SHARED TASK 3: Multi-class classification of effects of outdoor spaces on social anxiety symptoms in Reddit.

La tarea compartida pertenece al 'SMM4H-2024 The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks — Large Language Models (LLMs) and Generalizability for Social Media NLP' organizado por la Association for Computational Linguistics (ACL).

Preguntas de Investigación:

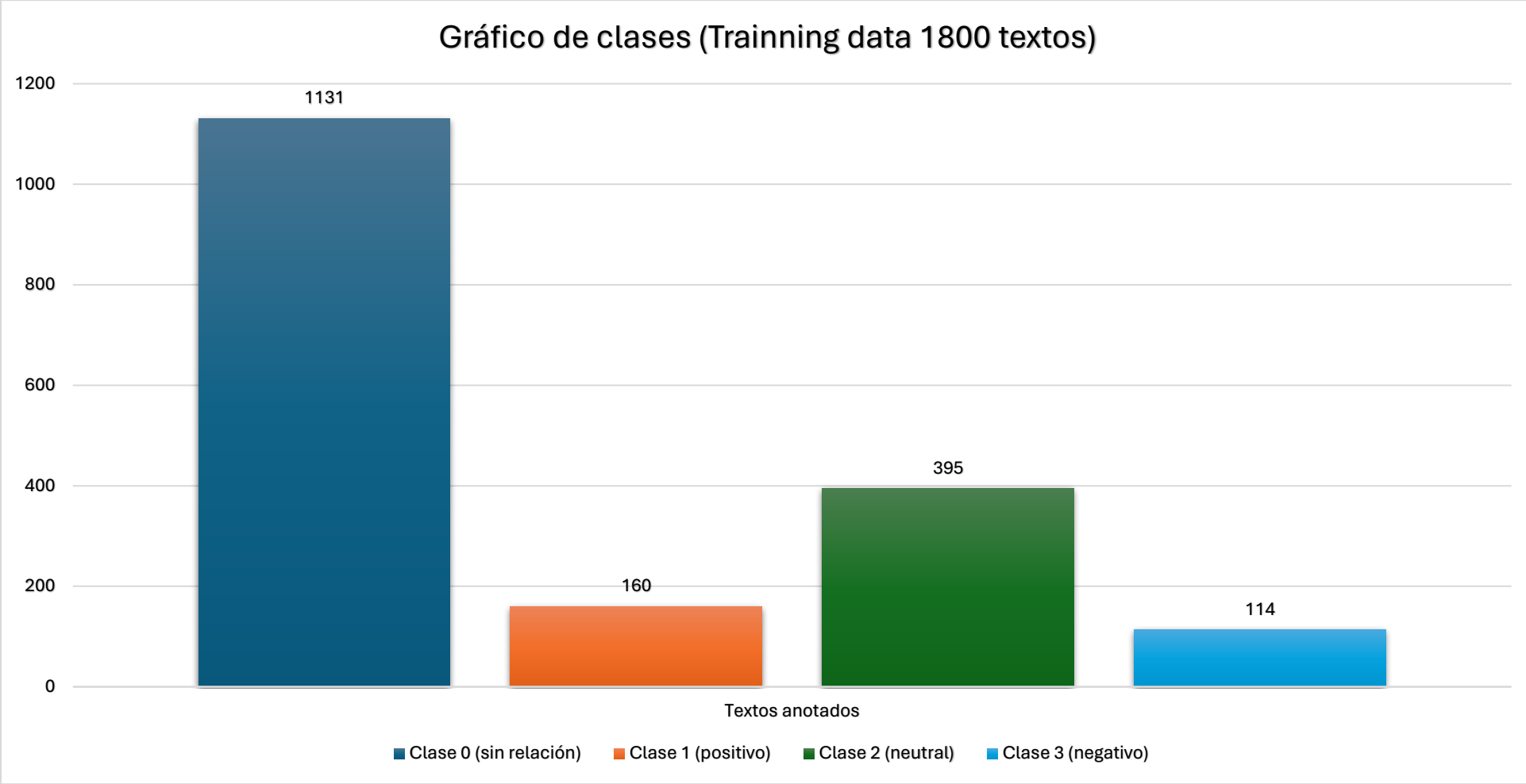
- ¿Cuál es la distribución de las clases en el conjunto de datos de entrenamiento?
- ¿Cuál es la longitud promedio de los textos en cada una de las clases?
- ¿Cuál es la distribución de la longitud de los textos en el conjunto de datos?
- ¿Cuál es la frecuencia de las keywords mencionadas en los textos?
- ¿Cuál es la relación entre la longitud del texto y la clase a la que pertenece?
- ¿Existen diferencias significativas en la frecuencia de palabras entre las clases?

