Semana 9
CI-0134: Investigación en Ciencias de la Computación
Paula Monge - B85066

Actividades de la semana:

- ☐ Revisión del dataset en español

Se hace la revisión del dataset con la idea de buscar algún patrón de escritura que haga que los algoritmos no puedan hacer bien la clasificación, buscamos el uso de "spanglish", frases sin sentido o algo por ese estilo.
Se encontraron los siguientes casos:

1. Uso de palabras específicas en inglés, ejemplos de tweets:

- "no todo tuyo pero no me des **unfollow**"
- "quesesto y donde esta mi **refollow** mira"
- "18 no sw quw dexir de ti si te di **follow** seria por algo pero ahora mismo"
- "diablo zury comentando mis fotos y todo y no me da **follow back**"
- "chau the **following** no puedo creer que la hayan cancelado todavia esto tiene que seguir"
- "y siguen sin aparecerme mis **followers** estoy llorando"
- "y yo que yo tuve el **follow** cuando tenias 9 000 **followers** ese **follow** valia oro y ya no lo tengo"
- "no di **unfollow** a nadie en serio twitter me dejo sin **following**"
- "a la verga por que lo hacen amo su **makeup"**
- **"**no me gusta poner fotos **fangirls d**e perfil de **whtsapp** bt hoy me ha podido una de harry abrazando a una fan es que socorro odio al mundo**"**
- **"**me pasa con mi **girlfriend** mas si querias arreglar para verla y justo no lo leyo"
- "el mae era un **fuckboy** yo lo miraba guapo"
- "zoy plaga lo ze pero te extrano jajaja te amo **baby** nunca me cambies"
- "os cuento y tal queridos **friends** digitales"
- "un dia totalmente **fitness happy**"
- "hola **happy birthday** muchas felicidades k cumplas muchos anos mas k la pase super un abrazo a la distancia"
-

Entre otros ejemplos, estas palabras en inglés son bastante utilizadas, sin embargo su presencia no evita entender el contexto del tweet. La palabra "follow" tiene 6094 recurrencias

2. Nombres de marcas, plataformas o series:

- "murio mi **instagram"**
- "martes adios telefono nooo que hare sin **whatsapp** una semana sin **instagram facebook** y **twitter** mi musica contactos todo sintelefono"
- "fui al **nike** de jumbo y no estaban las que yo queria jajaj"

- “corri mi mejor distancia con mi perro hoy yoelegicorrer historiasdeunrunner **nikeplus** soycorredora”
- “para cuando la segunda temporada de **gossip girl”**
-

Estos textos que mencionan marcas o plataformas con nombres en inglés también están muy presentes, sin embargo, al igual que les recurrencias anteriores no evitan el entendimiento del contexto del tweet

3. Frases completas en inglés:

- “hablando de eficacia y fotografia **vogue like a painting** muy recomendable”
- “un dia lo vi lo escuche y desde ese dia entro en mi corazon para nunca volver a salir **whyiloveliam**”
- “estan dando **crazy stupid love** en el warner”
- “fav si tienes las notificaciones **on all the love x”**
- “cuanto me gusta bajarme de un avion y leer estas 6 letras a mi alrededor **lovemadrid**”
- “no hay nada como la vida sencilla y plena en la montana **rurallovers** respirandoairepuro”
- “gracias por seguirme en breve te devuelvo **follow peace** amp **love** tuitutil”
- “listos para esquiar aramon panticosa verano2015 **happy loveit** panticosa pirineos”
- **“**59 dias mas de verdad que me muero por escucharlo kanxksmdka quiero escucha**r the girl who cried wolf”**
- **“**22 ella se cree que la tengo olvidada pero oye patitos como oath supremo **miss ma girl”**
- “siempre tambien desgracia **for life** jajajaja”
- **“i need sephora in my life but** maldita devaluacion”
- “**sorry baby** pero es para ver quien nos dice poque”
- “,a mi todos me dejan asi pero no entiendo como **my love from the star** no le fue guau”
- “que falta la de en la juventus **go back home** arquitecto”
- “**home sweet home** de vuelta a la realidad queestres”
- “**happyfriendshipday** francesca”
- “sabes que te amo pase lo que pase no por que internet **friends are te best friends**”
- “**i know that feel sister** ademas si son personas groseras un amigo mio lo tiene peor su madre desprecia los libros y todo”
- “pero que si la esta estudiando su maestria estoy orgullosa de ti fea dale con todo es lunes pero positiva **proudsister**”

Aquí se muestran solo algunos de los tweets que se pueden encontrar, donde es evidente que están mal escritos, con palabras o frases en inglés, parece que los datos no pasan por un proceso de limpieza.

