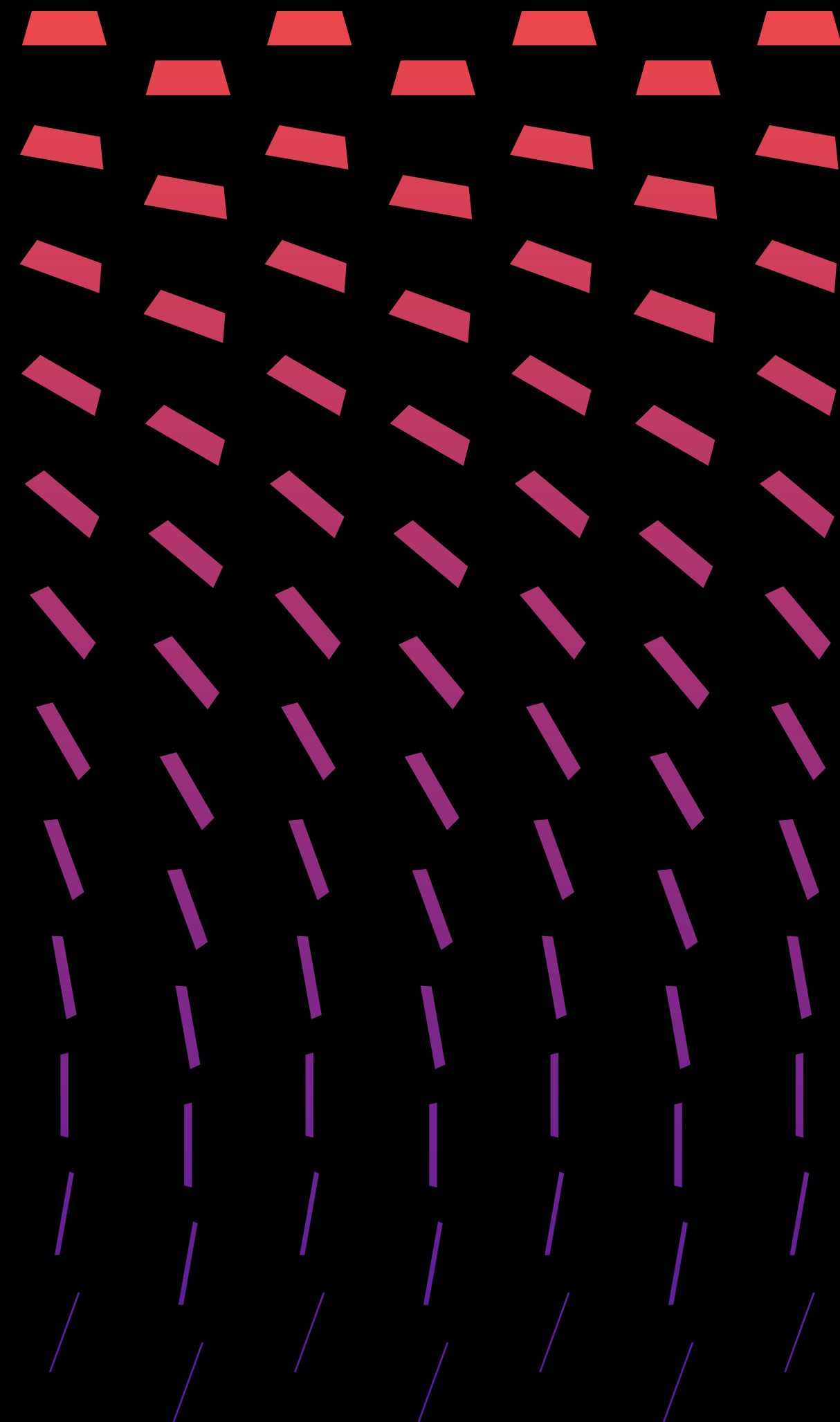


# Detección de Cyberbullying

Diego Hidalgo  
Brian Romero

---



# Contenidos

Introducción y contexto

---

Antecedentes

---

Procesamiento del lenguaje natural

---

Features

---

Limitaciones

---



# Introducción



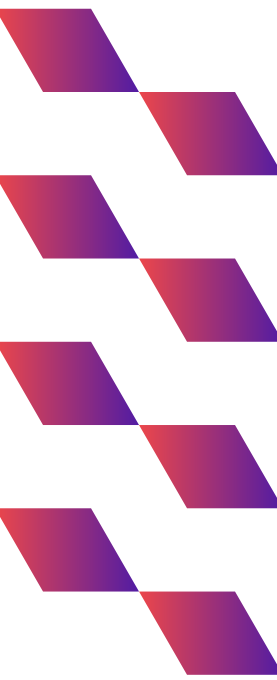
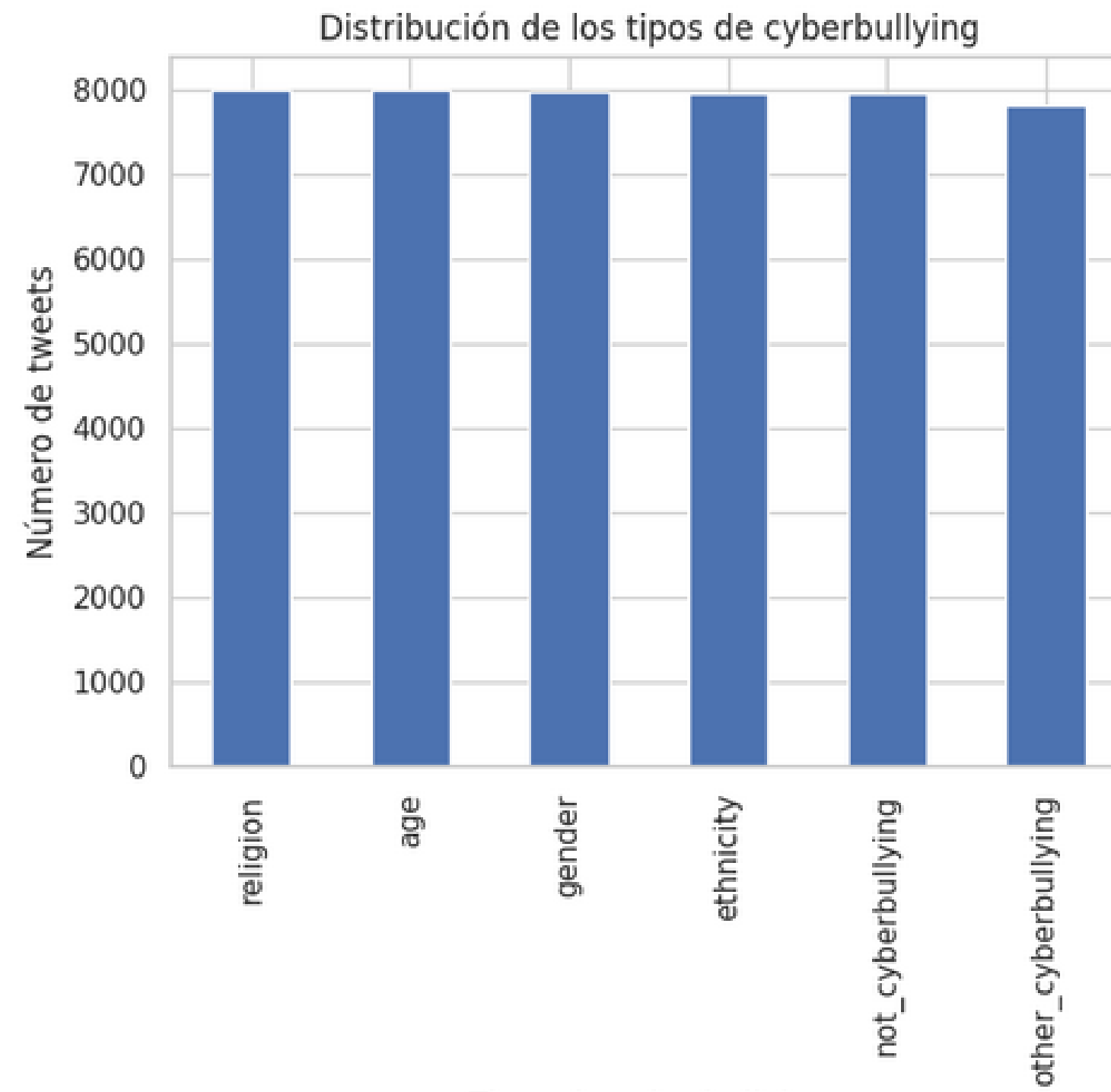
El cyberbullying, es una problemática que afecta cada vez a más personas, especialmente en las redes sociales. El impacto emocional que genera en las víctimas puede ser devastador y, en algunos casos, incluso llevar al suicidio. Por esta razón, resulta fundamental desarrollar herramientas que permitan detectar y prevenir el cyberbullying de manera temprana. Es por eso que hemos diseñado un proyecto que utiliza técnicas de procesamiento de lenguaje natural y aprendizaje automático para analizar mensajes en redes sociales y detectar patrones de comportamiento que indiquen la presencia de cyberbullying.