Datos del Corpus:

- Metadatos:
  - Total de textos: 3000
  - Textos para entrenamiento: 1800
  - Textos para validación: 600
  - Textos para prueba: 600
  - Número de clases: 4 (1=positive effect, 2=neutral or no effect, 3=negative effect, 4=unrelated)

- Edad de usuarios: 12-25 años

- Estadísticas, características y exploración de los datos.



ESTADÍSTICAS ANÁLISIS EXPLORATORIO DE DATOS TRAINING DATA (1800 textos)					
MÉTRICA	**DATASET COMPLETO**	**CLASE 0 (unrelated)**	**CLASE 1 (positive effect)**	**CLASE 2 (neutral or no effect)**	**CLASE 3 (negative effect)**
Promedio de palabras por textos	239.19988895058302	211.342175066313	196.0875	296.5848101265823	379.3421052631579
Longitud promedio de palabras	4.232379748394994	4.261957983762495	4.272794280238815	4.171931403527651	4.204795713328868
Promedio de palabras únicas por texto	110.04509283819628	110.04509283819628	105.88125	140.8860759493671	171.76315789473685
Promedio de números/dígitos por texto	2.363131593559134	2.058355437665782 4	1.90625	3.154403797468354	3.307017543859649
Promedio de verbos por texto	47.023320377556802	41.325375773651636	38.7625	58.450632911392404	75.96491228070175
Promedio de sustantivos por texto	38.775124930594412	34.407603890362 51	31.7	48.734177215189874	57.859649122807 02

Promedio de adjetivos por texto	16.313159355913338	14.528735632183908	13.16875	19.708860759493675	26.807017543859655
Palabras más comunes	('I', 17383) ('to', 12705) ('and', 11168) ('the', 8529) ('a', 8051) ('of', 5669) ('my', 5357) ('in', 4267) ('that', 4057) ('you', 3372)	('I', 8952) ('to', 6870) ('and', 6043) ('the', 4962) ('a', 4372) ('of', 3233) ('my', 2737) ('in', 2432) ('that', 2323) ('you', 2198)	('I', 1075) ('to', 1004) ('and', 772) ('a', 663) ('the', 566) ('you', 446) ('of', 391) ('my', 327) ('in', 301) ('that', 298)	('I', 5171) ('to', 3492) ('and', 3275) ('a', 2297) ('the', 2192) ('my', 1650) ('of', 1484) ('in', 1083) ('was', 1031) ('me', 1026)	('I', 2185) ('to', 1339) ('and', 1078) ('the', 809) ('a', 719) ('my', 643) ('of', 561) ('in', 451) ('that', 416) ('me', 405)
Palabras más comunes (sin stopwords)	('like', 2657) ('people', 2257) ('anxiety', 1462) ('get', 1461) ('feel', 1404) ('go', 1314) ('know', 1228) ('social', 1203) ('even', 1202) ('really', 1150)	('like', 1418) ('people', 1269) ('get', 843) ('anxiety', 828) ('feel', 724) ('social', 688) ('go', 642) ('time', 638) ('think', 637) ('really', 635)	('like', 187) ('people', 148) ('go', 145) ('anxiety', 117) ('get', 106) ('really', 98) ('feel', 95) ('would', 93) ('social', 93) ('time', 90)	('like', 766) ('people', 570) ('know', 383) ('feel', 381) ('go', 374) ('get', 354) ('even', 352) ('friends', 328) ('would', 322) ('anxiety', 320)	('like', 286) ('people', 270) ('feel', 204) ('anxiety', 197) ('get', 158) ('go', 153) ('even', 147) ('know', 138) ('social', 124) ('going', 121)
Número de keywords diferentes	86	53	50	48	33

