IDENTIFICACIÓN DE COMENTARIOS SEXISTAS NAS REDES SOCIAIS MEDIANTE TÉCNICAS DE INTELIXENCIA ARTIFICIAL

Rebeca Díaz Redondo[1] e Ana Fernández Vilas[1]
Mateo Ramos Merino e Sonia Valladares Rodríguez, DATASALUS SCG
[1]I&C Lab, EE de Telecomunicacións, Universidade de Vigo, rebeca@det.uvigo.es

Cátedra de Feminismos 4.0
DEPO - UVigo

DEPUTACIÓN PONTEVEDRA

Universidade de Vigo

DATASALUS

1. DESCRICIÓN

Aumento da **violencia de xénero dixital** na adolescencia [1]

Ámbito: RRSS e área Pontevedra

Problemática actual:

- Sexismos en comentarios nas RRSS
- Desigualdade de xénero nas RRSS
- Maior exposición das mulleres nas RRSS [2]
- Foco dos comentarios sexistas nas RRSS [2]
- Normalización deste tipo de ataques nas RRSS

Nova aproximación -> Intelixencia artificial contra o machismo dixital.

Partir dos avances acadados o ano pasado:

. Dataset con +87.000 comentarios procedentes de diversas RRSS

. Etiquetado manual de +4.300 comentarios en función da presencia de machismo

. Modelo de NLP/ML para a identificación automática de comentarios sexistas

IA4♀

2. OBXECTIVOS

. Incrementar o dataset actual

. Incrementar o % do etiquetado do dataset actual

. Mellorar e reentrenar os modelos de NLP/ML

. Xerar ferramenta Web, a modo de demostrador, para a identificación automatizada de comentarios sexistas.

. Fometnar a divulgación e a concienciación sobre esta problemática a través de talleres en institutos de secundaria e bacharelato da provincia

3. METODOLOXÍA

Tecnoloxías clave de IA [4]:

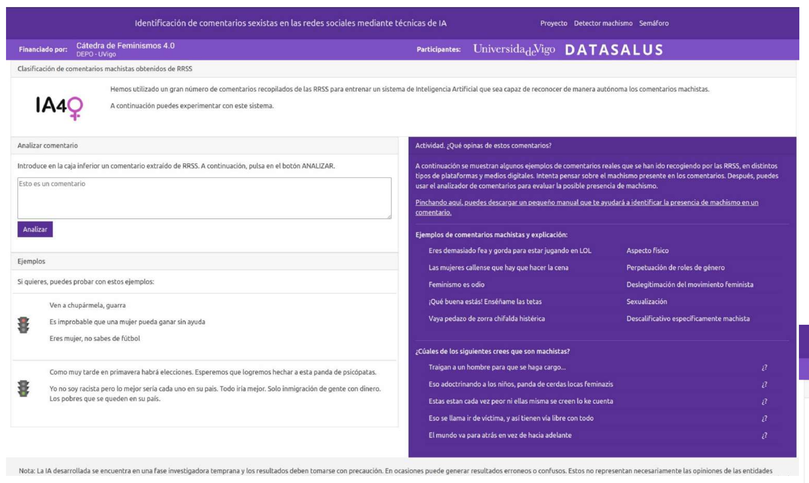
1. NLP
   - Procesado de linguaxe natural [3]
   - Extracción e identificación de palabras clave
2. ML
   - Machine learning [4]

1. Estudo, selección e adquisición dos "datasets"
2. Etiquetado dos "datasets"
3. Adestramento de modelos de ML (IA)
4. Creación da ferramenta Web
5. Divulgación en institutos
6. Identificación de posibles sesgos
7. Difusión dos resultados