# Contexto



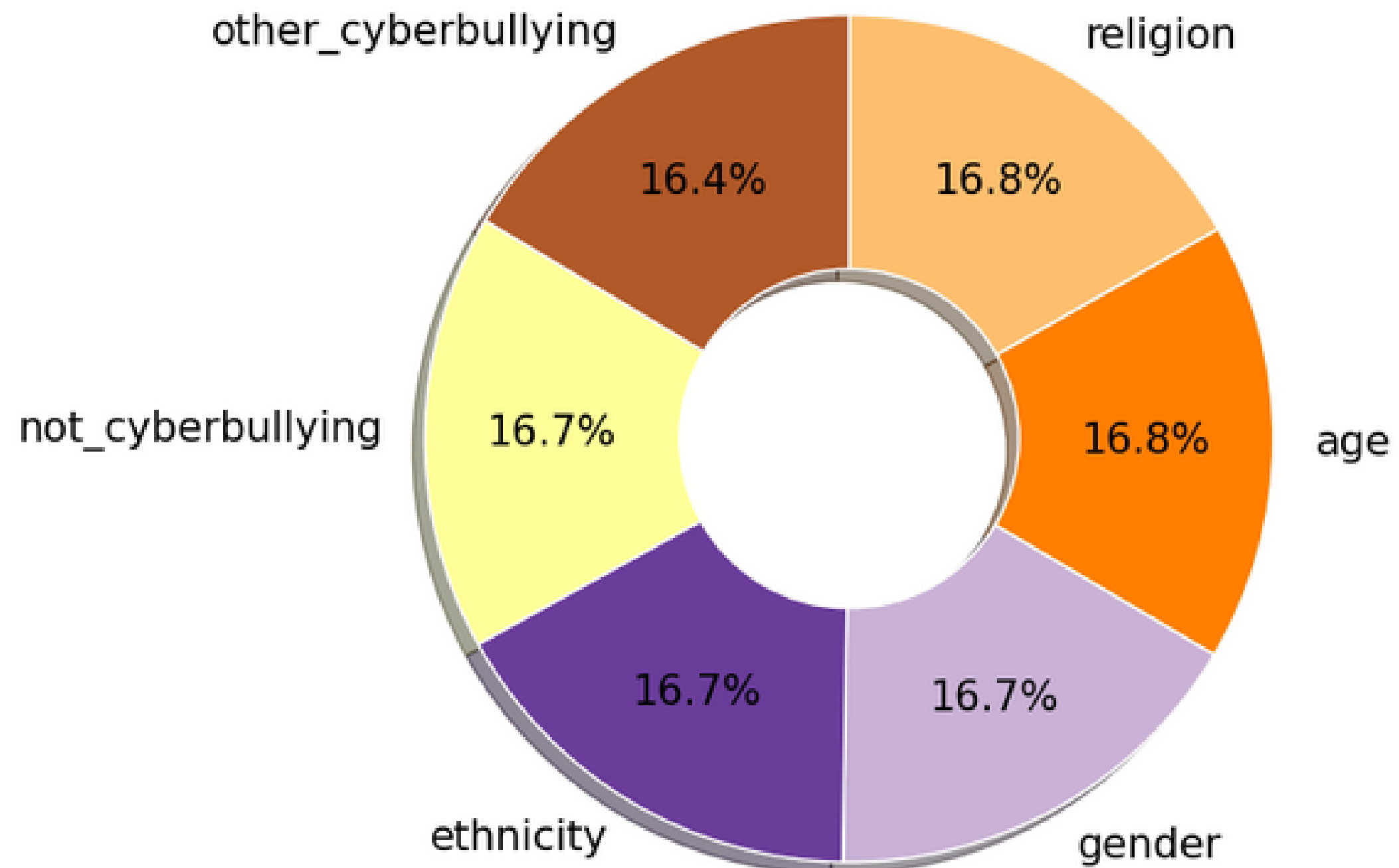
Para la realización de este proyecto hemos utilizado un dataset público que provee la plataforma kaggle. En este dataset se recopilan más de 40 mil tweets en ingles clasificados en distintos tipos de cyberbullying.



