TÍTULO DO TFG/TÍTULO DEL TFG:

Técnicas de Intelixencia Artificial Aplicadas á Detección de Ciberacoso en Redes Sociais

Resumo / Resumen (máximo 350 palabra):

O ciberacoso é un problema real no noso país e tamén a nivel global. De acordo con Unicef, 1 de cada 3 adolescentes sufre ciberacoso cada ano, o que deriva en numerosos trastornos de depresión, ansiedade, falta de autoestima e incluso un posible suicidio.

O obxectivo principal desde TFG é adestrar un modelo de Intelixencia Artificial capaz de detectar ciberacoso en comentarios de Redes Sociais, unha tarefa que resulta altamente complexa debido a que ditos comentarios se atopan escritos en Linguaxe Natural. Ademais, cómpre ter en conta que o rexistro empregado en redes sociais é moi informal, contando con numerosas palabras malsoantes, insultos escritos sen intención de insultar, faltas de ortografía e abreviaturas. Isto dificulta aínda máis a análise, xa que non basta un simple detector de palabras.

Partindo dun dataset existente de publicacións de Instagram cos seus respectivos comentarios, clasificadas segundo se conteñen ou non ciberacoso, realízase unha comparación exhaustiva de diferentes algoritmos de Procesamento de Linguaxe Natural (NLP). Estas técnicas (tales como FastText ou Doc2Vec), permiten adestrar de forma non supervisada modelos que converten texto a vectores, situando comentarios con significado similar cerca no espazo vectorial.

Empregando estas técnicas de NLP, conséguense xerar novos datasets de vectores que poden ser “entendidos” por un algoritmo de clasificación (como Random Forest ou Support Vector Classifier). Deste xeito, por cada novo dataset (xerado a partir dunha combinación de algoritmos de aprendizaxe non supervisados), adéstrase un modelo clasificador. Todos estes novos modelos son comparados empregando métricas estándar (accuracy, recall, F1 e precision), determinando que combinación de algoritmos é quen de detectar o máximo número de casos de acoso posible, minimizando a cantidade de falsos positivos.

Tras conseguir adestrar un modelo capaz de detectar de forma correcta máis dun 81% dos casos de acoso (e capaz de descartar case un 90% dos casos sen acoso), créase un servizo REST cunha petición que devolva a partir dun texto dado a probabilidade de que o mesmo constitúa ciberacoso.

Finamente, prográmase unha aplicación móbil para iOS, que permite empregar este servizo REST a través dunha interface coidadosamente deseñada.

Posibles aplicacións / Posibles aplicaciones (máximo 250 palabra):
Tendo en conta que o modelo empregado polo servizo REST é capaz de detectar un 81% dos casos de acoso en redes sociais, e sabendo que 1 de cada 3 adolescentes sofre ciberacoso cada ano, poderíase deseñar unha aplicación que obteña automaticamente comentarios de publicacións aleatorias en Instagram e empregue este servizo REST para detectar se conteñen ou non ciberacoso. No caso de que se detecte unha publicación con ciberacoso, esta podería ser eliminada de forma automática, reducindo a probabilidade de que a vítima sufra as devastadores consecuencias psicolóxicas e incluso físicas que causa o ciberacoso.

A aplicación móbil, creada a modo de demo, tamén pode ser útil para que un adolescente que crea que sufre ciberacoso, poida pegar os comentarios ferintes na mesma e obter unha confirmación de que está sufrindo acoso, animándoo a dar o paso a denunciar os feitos antes de que sexa tarde. Ademais, esta aplicación permite chamar ou abrir un chat co servizo de axuda ó menor da asociación ANAR.

Etapas para o seu desenvolvemento futuro / Etapas para su desarrollo futuro (máximo 250 palabras):

- Obter datasets en diversos idiomas que permitan adestrar novos modelos, de xeito que o servizo sexa multilíngue.
- Deseñar un detector automático de idioma, que empregue un modelo ou outro en función do resultado.
- Implementar un sistema que permita facer *scrapping* en redes sociais, analizando e bloqueando publicacións con ciberacoso.

Imaxes representativas / Imágenes representativas (máximo 2):