EMO Project

Regresión de intensidad emocional cross-lingüe (EMO)

Pablo García Quintas

Abstract

En esta tarea de regresión para determinar la intensidad de las emociones en textos cortos, como tweets, enfrentamos la limitación de la falta de datos etiquetados en español y en inglés. Para abordar este desafío, proponemos dos estrategias de transferencia de conocimientos.

La primera estrategia implica utilizar la traducción automática para traducir el conjunto de datos de evaluación en español al inglés. Posteriormente, evaluamos un modelo entrenado en el conjunto de entrenamiento en inglés en estos textos traducidos. Esta aproximación se basa en la premisa de que las estructuras subyacentes de las emociones son universales y transferibles entre idiomas, permitiendo que un modelo entrenado en inglés generalice adecuadamente a textos en español.

La segunda estrategia consiste en utilizar la traducción automática para convertir el conjunto de entrenamiento y desarrollo en inglés al español. Luego, evaluamos un modelo entrenado en el conjunto traducido en textos en español. Esta estrategia busca adaptar el modelo a las particularidades del español, considerando las diferencias culturales y lingüísticas que puedan influir en la expresión de las emociones.

Ambas estrategias aprovechan la capacidad de la traducción automática para mitigar la falta de datos etiquetados en español e inglés. La evaluación de modelos en textos traducidos busca preservar la capacidad del modelo para capturar patrones emocionales relevantes, independientemente del idioma. Estas estrategias proporcionan enfoques prácticos para superar las barreras idiomáticas y mejorar la aplicabilidad de modelos de regresión de emociones, incluso en ausencia de conjuntos de datos específicos para este idioma.

---

1 Introducción:

Este proyecto se inicia con la adquisición de archivos fundamentales alojados en el repositorio de GitHub proporcionado por los profesores: https://github.com/jerbarnes/hp-emo

La descarga de estos archivos no solo implica la obtención de código, la documentación adjunta a estos archivos actúa como una guía esencial que desentraña la complejidad del código, proporcionando información detallada sobre la funcionalidad y el propósito de cada componente.

---

2 Objetivo Z1:

Para comenzar con el proyecto, haciendo uso de la web LibreTranslate, he realizado las distintas traducciones de los archivos que se encuentran en las carpetas “en” y “es”, que a su vez se encuentran dentro de la carpeta “datasets”. He creado dos carpetas distintas, una para cada estrategia:

Estrategia 1: Utilizar la traducción automática, traducir el conjunto de datos de evaluación en español al inglés y evaluar un modelo entrenado en el conjunto de entrenamiento en inglés.

En este caso he creado la carpeta “Español -> Ingles”

Estrategia 2: Utilizar la traducción automática, traducir el conjunto de entrenamiento y desarrollo en inglés al español y evaluar un modelo entrenado en el conjunto de traducido en textos en español.

En este caso he creado la carpeta “Ingles -> Español”

Por otro lado he modificado el código que nos ha sido proporcionado de manera que podemos elegir desde la terminal, al ejecutar el programa, tanto la emoción que queremos evaluar, como el idioma que queremos que tengan la data del train y del dev por un lado y la data del test por el otro.

```
if __name__ == '__main__':
    print("Elige de que emoción se evaluará entre las que hay: anger, joy, sadness, fear")
    emocion = input()

    print("Elige idioma del train/dev data: ingles->1 o castellano->2")
    idiomaTrain = input()
    if idiomaTrain == "1":
        aux1 = "../dataset/en"
    else:
        aux1 = "../dataset/Ingles -> Español"

    print("Elige idioma del test data: ingles->1 o castellano->2")
    idiomaTrain2 = input()
    if idiomaTrain2 == "1":
        aux2 = "../dataset/Español -> Ingles"
    else:
        aux2 = "../dataset/es"

    aux3 = "1"
    opciones = []
    while aux3 !="0":
        print("Elige las opciones opcionales: ngrams, char_ngrams, NRC_sent, NRC_emo o NRC_hash:")
        opciones.append(input())
        print("Seguir añadiendo->1; Salir->0")
        aux3 = input()
```

```
parser = argparse.ArgumentParser()
parser.add_argument('-sd', '--src_dataset', default= aux1)
parser.add_argument('-td', '--trg_dataset', default= aux2)
parser.add_argument('-emo', '--emotion', default=emocion)
parser.add_argument('-f', '--features', nargs='+', default=opciones)

args = parser.parse_args()
```

---

3 Resultados del modelo sin aplicar ninguna estrategia:

TEST	ANGER	JOY	SADNESS	FEAR
NGRAMS	0,06	0,23	-0,03	0
CHAR_NGRAMS	0,13	0,21	-0,11	-0,01
NGRAMS-NRC_EMO	0,06	0,17	0,02	-0,02
CHAR_NGRAMS-NRC-EMO	0,13	0,2	-0,08	-0,01
NGRAMS-NRC_SENT	0,03	0,15	-0,05	0,02
CHAR_NGRAMS-NRC_SENT	0,07	0,2	-0,1	-0,02
NGRAMS-NRC_HASH	0	0,09	0,04	0,07
CHAR_NGRAMS-NRC_HASH	0,04	0,21	-0,03	0,02
MEDIAS	0,065	0,1825	-0,0425	0,00625

Ahora que hemos probado varias combinaciones básicas vamos a probar a trabajar con la Estrategia 1. Como he mencionado antes, he preparado una carpeta con los datos de test que

estaban previamente en español pero traducidos a inglés. Para traducirlos he usado la web de LibreTranslate que además de traducir textos, permite traducir archivos con formato .txt entre otros.