Promedio de keywords por texto	1.2903942254303165	1.1679929266136162	1.575	1.4253164556962026	1.6403508771929824
Frecuencia de keywords	outside: 536 walk: 341 run: 331 running: 229 park: 106 nature: 49 runs: 47 beach: 45 camp: 40 sun: 40 pool: 40 basketball: 40 soccer: 39 bike: 33 swimming: 26 hiking: 24 riding: 21 tree: 21 swim: 18 boat: 18 horse: 16 garden: 16 mountain: 15	run: 280 outside: 260 running: 163 walk: 133 park: 41 runs: 39 basketball: 38 nature: 34 sun: 32 pool: 31 camp: 22 riding: 19 tree: 18 boat: 18 soccer: 14 horse: 13 beach: 13 mountain: 13 swim: 12 swimming: 12 stream: 11 bike: 10 climb: 9	walk: 53 outside: 38 running: 27 run: 21 park: 16 nature: 10 bike: 8 hiking: 6 beach: 5 jog: 4 backyard: 4 biking: 3 Fresh air: 3 trees: 3 backpacking: 3 Go for a walk: 3 hike: 2 runs: 2 jogs: 2 parkour: 2 skating: 2 soccer: 2 outside : 2	outside: 176 walk: 109 park: 40 running: 28 run: 24 beach: 23 soccer: 19 camp: 17 hiking: 15 bike: 12 biking: 10 garden: 8 swimming: 7 sun: 7 nature: 5 pool: 5 runs: 4 jog: 4 mountains: 4 tree: 3 forest: 3 hikes: 3 climb: 3	outside: 62 walk: 46 running: 11 park: 9 swimming: 6 run: 6 swim: 4 soccer: 4 beach: 4 pool: 3 bike: 3 pool: 3 runs: 2 jogging: 2 garden: 2 grass: 2 outside : 2 parkour: 1 lake: 1 jog: 1 riding: 1 camp: 1 basketball: 1

	biking: 14	bench: 9	Hiking: 2	horse: 2	skating: 1

ESTADÍSTICAS ANÁLISIS EXPLORATORIO DE DATOS VALIDATION DATA (600 textos)	
MÉTRICA	**DATASET COMPLETO**
Promedio de palabras por textos	308.22333333333336
Longitud promedio de palabras	4.26373681598603
Promedio de números/dígitos por texto	3.3033333333333332
Promedio de verbos por texto	59.855
Promedio de sustantivos por texto	51.406666666666666
Promedio de adjetivos por texto	21.096666666666668
Palabras más comunes	('I', 7753) ('to', 5347) ('and', 4661) ('the', 3632) ('a', 3455) ('my', 2574) ('of', 2472) ('in', 1852) ('that', 1693)

	('was', 1450)
Palabras más comunes (sin stopwords)	('like', 1173) ('people', 958) ('feel', 656) ('anxiety', 620) ('social', 594) ('get', 581) ('even', 528) ('know', 525) ('time', 524) ('really', 489)
Número de keywords diferentes	123
Promedio de keywords por texto	1.4683333333333333
	outside: 89 walk: 53 run: 49 running: 37 outside : 34

Frecuencia de keywords	bike: 24 beach: 22 bench: 22 run : 21 soccer: 19 basketball: 18 mountain: 18 park: 18 swimming: 17 sun: 17 waves: 14 skate: 13 outdoor: 12 golf: 11 pool: 11

Posibles Estrategias de Solución:

- Modelos de Clasificación: Entrenar modelos de clasificación multiclase utilizando técnicas como Regresión Logística, Support Vector Machines, Random Forest, y Redes Neuronales.
- Procesamiento del Lenguaje Natural (NLP): Utilizar técnicas de NLP para preprocesar los textos, como tokenización, eliminación de stopwords, lematización, y extracción de características relevantes.
- Aprendizaje Profundo: Explorar el uso de modelos de aprendizaje profundo, como Recurrent Neural Networks (RNNs) o Transformers, para capturar mejor la estructura y el contexto de los textos.
- Ensamble de Modelos: Probar estrategias de ensamble para combinar las predicciones de múltiples modelos y mejorar el rendimiento general.
- Ajuste de Hiperparámetros: Realizar búsqueda de hiperparámetros para optimizar el rendimiento de los modelos en el conjunto de validación.
- Validación Cruzada: Utilizar técnicas de validación cruzada para evaluar la generalización de los modelos y evitar el sobreajuste.