Hito 2

"Análisis de comentarios sarcásticos en Reddit"

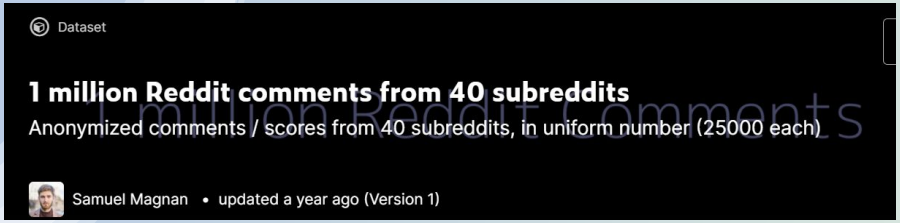
Integrantes:

- Nicolás García
- Pablo Gutiérrez
- Javier Lavados
- Sebastián Salinas
- José Triviño

Mejoras al Hito 1

Adición de nuevos datos:

Se incorporó una nueva base de datos con comentarios no sarcásticos para poder comparar con los comentarios sarcásticos originales. Estos datos fueron preprocesados y unidos en una misma tabla para facilitar el entrenamiento de un clasificador.



subreddit	body	controversiality	score
(categorical) on which subreddit the comment was posted	(str) comment content	(binary) a reddit aggregated metric of how controversial was the comment	(scalar) upvotes minus downvotes
40 unique values	971729 unique values	0 1	-889 35.6k
gameofthrones	Your submission has been automatically removed because all post titles must begin with one hard-brac...	0	1
aww	Dont squeeze her with you massive hand, you mean giant.	0	19

Los datos fueron obtenidos desde Kaggle, y fueron incorporados a los datos que se tenían previamente mediante Excel.

```
yes_or_no = pd.read_excel("DatosHito2.xlsx")
yes_or_no.head()
```

	comment	subreddit	score	sarcasm
0	I don't pay attention to her, but as long as s...	AskReddit	0	yes
1	Trick or treating in general is just weird...	AskReddit	1	yes
2	what the fuck	AskReddit	22	yes
3	This would make me cry.	AskReddit	1	yes
4	My stuffed animal I've had since I was born.	AskReddit	1	yes

Fuente: https://www.kaggle.com/smagnan/1-million-reddit-comments-from-40-subreddits

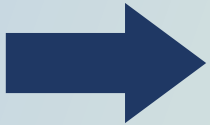
Mejoras al Hito 1

Modificación a las preguntas y problemas:

Se concluyó que varias de las interrogantes planteadas, o bien podían ser respondidas de forma trivial, o eran de un nivel de complejidad mayor al que podía ser respondido con los conocimientos que se tienen hasta el momento

Preguntas Originales

1. ¿Qué tan necesario es el contexto? ¿Es necesaria la presencia del comentario padre? **(Demasiado compleja)**
2. ¿En que subreddits el uso de sarcasmo es proporcionalmente mayor? **(Trivial, no requiere métodos de minería de datos)**
3. ¿Es posible entrenar un clasificador? **(Demasiado general)**



Preguntas Nuevas

1. ¿A través de qué método se debe entrenar un clasificador para que presente un buen desempeño?
2. ¿Qué factores influyen más en el desempeño de un clasificador? ¿Longitud, palabras, puntaje?
3. Si se entrena un clasificador en un subreddit específico, ¿Cambia el desempeño al usar este clasificador en un subreddit distinto?

Propuesta Experimental

Pregunta 1: Método de entrenamiento de un clasificador

- Se segmenta el dataset en distintos subreddits.
- Se dividirán estos segmentos en grupos de training y de testing, para luego entrenar 5 clasificadores distintos.
- Después de entrenar los datos, se comparará el accuracy para determinar el desempeño.

Tipos de clasificadores a utilizar:

- Decision Tree
- Support Vector Machine
- Linear Support Vector Machine
- K-Nearest Neighbours
- Dummy Classifier

Propuesta Experimental

Pregunta 2: Factores que influyen en el desempeño

- Se filtraran los comentarios bajo o sobre una cierta longitud y se entrenará sobre estos datos para determinar si hay cambios en el desempeño
- Se filtraran ciertas palabras con el mismo objetivo de determinar cambios en el desempeño.
- Se eliminarán columnas como el score para analizar si existen cambios en el desempeño del clasificador.

Propuesta Experimental

Pregunta 3: Clasificación dependiente al subreddit

- Se filtrará el dataset para que solo contenga datos correspondientes a AskReddit y WorldNews.
- Se realizará una vectorización de este dataset en conjunto.
- Luego se separará la información en grupos de training y testing.
- Se entrenarán los 5 clasificadores con los datos ya separados de ambos subreddits.
- Se probarán estos clasificadores utilizando datos del subreddit opuesto.
- Finalmente, se analizarán los resultados obtenidos utilizando la métrica de accuracy.

Resultado Preliminar

Se decidió responder a la siguiente pregunta:

¿A través de qué método se debe entrenar un clasificador para que presente un buen desempeño?

Resultado Preliminar

- La proporción de comentarios sarcásticos dentro de un subreddit no influye al momento de entrenar un clasificador. El desempeño se mantiene tanto cuando la proporción de comentarios sarcásticos es del 50% como cuando están desbalanceados.
- El clasificador que presenta el mejor desempeño es el SVMC, arrojando un accuracy cercano al 80% para ambos subreddits evaluados, pero el costo computacional es muy alto. Por otro lado, Linear SVM genera resultados similares en menor tiempo.