# Contexto



Distribución de las clases



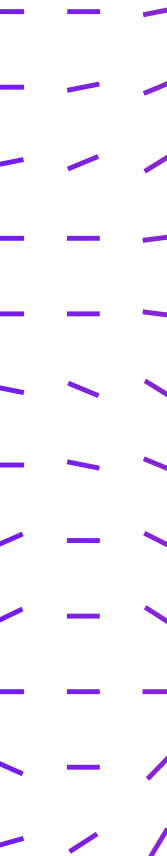
# Antecedentes

# Antecedentes

## Cyberbullying Identification System Based Deep Learning Algorithms



Un proyecto creado por los investigadores Theyazn H. H. Aldhyani, Mosleh Hmoud Al-Adhaileh y Saleh Nagi Alsubari. El proyecto consiste en dos modelos de identificación de Cyberbullying basadas en variaciones de algoritmos de Deep Learning. El primer algoritmo implementado consiste de redes neuronales convolucionales integradas con redes neuronales de memoria a corto y largo plazo bidireccionales (CNN-BiLSTM) y el segundo unicamente utiliza Red Neuronal de Memoria a Corto y Largo Plazo Bidireccional (BiLSTM). Los modelos se entrenaron con más de 200 mil datos recolectados de diferentes redes sociales consiguiendo que el primer modelo lograra una precisión de detección del 94% y el segundo modelo logra un 99% de precisión de detección.

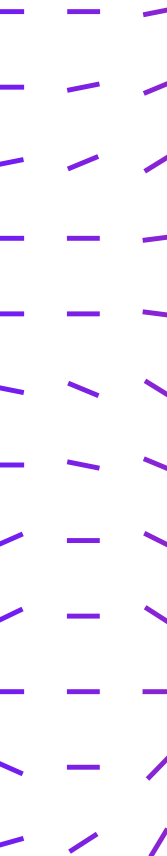


# Antecedentes

## Cyberbullying detection kaggle projects



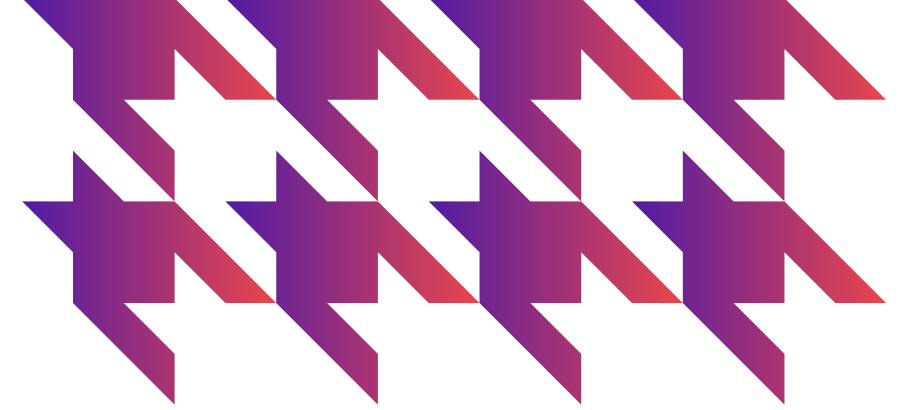
Conjunto de proyectos publicados en kaggle que utilizan principalmente librerías de procesamiento de lenguaje natural y modelos de clasificación como la regresión logística. En general, la precisión de estos modelos se encuentra entre el 78% y 82% de precisión en la identificación del Cyberbullying.





