POS, probas={POS: 0.998, NEG: 0.002, NEU: 0.000})

Emotion Analysis (Spanish, English)

Otorga una clasificación del texto determinando su posible emoción en el contexto: Sorpresa, Diversión, Disgusto, Tristeza, Miedo, Ira y otros con sus respectivas probabilidades.

Ejemplo:

returns AnalyzerOutput(output=joy, probas={joy: 0.723, others: 0.198, surprise: 0.038, disgust: 0.011, sadness: 0.011, fear: 0.010, anger: 0.009})

emotion_analyzer.predict("fuck off")

Hate Speech Detection (Spanish, English)

Otorga una clasificación del texto determinando un posible escrito de odio en el contexto

Ejemplo:

returns AnalyzerOutput(output=[], probas={hateful: 0.022, targeted: 0.009, aggressive: 0.018})

¿Cómo está construido?

Pysentimiento es un una librería que utiliza modelos pre-entrenados de transformers (mencionado anteriormente) para distintas tareas de SocialNLP. Usa como modelos bases a BETO (Nacido a partir de BERT) y RoBERTuito en Español, y BERTweet en inglés. Estos son modelos de Machine learning e IA.

El modelo desarrollado por el equipo de pysentimiento fue entrenado con Tweets, por lo que está especialmente diseñado para analizar texto proveniente de redes sociales.

Los mencionados modelos fueron creados por Google y Facebook respectivamente, son bibliotecas como software gratuito y de código abierto para cualquier persona interesada en utilizarlo con fines de investigación.

El modelo BETO nace a partir de BERT, mismo que se explica a continuación:

Bert es una red neuronal de código abierto desarrollada por Google. Esta tecnología permite entrenar un sistema para mejorar y aumentar la capacidad para responder preguntas. Este modelo de lenguaje "enmascarado", cuenta con una arquitectura que supera todos los puntos de referencia del PNL.

Bert posee la característica de ser "bidireccionalidad", lo cual consiste en analizar una oración en dos direcciones. Es decir, analiza las palabras que se encuentran tanto a la izquierda como a la derecha de una palabra clave, y esto permite entender en profundidad el contexto y la temática de toda la frase que introduce un usuario.

Los modelos pre-entrenados con enmascaramiento de palabras completas, tienen una estructura y un vocabulario idénticos a los modelos originales.

La red neuronal está compuesta por:

• *BERT-Large, Uncased (Whole Word Masking):*

- 24 capas,
- 1024 ocultos,
- 16 cabezales,
- parámetros 340M.

• *BERT-Large, Cased (Whole Word Masking):*

- 24 capas,
- 1024 ocultos,
- 16 cabezales,
- parámetros 340M.

Modelo RoBERTa

Implementada en Facebook, basada en Bert con una metodología de entrenamiento mejorada, un 1000% más de datos y potencia de cálculo.
Para mejorar el procedimiento de entrenamiento, elimina la tarea de predicción de siguiente oración, como funcionaba Bert e introduce un enmascaramiento dinámico para que el token enmascarado cambie durante el tiempo de entrenamiento. Usa los tamaños por lotes más grandes para mejorar el procedimiento de entrenamiento.
https://www.ibidemgroup.com/edu/bert-nlp-machine-translation/

¿Cómo funcionan los modelos de lenguaje enmascarado?

El objetivo del entrenamiento del Modelo de lenguaje enmascarado (MLM) es ocultar una palabra en una oración y luego hacer que el programa prediga qué palabra se ha ocultado (enmascarado) según el contexto de la palabra oculta. El objetivo del entrenamiento de predicción de la siguiente oración es hacer que el programa prediga si dos oraciones dadas tienen una conexión secuencial lógica o si su relación es simplemente aleatoria.

¿Por qué se seleccionó trabajar con Pysentimiento?

Principalmente por una clara comparativa con diversos modelos de clasificación:

Modelos	IDIOMA	Que predice	Matriz	Precisión
sentiment-analysis	ES	NEGATIVE / POSITIVE	Predicted NEGATIVE POSITIVE Actual NEG 1698 309 NEU 2862 897 POS 171 62	78.57%
finiteautomata/bertweet-base-sentiment-analysis	ES	NEG / NEU / POS	Predicted NEG NEU POS Actual NEG 68 1914 25 NEU 50 3649 60 POS 2 194 37	62.5%
sagorsarker/codeswitch-spaeng-sentiment-analysis-lince	ES	Label 0 = Negativo Label 1 = Positivo	Predicted LABEL_0 LABEL_1 Actual NEG 756 1251 NEU 901 2858 POS 7 226	43.8%
j-hartmann/sentiment-roberta-large-english-3-classes	INGLÉS (Tweets traducidos)	negative / neutral / positive	Predicted negative neutral positive Actual NEG 1128 823 56 NEU 1030 2550 179 POS 6 142 85	62.7%
pysentimiento	ES	NEU / NEG/ POS	Predicted NEG NEU POS Actual NEG 1757 246 4 NEU 1432 2247 80 POS 14 90 129	68.8%

Accuracy por categoría:
Neg = 87.54%
Pos = 55.36%
Neu = 59.77

Es importante mencionar que a pesar de no tener precisiones altas por categoría de positivo y neutral , la precisión de la categoría de Neg aporta un valor del 87.54%, mismo que apoya al interés y al objeto de estudio del reto proporcionado

Aplicaciones, papers, menciones

https://labtecnosocial.org/opiniones-creencias-vacunas-rrss-bolivia/
https://www.gub.uy/agencia-uruguaya-cooperacion-internacional/sites/agencia-uruguaya-cooperacion-internacional/files/documentos/publicaciones/Informe_CuantificacionViolenciaMujeresPoliticasRS_UY-8Mar.pdf

Uso del Machine Learning en Pysentimiento

Pysentimiento al incorporar modelos preexistentes y pre entrenados por millones de datos, continua aprendiendo conforme el uso. En el caso del modelo BERT desarrollado por Google, va aprendiendo sin supervisión del texto sin etiquetar y mejorando incluso mientras se usa en aplicaciones de Google. Es de esta forma, que sus algoritmos base pueden adaptarse cada vez más en el contenido de búsqueda y consultas.

“permite utilizar una red neuronal basada en la arquitectura de Transformers, para aplicaciones del procesamiento del lenguaje natural. La mayor ventaja de esta herramienta es que permite trabajar con modelos ya entrenados permitiendo analizar los comentarios en español. Logrando así realizar el análisis de Sentimiento, Emociones y Odio.

El modelo aplicado fue entrenado con diversas fuentes de datos. Donde a cada comentario se le asigna una probabilidad de pertenecer a una categoría o de expresar alguna emoción. Por ejemplo, para el tema del Análisis de Sentimiento el comentario: “Esto es pésimo” resulta en un 99.9% como comentario negativo, 0.01 como comentario positivo y un 0.0 neutro. ¿Y en caso del comentario “Que es esto?” presentaría un 0.993 como comentario neutro, 0.005 como negativo y un 0.002 como positivo.”

Peredo, Valeria. (2022). (https://labtecnosocial.org/opiniones-creencias-vacunas-rrss-bolivia/)

Cita del modelo:

```
@misc{perez2021pysentimiento,
    title={pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks},
    author={Juan Manuel Pérez and Juan Carlos Giudici and Franco Luque},
    year={2021},
    eprint={2106.09462},
    archivePrefix={arXiv},
    primaryClass={cs.CL}
```

Pérez, J., Giudici, J., & Luque, F. (2021). pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. arXiv. https://doi.org/10.48550/arXiv.2106.09462