"No tendría que haber muerto": suicidio de adolescente gay de 13 años conmociona a Francia

Vía: El Universal MX

Una chica se suicida en Jaén por el ciberacoso sufrido por ser lesbiana

Vía: ovejarosa

Actualidad

El ciberacoso provoca el suicidio del influencer trans Paulo Vaz

Vía: MARCA

SUCESOS • Denuncia de los colectivos LGTBI

Una joven de 17 años se quita la vida en Galicia tras sufrir acoso por su orientación sexual

Vía: El Mundo

Vía: Observatorio Andaluz

ETSI
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA
uhu.es

Técnicas de Deep Learning para Identificar y Clasificar Mensajes de Odio en Redes Sociales Dirigidos a la Comunidad LGTBQ+

Antonio Morano Moriña - Javier Román Pásaro (Grupo I^{2}C-UHU)

Septiembre de 2023

Tabla de contenidos

LGBT+Fobia

Discriminación contra la comunidad LGBT+

Comunidad vulnerable a abusos, trastornos y discriminación.

Detección automática y clasificación de discursos de odio

Promoción de la inclusión, el respeto y la igualdad en línea.

01

Descripción de la tarea

IberLEF-2023 HOMO-MEX

Descripción de la Tarea

Subtarea 1: Detección de discursos de odio (Multi clase)

Subtarea 2: Detección de discursos de odio de granularidad fina (Multi etiqueta)

Subtarea 1: Detección de discursos de odio (Multi clase)

Clases

- P (LGBT+fóbico)
- NP (No LGBT+fóbico)
- NA (No relacionado)

7000 instancias
80% Train
14% Valid
6% Test

Formato del conjunto de datos

Indice	Tweet	Clase
104	@neymarjr tiene que jugar en liga de Maricas!	P

Subtarea 2: Detección de discursos de odio de granularidad fina (Multi etiqueta)

Etiquetas

- L (Lesbofobia)
- G (Gayfobia)
- B (Bifobia)
- T (Transfobia)
- O (Otro tipo de LGBT+fobia)

863 instancias
80% Train
14% Valid
6% Test

Formato del conjunto de datos

Tweet	G	L	B	T	O
@edith_gdl jajaja no sufras. No pasa nada, el punto es una prenda. Hay chicas que les encanta travestirse de hombres (se llaman lesbianas)	0	1	0	1	0

Métricas de evaluación

Precision

$$precision = \frac{TP}{TP + FP}$$

Recall

$$recall = \frac{TP}{TP + FN}$$

F1 macro

$$F1\ macro = 2 * \frac{precision * recall}{precision + recall}$$

Matriz de confusión

		Predicciones	
		Positivos	Negativos
Valores Reales	Positivos	Verdaderos Positivos (VP)	Verdaderos Negativos (VN)
	Negativos	Falsos Positivos (FP)	Falsos Negativos (FN)

02

Fases de Desarrollo

Metodología aplicada. Entrenamiento de los modelos

Fases de Desarrollo

01

Baseline

02

Preprocesamiento

03

Aumento de Datos y Búsqueda de Hiperparámetros

04

Enfoque de ensemble

01

Referencia Base

Objetivo

Establecer un punto de referencia inicial y evaluar la viabilidad del problema

Modelos preentrenados

- dccuchile/bert-base-spanish-wwm-uncased (BETO)
- PlanTL-GOB-ES/roberta-base-bne (RoBERTa)
- xlm-roberta-base (XLM)

Hugging Face

01

Referencia Base

Objetivo

Establecer un punto de referencia inicial y evaluar la viabilidad del problema

Subtarea 1 (Multiclase)

Modelo	F1 Macro
	Baseline
BETO	0,8172
RoBERTa	0,8011
XLM	0,8227

01

Referencia Base

Objetivo

Establecer un punto de referencia inicial y evaluar la viabilidad del problema

Subtarea 2 (Multietiqueta)

Modelo	F1 Macro
	Baseline
BETO	0,6412
RoBERTa	0,6318
XLM	0,6128

02

Preprocesamiento

Objetivo

Estandarizar los mensajes y eliminar el posible ruido

Limpieza de datos

Tweet original	Ser valiente ponga #Foto y no Sea #Maricon yo tengo mi foto @MarceTyS aunque también algo tapada pero me pongo con ud url
Tweet procesado	ser valiente ponga foto y no sea maricon yo tengo mi foto aunque también algo tapada pero me pongo con ud

Dictionary 

02

Preprocesamiento

Objetivo

Estandarizar los mensajes y eliminar el posible ruido

Subtarea 1 (Multiclase)

Modelo	F1 Macro	
	Baseline	Preprocessing
BETO	0,8172	0,8281
RoBERTa	0,8011	0,8197
XLM	0,8227	0,8228