# Procesamiento del lenguaje natural

# Procesamiento del lenguaje natural

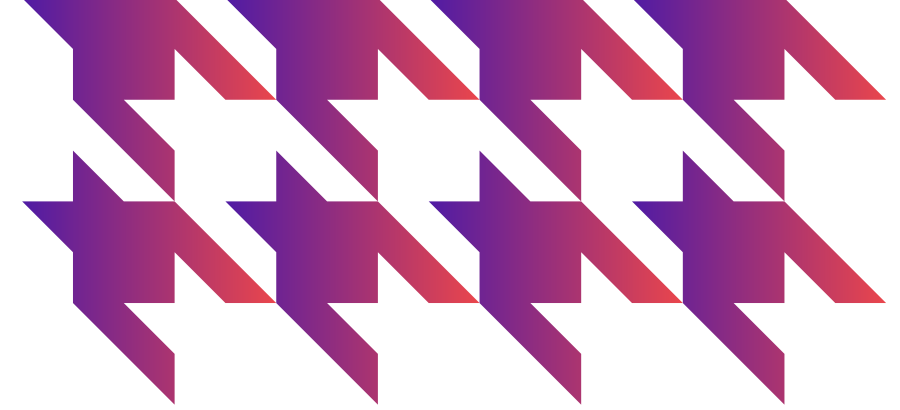


## Librería NLTK



La librería NLTK (Natural Language Toolkit) es una herramienta de procesamiento de lenguaje natural en Python que permite realizar una amplia variedad de tareas de análisis de texto. Esta librería contiene una gran cantidad de recursos para procesar el lenguaje natural, como por ejemplo: tokenizadores de palabras y frases, etiquetadores de partes del discurso, analizadores de sentimientos, y herramientas para la desambiguación de palabras.

# Procesamiento del lenguaje natural



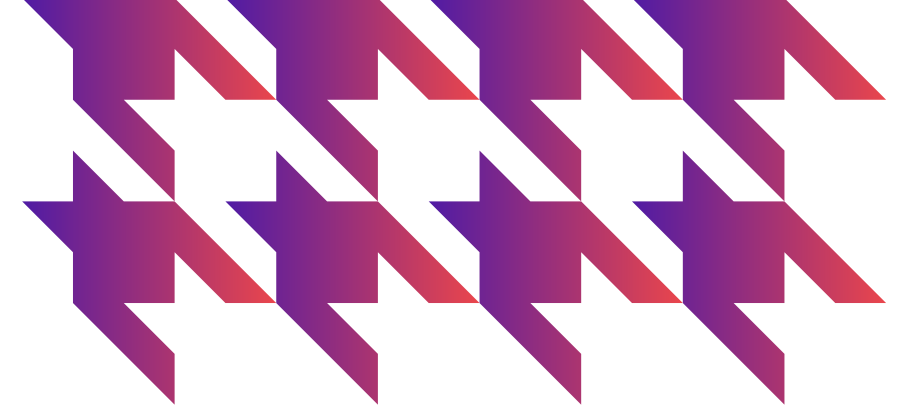
## Librería NLTK: Stopwords



En el procesamiento de lenguaje natural, las palabras de parada o stopwords son palabras comunes que generalmente se eliminan del análisis de texto porque no aportan mucho significado por sí mismas. Algunos ejemplos de palabras de parada en inglés son "a", "an", "the", "in", "of", "and", entre otros.

NLTK incluye una lista de palabras de parada predefinidas para varios idiomas, incluyendo inglés, español, francés, alemán, entre otros. La eliminación de palabras de parada puede ayudar a reducir el ruido en el análisis de texto y mejorar la precisión del modelo.

# Procesamiento del lenguaje natural



## Librería NLTK: PorterStemmer

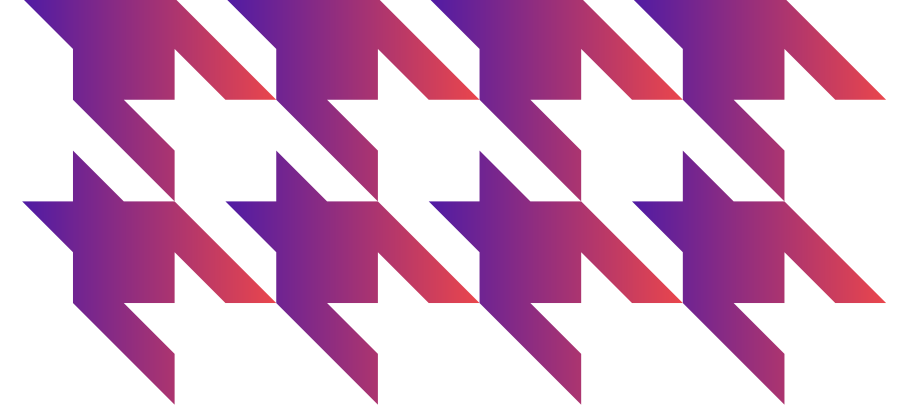


El porterStemmer es un algoritmo de stemming utilizado para normalizar las palabras en un texto. El stemming es el proceso de reducir una palabra a su raíz o base, lo que ayuda a simplificar el análisis de texto y reducir la cantidad de datos que se necesitan para procesar.

