Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Materia: Temas de Procesamiento del Lenguaje Natural - NLP Aplicado

Trabajo práctico

Objetivo

El objetivo del trabajo práctico es trabajar en una tarea de anotación de textos, evaluar la performance entre anotadores, armar un *data statement* y evaluar la anotación automática a través de prompts a un LLM (como ser ChatGPT, Gemini, DeepSeek u otro).

Introducción

El discurso de odio contra las mujeres, los inmigrantes y muchos otros grupos es un fenómeno generalizado en Internet, particularmente desde la eclosión de las redes sociales. Numerosos trabajos han estudiado este discurso prevalente. A pesar de esta prevalencia general, suelen tener lugar "picos" en la discriminación en las redes sociales luego de eventos disparadores, como han sido crímenes de odio (asesinatos con motivaciones religiosas en Woolwich, atentados terroristas, etc) u otros eventos.

Las redes sociales y las plataformas digitales manejan enormes cantidades de contenido generado por los usuarios. Si bien democratizan el acceso y la difusión de opiniones, y contribuyen al debate público, en ocasiones también fomentan el odio y la discriminación. El discurso de odio, definido como cualquier comunicación que degrada a un grupo de personas por una característica común como raza, religión, etnia, género o preferencia política, entre otros [3] , se ha convertido en un problema global y generalizado [4].

Entre las diversas manifestaciones de discurso de odio, el antisemitismo constituye una forma significativa de discurso de odio dirigido contra los judíos. Si bien este fenómeno tiene raíces históricas profundas, en la actualidad persiste y se expresa de manera renovada a través de plataformas digitales, redes sociales y espacios de participación en línea como los comentarios de noticias.

En este trabajo nos proponemos anotar un dataset pequeño de comentarios asociados a notas periodísticas para determinar si son de carácter antisemita o no. Los comentarios (COMM) hacen referencia a notas periodísticas y estarán acompañados por el título de la nota a partir de la cuál se originan, como para poder interpretarlos con un contexto mínimo. Se trata de notas de medios de Argentina (Clarín, La Nación, Infobae), Chile (Biobio, Emol, La Tercera), Colombia (Dos Orillas, El Tiempo), Costa Rica (CR Hoy, Diario Extra), Panamá (La Estrella, La Prensa) y Uruguay (El Observador, El País, Montevideo Portal, Subrayado).

Datos

Se trabajará con un dataset para detectar discursos antisemitas en textos escritos en español.

Tareas

1. Cada integrante del equipo tiene que anotar 40 Comentarios. 10 de ellos tienen que coincidir entre todos. Las posibles clases son **antisemita, negativo, indefinido** (sólo usarla en casos excepcionales), para el caso en que no pueden anotarlo como una de las dos otras opciones.
   (dataset 1)
2. Leer los **criterios de anotación** y anotar 40 comentarios más de cada conjunto con 10 de intersección, al igual que en el punto 1 (**dataset 2**)
3. En los puntos 1 y 2 se permite y se incentiva buscar información para aclarar significados de términos con los que no estén familiarizados, si eso los ayuda a hacer el trabajo con más información. No facilitamos el link de la nota periodística original y se espera que no la busquen.
4. Para los puntos 1 y 2,
   a. calcular el acuerdo entre anotadores (Inter Annotator Agreeement -IAA-) con alpha de Krippendorf.
   b. Para los datos anotados por más de un anotador, determinar cómo decidirán con cuál se quedan.
5. Armar el resumen ejecutivo de un data statement [1, sección 2], un breve *curation rational* [1, sección 3], annotator demographic y descripción del dataset (como en tabla 3 de la referencia [3]. Pueden incluir otras secciones del data statement, si así lo desean.
6. Construir un prompt (de manera iterativa) en algún LLM , probando distintas técnicas (zero shot, few shot, etc). El objetivo del prompt es poder obtener una anotación automática de todos los comentarios anotados por el equipo (sin poner como entrada el resultado de vuestra anotación). Para hacer few shot usar datos del **dataset 3.**. Una vez que se tiene un prompt definido y estabilizado, probarlo con 40 comentarios del dataset 3 (distintos a los usados para hacer el few shot). Estos datos vieenen anotados. . Una vez que “funciona razonablemente”, probarlo con vuestros datos anotados **(datasets 1 y 2)**.
7. Calcular IAA para
   a. Anotación original del corpus vs vuestra (para la vuestra tomar la de los puntos 1 y 2) (la anotación original se libera en unos días)
   b. Anotación original del corpus vs LLM (tomar la de los puntos 1 y 2)
   c. Anotación vuestra vs LLM (para la vuestra tomar la de los puntos 1 y 2)
8. Armar un informe, estilo paper, de entre no más de 4 hojas (puede incluir apéndices, y en ese caso superar las 4 hojas), contando los experimentos realizados y los data statements. Se sugiere template de ACL. Debe incluir
   a. Introducción

    i. Cuál es el problema, Por qué es importante, qué hacen, en qué se diferencia de otros trabajos
b. Trabajos previos para Español si encuentran (distintas variantes). Sino en inglés y/ o otras lenguas.
c. Metodología
    i. Cómo seleccionaron los subconjuntos de datos a anotar
    ii. Data Statements escritos por ustedes (de acuerdo a lo pedido más arriba).
    iii. Detalles técnicos (decisiones de implementación, por ej. cómo se resuelven discrepancias entre asignación de anotadores) y cómo implementaron el IAA.
    iv. Cómo hicieron el prompting
        1. Método de prompting
        2. Prompt utilizado (agregar ejemplos si los hubo) (en anexo)
d. Resultados del IAA (vuestro cálculo de **la tabla 3 de la referencia [3])** y del cálculo hecho en el punto 6 y análisis de resultados. **Adjuntar link** a resultados de anotaciones hechas por anotadores (previos), las vuestras, y las de ChatGPT.
e. Conclusiones
f. Referencias bibliográficas

Fecha de entrega: Domingo 6 de julio, 23:59. Envio por mail a vivianacotik@gmail.com. Con subject: TP Temas NLP 2025 - Grupo X, en donde X es el nro del grupo.

Otras fechas:

- Release dataset 1: 27/06: https://docs.google.com/spreadsheets/d/1bi7g7oKL_cx9MJkzfOozu0liP75_j-W3/edit?gid=1667212715#gid=1667212715
- Criterio de anotación. Ya tienen que tener los primeros comentarios anotados (punto 1): 01/ 07
- Release dataset 2 (posterior a punto 1): 01/07
- Release dataset 3 (posterior a punto 2): 03/07
- Release resultados anotación 1 y 2: 04/07

Referencias

[1] Bender, E.M. and Friedman, B., 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, pp.587-604. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00041/43452/Data-Statements-for-Natural-Language-Processing

[2] https://techpolicylab.uw.edu/wp-content/uploads/2021/11/Data_Statements_Printer_Guide_V2.pdf

[3] Pérez, J.M., Luque, F.M., Zayat, D., Kondratzky, M., Moro, A., Serrati, P.S., Zajac, J., Miguel, P., Debandi, N., Gravano, A. and Cotik, V., 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, *11*, pp.30575-30590. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10076443

[4] Pérez, J.M., Luque, F.M., Zayat, D., Kondratzky, M., Moro, A., Serrati, P.S., Zajac, J., Miguel, P., Debandi, N., Gravano, A. and Cotik, V., 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, *11*, pp.30575-30590. https://openreview.net/forum?id=01eOESDhbSW