En el código que he puesto más arriba en este documento, no hay que cambiar nada. Es en la ejecución donde tendremos que especificar que queremos los datos en inglés.

Por tanto, para comenzar vamos a ver como quedarían las tablas usando la ESTRATEGIA 1.

la ESTRATEGIA 1:

En esta estrategia lo que vamos a hacer es traducir los datos de evaluación que están en castellano a inglés con un traductor automático. Posteriormente, entrenaremos el modelo con los datos de entrenamiento en inglés y lo evaluaremos con los datos de evaluación traducidos a inglés.

Tras hacer los pasos anteriores obtenemos una tabla como esta:

---

4 Resultados del modelo aplicando

TEST	ANGER	JOY	SADNESS	FEAR
NGRAMS	0,24	0,16	-0,15	0,13
CHAR_NGRAMS	0,21	0,27	-0,19	0,08
NGRAMS-NRC_EMO	0,31	0,29	-0,06	0,29
CHAR_NGRAMS-NRC-EMO	0,25	0,33	-0,15	0,15
NGRAMS-NRC_SENT	0,31	0,31	0,09	0,24
CHAR_NGRAMS-NRC_SENT	0,25	0,34	-0,1	0,12
NGRAMS-NRC_HASH	0,31	0,28	0,04	0,28
CHAR_NGRAMS-NRC_HASH	0,26	0,31	-0,11	0,16

MEDIAS	0,2675	0,28625	-0,07875	0,18125

He probado también combinaciones de tres elementos y dan resultados muy parecidos.

La mejor combinación según la tabla sería la de ngrams junto con NRC_sent.

En conclusión, la aplicación de la estrategia de utilizar la traducción automática para llevar a cabo la traducción del conjunto de datos de evaluación de español a inglés ha demostrado ser una aproximación medianamente eficaz en la evaluación de modelos entrenados en un conjunto de entrenamiento en inglés. A lo largo de este estudio, hemos observado cómo esta estrategia proporciona una solución algo viable pese a que los resultados son algo pobres.

Es crucial tener en cuenta las posibles pérdidas de información durante el proceso de traducción automática y evaluar cuidadosamente su impacto en los resultados finales. Además, la implementación de esta estrategia resalta la importancia de considerar enfoques innovadores y flexibles en el ámbito de la inteligencia artificial, particularmente cuando se enfrenta a desafíos relacionados con la diversidad de idiomas. Aunque la traducción automática puede introducir ciertos sesgos y limitaciones, su aplicación estratégica puede allanar el camino para investigaciones futuras que busquen mejorar y refinar aún más este enfoque.

---

5 Resultados del modelo aplicando la ESTRATEGIA 2:

Anteriormente, en la estrategia 1 hemos visto los resultados del modelo de regresión de emociones al evaluar los resultados del modelo entrenado en inglés en textos en español. Tradujimos el conjunto de datos de evaluación al inglés mediante traducción automática y luego evaluamos el modelo entrenado en inglés en estos textos traducidos. En la segunda estrategia, utilizamos traducción automática para convertir el conjunto de entrenamiento y desarrollo en inglés al español, evaluando posteriormente el modelo en el conjunto traducido en textos en español.

Para traducir los archivos he usado la página web Onlinedoctranslator.com

Tras entrenar el modelo con los datos que hemos traducido al castellano, evaluamos con los datos de evaluación en castellano y obtenemos la siguiente tabla:

TEST	ANGER	JOY	SADNESS	FEAR
NGRAMS	0,11	0,14	-0,07	0,08
CHAR_NGRAMS	0,18	0,26	0,04	0,13

NGRAMS-NRC_EMO	0,11	0,13	-0,08	0,07
CHAR_NGRAMS-NRC-EMO	0,18	0,26	0,03	0,12
NGRAMS-NRC_SENT	0,15	0,16	-0,03	0,11
CHAR_NGRAMS-NRC_SENT	0,21	0,27	0,05	0,13
NGRAMS-NRC_HASH	0,08	0,13	-0,07	0,1
CHAR_NGRAMS-NRC_HASH	0,18	0,26	0,03	0,13
MEDIAS	0,15	0,20125	-0,0125	0,10875

¿Qué podemos observar? En general los resultados son significativamente peores a la primera estrategia, esto podemos verlo en los resultados tan bajos que han salido. La combinación que mejores valores da es la de char_ngrams junto con NRC_sent

---

6 CONCLUSIONES:

En el análisis de las tres estrategias en diferentes emociones, la Estrategia 1 emerge como la mejor elección. En las emociones anger y joy, esta estrategia exhibe valores más altos, acercándose en mayor medida al objetivo óptimo de 1. La consistencia en la superioridad de la Estrategia 1 en estas emociones sugiere un desempeño más confiable. Aunque en sadness ninguna estrategia alcanza el valor deseado, la Estrategia 2 se posiciona ligeramente mejor que la Estrategia 1 en esta emoción en particular. Sin embargo, los valores en ambos casos son prácticamente 0. Al considerar el panorama general, incluida también la emoción fear, la Estrategia 1 vuelve a destacarse con una aproximación más cercana a nuestro objetivo.

Teniendo en cuenta lo anterior, diríamos que es mejor traducir los datos de evaluación a inglés.

---

7 Bibliografia

[1] Irean Navas Alejo, Toni Badia, and Jeremy Barnes. 2020. Cross-lingual Emotion Intensity

Prediction. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 140–152, Barcelona, Spain (Online). Association for Computational Linguistics.

[2] LibreTranslate: https://libretranslate.com/

[3]Onlinedoctranslator: https://www.onlinedoctranslator.com/es/translationform