02

Preprocesamiento

Objetivo

Estandarizar los mensajes y eliminar el posible ruido

Subtarea 2 (Multietiqueta)

Modelo	F1 Macro	
	Baseline	Preprocessing
BETO	0,6412	0,6503
RoBERTa	0,6318	0,6726
XLM	0,6128	0,6539

03

Aumento de Datos y Búsqueda de Hiperparámetros

Objetivo

Equilibrar las clases y encontrar los mejores hiperparámetros para mejorar el rendimiento

Back-translation

Original	Cuando los hombres dan a luz en una sociedad pervertida y **degenerada** dominada por **transgéneros**
Primera traducción	When men give birth in a perverted and **degenerate** society dominated by **transgenders**
Back-translation	Cuando los hombres dan a luz en una sociedad pervertida y **depravada** dominada por **transexuales**

50% de las instancias positivas

03

Aumento de Datos y Búsqueda de Hiperparámetros

Objetivo

Equilibrar las clases y encontrar los mejores hiperparámetros para mejorar el rendimiento

Espacio de Hiperparámetros

Hiperparámetro	Valores
Batch Size	[16, 32, 64]
Learning Rate	[2e-5, 3e-5, 5e-5]
Max. Length	[64, 128, 256]
Weight Decay	[0.001, 0.01, 0.1]

W&B

03

Aumento de Datos y Búsqueda de Hiperparámetros

Objetivo

Equilibrar las clases y encontrar los mejores hiperparámetros para mejorar el rendimiento

Mejores hiperparámetros por modelo

Hiperparámetro	BETO	RoBERTa	XLM
Batch Size	32	32	16
Learning Rate	5e-5	3e-5	2e-5
Max. Length	128	128	256
Weight Decay	0,01	0,01	0,01

03

Aumento de Datos y Búsqueda de Hiperparámetros

Objetivo

Equilibrar las clases y encontrar los mejores hiperparámetros para mejorar el rendimiento

Subtarea 1 (Multiclase)

Modelo	F1 Macro		
	Baseline	Preprocessing	Data Aug. e Hiperparámetros
BETO	0,8172	0,8281	0,8566
RoBERTa	0,8011	0,8197	0,8451
XLM	0,8227	0,8228	0,8228

03

Aumento de Datos y Búsqueda de Hiperparámetros

Objetivo

Equilibrar las clases y encontrar los mejores hiperparámetros para mejorar el rendimiento

Subtarea 2 (Multietiqueta)

Modelo	F1 Macro		
	Baseline	Preprocessing	Data Aug. e Hiperparámetros
BETO	0,6412	0,6503	0,6674
RoBERTa	0,6318	0,6726	0,6960
XLM	0,6128	0,6539	0,6714

04

Enfoque de ensemble

Objetivo

Incrementar la robustez y fiabilidad.
Aprovechar la diversidad de enfoques

Subtarea 1 (Multiclase)

Modelo dominante	Valor F1 (Media Macro)
BETO	0,8325

Subtarea 2 (Multietiqueta)

Modelo dominante	Valor F1 (Media Macro)
RoBERTa	0,6960

03

Análisis de Resultados

Exposición y comparación de resultados finales

Competición

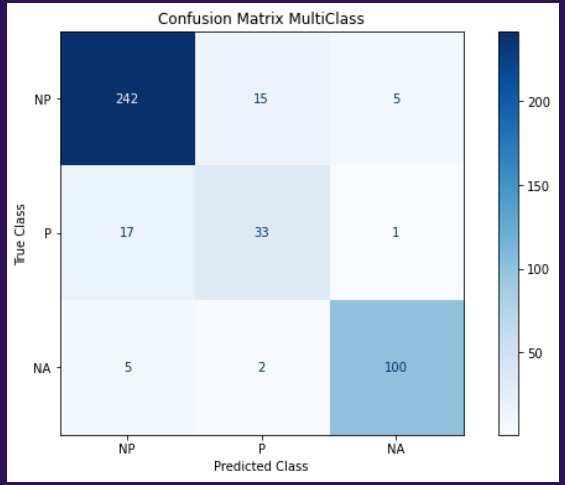
Subtarea 1

Posición	Usuario	Valor F1 Predicción
1	bayesiano98	0,8847
2	carfer	0,8432
-	-	-
6	**I²C-UHU**	**0,8325**
-	-	-
11	moeintash	0,7326

Subtarea 2

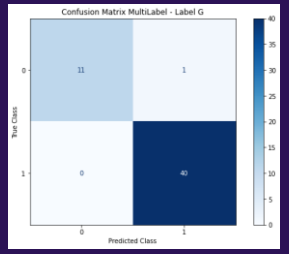
Posición	Usuario	Valor F1 Predicción
1	**I²C-UHU**	**0,6960**
2	carfer	0,6847
-	-	-
9	cesar_m	0,6550