En el caso de PorterStemmer, es un algoritmo que aplica una serie de reglas para reducir las palabras a su forma base. Por ejemplo, las palabras "running", "runs", y "ran" se reducirían todas a "run".

Este proceso es útil para aplicaciones como la recuperación de información, análisis de sentimientos y clasificación de texto, ya que permite agrupar diferentes formas de una palabra en una única categoría, lo que puede facilitar el procesamiento del texto.

# Procesamiento del lenguaje natural



## Corpus

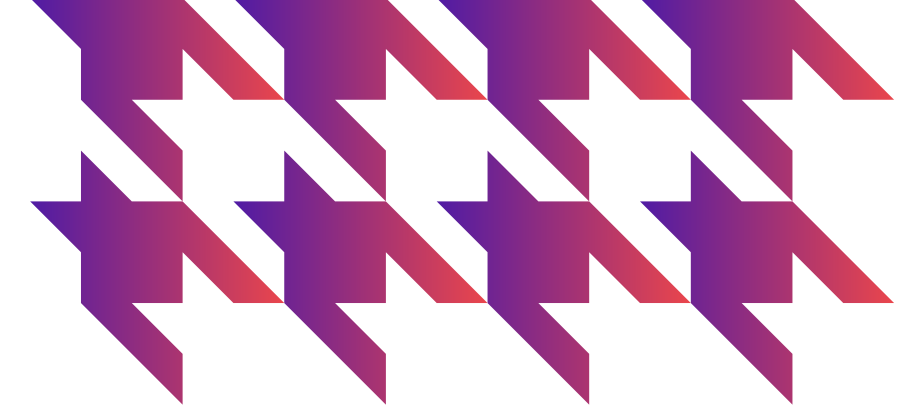


El término "corpus" se refiere a una colección de textos escritos en un determinado idioma o lenguaje que se utiliza como fuente de datos para análisis lingüísticos y de procesamiento de lenguaje natural. En el contexto de la biblioteca NLTK de Python, el corpus se refiere a una colección de textos y recursos lingüísticos que se pueden utilizar para tareas como análisis de sentimientos, clasificación de textos, etiquetado de partes del discurso y más.

En el contexto de nuestro problema el corpus se construye con todos los tweets del dataset.

# Extracción de features

# Extracción de features



## CountVectorizer y fit\_transform



En aprendizaje automático, las características se refieren a los atributos o variables que se utilizan para describir los datos y ayudar a entrenar el modelo.

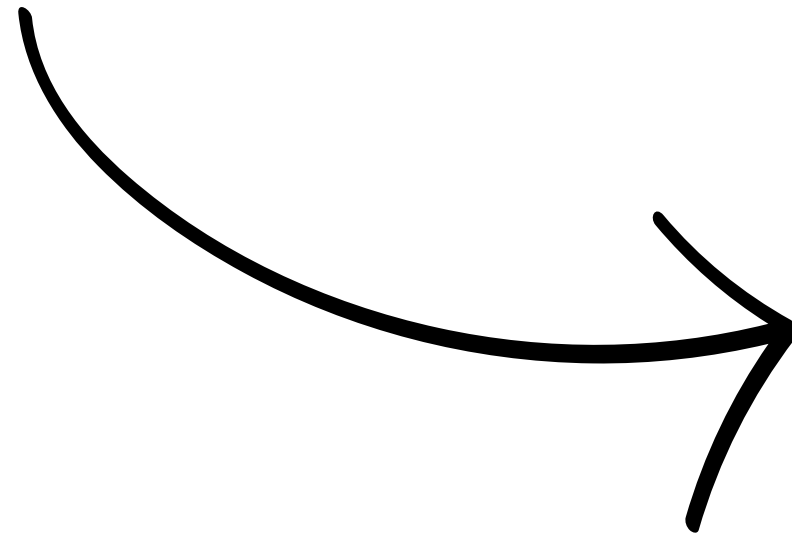
```
# extraer las features
vectorizer = CountVectorizer(analyzer = tweets_cleaner, dtype =
                             'uint8')
X_cv = vectorizer.fit_transform(df['tweet_text'])
```