4. PROMOCIÓN DA IGUALDADE DE XÉNERO

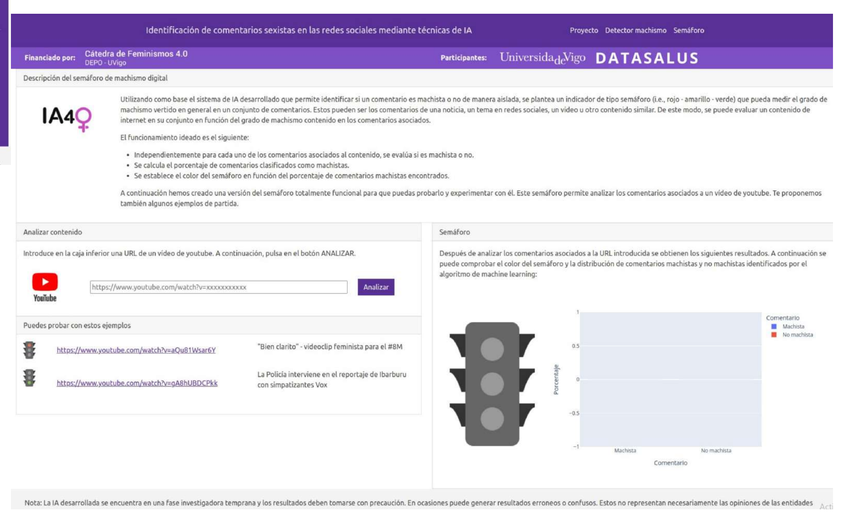
Potenciais ámbitos de mellora derivados dos resultados do estudo:

1. Dotar de ferramentas para a regulación automática de comentarios sexistas
2. Estudar os perfís das persoas usuarias máis propensas a realizar este tipo de comentarios nas RRSS
3. Contribuír á elaboración de sistemas de detección de ciber-acoso ou semellantes.
4. Identificación automática de conversas con este tipo de contidos para evitar e que se volvan virais
5. Creación de indicadores que amosen o grao xeral de comentarios sexistas
7. Mellorar os protocolos de identificación de risco real de violencia de xénero [5] (e.g., mensaxería ou RRSS) dunha potencial vítima



Ferramenta Web. Clasificación de comentarios introducidos de xeito manual e proposta de actividades sobre a presenza de sexismo en comentarios de RRSS.

Ferramenta Web. Semáforo de machismos dixitais. Permite analizar os comentarios asociados a un vídeo de youtube. En función da porcentaxe de comentarios identificados como machistas ou non, establécese unha cor para o semáforo.



5. IMPACTO E DIFUSIÓN

Data	Tarefas planificadas
Xuño-Setembro 2022	- Refinamento e actualización dos criterios para o etiquetado de comentarios*. Aumento do etiquetado. - Mellora dos modelos de NLP/ML. - Xeración da ferramenta Web
Outubro 2022	-Contacto con institutos e preparación das sesións a impartir
Novembro 2022	-Impartir sesións nos institutos
Decembro 2022	-Avaliación resultados e da impartición das sesións de divulgación. -Informe de xustificación

* Refinamento dos criterios para o etiquetado de comentarios asesorado e supervisado pola técnica en igualdade de xénero Amara Pérez Davila.

6. REFERENCIAS

- [1] https://www.unicef.org/es/end-violence/ciberacoso-que-es-y-como-detenerlo .
- [2] La desigualdad de género y el sexismo en las redes sociales: una aproximación cualitativa al uso que hacen de las redes sociales las y los jóvenes de la CAPV. Estébanez, Ianire; Vázquez, Norma; Observatorio Vasco de la Juventud (coord.) (2013) ISBN (13): 978-84-457-3295-3
- [3] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.
- [4] Mabrouk, A., Redondo, R. P. D., & Kayed, M. (2020). Deep learning-based sentiment classification: A comparative survey. *IEEE Access*, *8*, 85616-85638. .
- [7] González Álvarez, J. L., López Ossorio, J., & Munoz Rivas, M. (2018). La valoración policial del riesgo de violencia contra la mujer pareja en España–Sistema Viogén.