Matrices de confusión

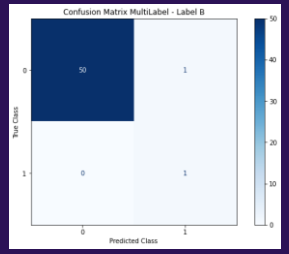


Matrices de confusión

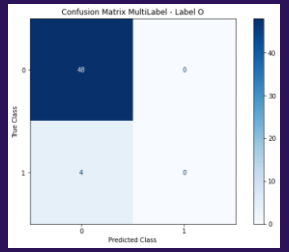
G



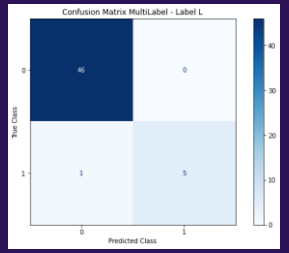
B



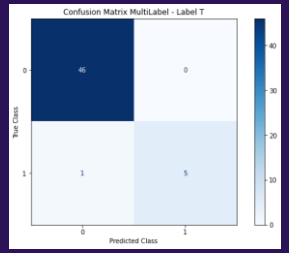
O



L



T



Análisis de errores

Tweet	Clase	Predicción
Puto el primero que se contagie del coronavirus!	P	NA
Ese ruido que hacían los Transformers en la serie animada al transformarse, que no tenía nada que ver con la transformación.	NA	P
Ser homosexual es una actitud frente a la cama, ser puto es una actitud frente a la vida.	NP	P

Análisis de errores

Tweet	Etiquetas	Predicción
O mejor "todos", q incluye femenino, masculino, transgénero, homosexual, bisexual y lo q esta semana agregue la corrección política.	[0,0,0,0,1]	[1,1,0,1,1]
Los vatos sacan el lado marica y las morras el lado sharmuta.	[1,0,0,0,0]	[1,1,0,0,0]
Yo le hacia el cambio de sexo gratis a #Daniel por maldito joto cobarde #YoNoCreoEnLosHombres.	[1,0,0,0,0]	[1,0,0,1,0]

04

Mejoras y conclusiones

Mejoras desarrolladas. Conclusiones y posibles usos

Mejoras desarrolladas

Undersampling (Subtarea 1)

Muestreos de los conjuntos con distintas proporciones

Bootstrap

Generación de conjuntos de datos adicionales

Mejoras desarrolladas

Undersampling

Distribución de clases

Proporción			Train Dataset			Valid Dataset			Test Dataset		
P	NP	NA	P	NP	NA	P	NP	NA	P	NP	NA
1	5	2	690	3488	1422	121	610	249	51	262	107
1	1	1	690	690	690	121	121	121	51	51	51
1	2	1,3	690	1380	897	121	242	158	51	102	67
1	3,5	1,6	690	2415	1104	121	424	197	51	179	82

Mejoras desarrolladas

Undersampling

Resultados (F1 Macro)

Modelo	Proporciones (P:NP:NA)			
	Original	1:1:1	1:2:1.3	1:3.5:1.6
BETO	0,8566	0,7724	0,7909	0,8246
RoBERTa	0,8451	0,7348	0,7758	0,8110
XLM	0,8228	0,7327	0,7671	0,8002

Mejoras desarrolladas

Bootstrap

Resultados (F1 Macro)

Modelo	Atributos y valores			
	Tamaño subconjuntos	Subconjuntos analizados	Subtarea 1	Subtarea 2
BETO	80	18	0,814	0,6162
RoBERTa	100	20	0,823	0,6178
XLM	200	5	0,773	0,5992

Posibles usos

Moderación de contenidos, Alertas y notificaciones para identificar y filtrar la incitación al odio y la discriminación.

Recogida de datos. Identificar tendencias y pautas específicas de discriminación

Generar información y recursos educativos sobre los retos a los que se enfrenta la comunidad LGBTQ+.

Apoyo emocional a las personas afectadas

Antonio José Morano Moriña

www.linkedin.com/in/antoniojosemoranomor/

antoniojmorano@gmail.com

Javier Román Pásaro

www.linkedin.com/in/javierromanpasaro/

javier.roman.pasaro@gmail.com

ETSI
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA
uhu.es

Técnicas de Deep Learning para Identificar y Clasificar Mensajes de Odio en Redes Sociales Dirigidos a la Comunidad LGTBQ+

Antonio Morano Moriña - Javier Román Pásaro (Grupo I²C-UHU)

Septiembre de 2023