# Extracción de features

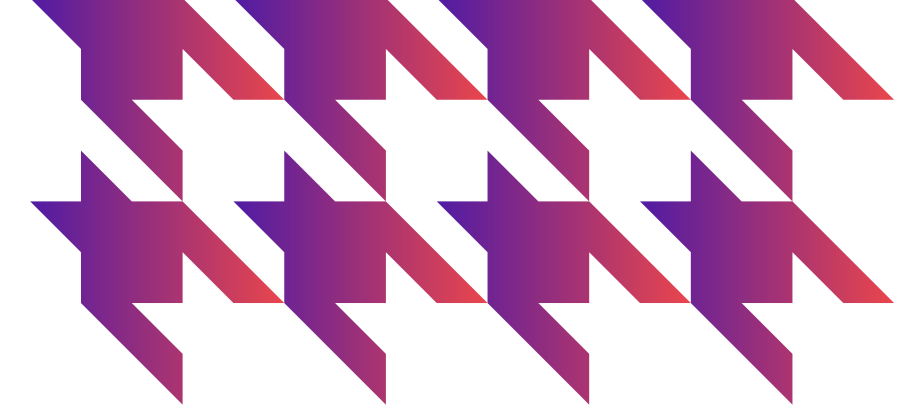
fit\_transform



classi whore red velvet cupcak

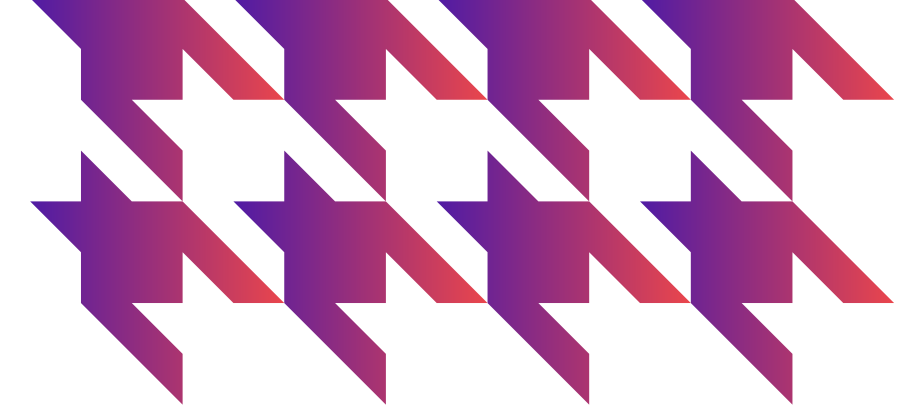


1
1
1
1
1





# Extracción de features



## TfidfTransformer



TfidfTransformer es una herramienta de la biblioteca scikit-learn de Python que se utiliza para transformar la representación numérica de los datos de un conjunto de documentos de texto.

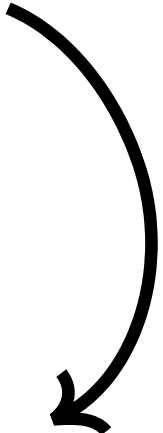
Tfidf significa "Term Frequency-Inverse Document Frequency" (frecuencia de término-frecuencia inversa de documentos), y es una técnica comúnmente utilizada para ponderar la importancia de las palabras en un documento en función de la frecuencia con la que aparecen en ese documento y en otros documentos en el corpus.