- ☐ Se prueba el algoritmo de Regresión Logística con el nuevo dataset sobre reseñas, estos son los resultados de diversas experimentaciones

Prueba 1.

Hiperparametros:

- max_iters: `4000`
- ngram_range: `(1,2)`
- max_features: `50000`
- C: `1`

Resultados con datos en inglés:

	precision	recall	f1-score	support
negative	0.91	0.90	0.90	4929
positive	0.90	0.91	0.91	4991
accuracy			0.90	9920
macro avg	0.90	0.90	0.90	9920
weighted avg	0.90	0.90	0.90	9920

```
Precision: 0.90
Accuracy: 0.90
Recall: 0.91
F1 Score: 0.91
```

Resultados con datos en español:

	precision	recall	f1-score	support
negative	0.90	0.88	0.89	4929
positive	0.88	0.90	0.89	4991
accuracy			0.89	9920
macro avg	0.89	0.89	0.89	9920
weighted avg	0.89	0.89	0.89	9920

```
Precision: 0.88
Accuracy: 0.89
Recall: 0.90
F1 Score: 0.89
```

Prueba 2.

Hiperparametros:

- max_iters: `1000`
- ngram_range: `(1,2)`
- max_features: `1000`
- C: `1`

Resultados con datos en inglés:

```
              precision    recall  f1-score   support

    negative       0.86      0.85      0.86      4929
    positive       0.86      0.86      0.86      4991

    accuracy                           0.86      9920
   macro avg       0.86      0.86      0.86      9920
weighted avg       0.86      0.86      0.86      9920

Precision: 0.86
Accuracy: 0.86
Recall: 0.86
F1 Score: 0.86
```

Resultados con datos en español:

```
              precision    recall  f1-score   support

    negative       0.84      0.82      0.83      4929
    positive       0.82      0.85      0.84      4991

    accuracy                           0.83      9920
   macro avg       0.83      0.83      0.83      9920
weighted avg       0.83      0.83      0.83      9920

Precision: 0.82
Accuracy: 0.83
Recall: 0.85
F1 Score: 0.84
```

Prueba 3.

Hiperparametros:

- max_iters: `4000`
- ngram_range: `(1,2)`
- max_features: `50000`
- C: `0.01`

Resultados con datos en inglés:

```
              precision    recall  f1-score   support

    negative       0.84      0.79      0.81      4929
    positive       0.80      0.85      0.83      4991

    accuracy                           0.82      9920
   macro avg       0.82      0.82      0.82      9920
weighted avg       0.82      0.82      0.82      9920

Precision: 0.80
Accuracy: 0.82
Recall: 0.85
F1 Score: 0.83
```

Resultados con datos en español:

```
              precision    recall  f1-score   support

    negative       0.84      0.78      0.80      4929
    positive       0.79      0.85      0.82      4991

    accuracy                           0.81      9920
   macro avg       0.81      0.81      0.81      9920
weighted avg       0.81      0.81      0.81      9920

Precision: 0.79
Accuracy: 0.81
Recall: 0.85
F1 Score: 0.82
```

Se están realizando las mismas pruebas que en la experimentación con el dataset anterior para poder ver el comportamiento, además de que se experimenta con datos en inglés.
Si bien se obtienen mejores métricas con los datos en español del nuevo dataset es interesante que, los resultados con las pruebas de inglés tienen un porcentaje (bajo) mejor en sus métricas. Considerando que se trata de traducciones es de interés investigar por qué esa leve diferencia de métricas.

Actividades para la próxima semana:

Todas estas actividades pueden ser realizadas por mi compañero o por mi.

1. Investigación sobre los distintos hiperparametros de los 3 algoritmos
2. Más experimentación con los datos