# Extracción de features

TfidfTransformer

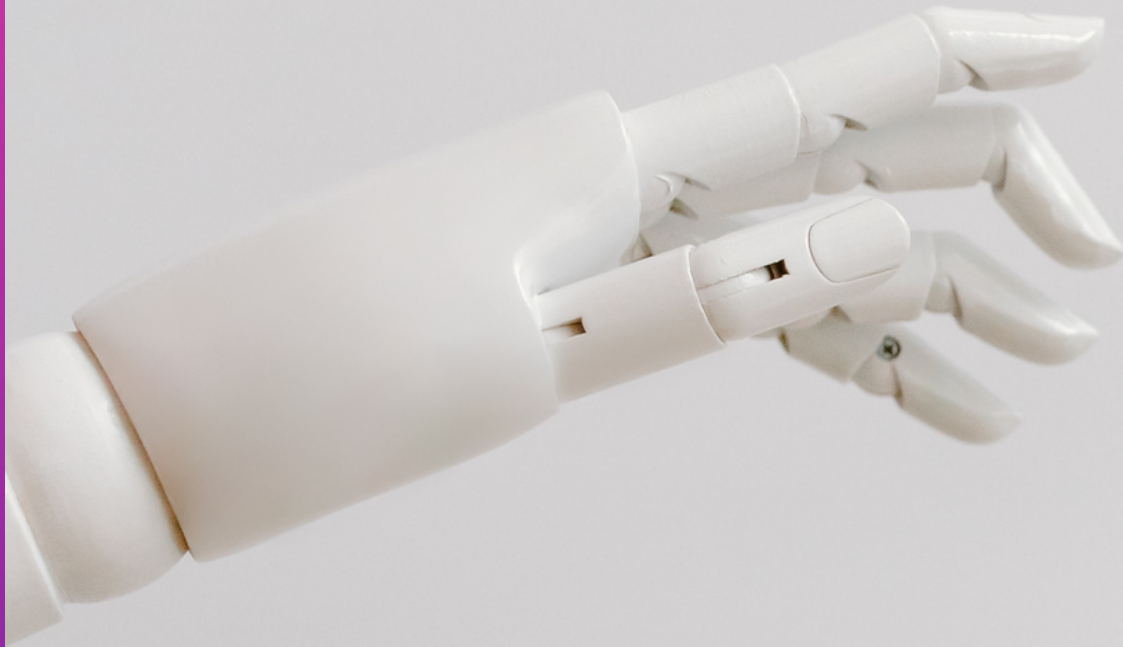


bro u got ta chill rt dog fuck kp dumb nigger

A black curved arrow pointing from the input text 'bro u got ta chill rt dog fuck kp dumb nigger' to the first cell of the table.

	bro	u	gotta	chill	rt	dog	fuck	dumb	nigger
bro u got ta chill rt dog fuck kp dumb nigger	0.2	0.1	0	0.1	0	0.4	0.7	0.8	0.9

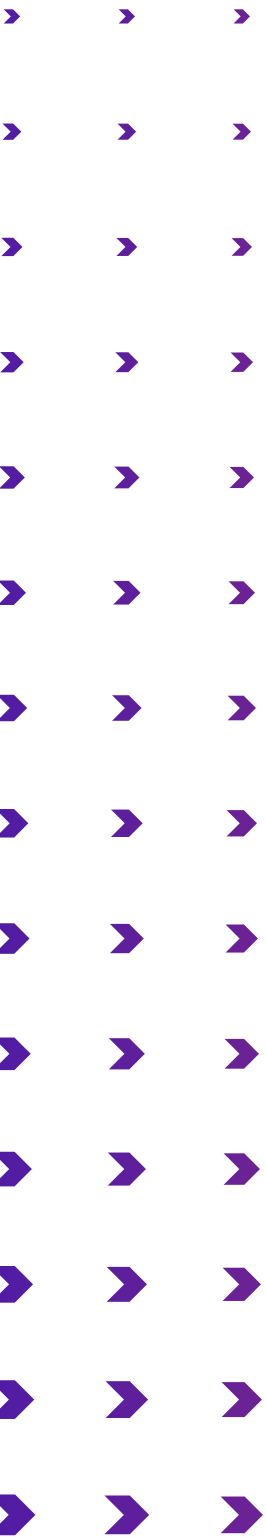
# Limitaciones



# Limitaciones



La principal limitación que hemos identificado en el desarrollo del proyecto es el dataset utilizado. El dataset contiene datos exclusivamente en ingles por lo cuál el modelo entrenado solo es capaz de identificar el cyberbullying en un tweets que se encuentren en el idioma ingles y por tanto desplegarlo para su uso en latino américa no es posible ya que es incapaz de identificar los modismo y demás usos del lenguaje necesarios para detectar el cyberbullying.



# Preguntas



**¿Qué pasos debemos aplicar para procesar los tweets?**

**¿Qué modelos podemos de machine learning podemos utilizar para la clasificación?**

**¿Cómo podemos mejorar la capacidad de predicción del modelo seleccionado?**

# Código

