Desarrollo de técnicas de aprendizaje profundo para la clasificación de memes empleados en campañas de desinformación

David Arroyo Guardeño

5 de octubre de 2021

Introducción

Unos de los principales retos en el ámbito cibernético viene determinado la dificultad de evaluar la fiabilidad de fuentes de información. En este proyecto se diseñará una técnica basada en aprendizaje profundo (Deep Learning) para identificar dinámicas y campañas de desinformación mediante análisis multimodal de texto y de imágenes. Se partirá del estado del estado en lo concerniente al uso de memes en campañas de desinformación [1] y operaciones de información [2], y se trabajará en incorporar funcionalidad de análisis de imágenes a la herramienta MsW desarrollada en el contexto del proyecto europeo TRESCA.

El plan de formación a llevar a cabo se centrará en la identificación y/o creación de conjunto de datos y en el entrenamiento/validación de redes neuronales recurrentes y convolucionales para la clasificación de memes usados en operaciones de información científica. El equipo de trabajo encargado de supervisar el plan de formación forma parte tanto del proyecto **TRESCA** como del proyecto **XAI-Disinfodemics** adscrito a la temática 18 ("Desinformación, engaños y noticias falsas a través de canales públicos y privados") de la convocatoria de proyectos en líneas estratégicas 2021 de la Agencia Estatal de Investigación.

Proyecto formativo, actividades y competencias a desarrollar

El proyecto tendría por objetivo el explora técnicas de Deep Learning de acuerdo con la implementación en Python proporcionada a través de Tensorflow y Keras. Para ello, se tendrá acceso a los recursos de cómputo intensivo del Grupo de investigación en Criptografía y Seguridad de la Información del CSIC.

Se desarrollarán, pues, competencias en el diseño de soluciones de aprendizaje profundo mediante Python. Asimismo, se adquirirán conocimientos relacionados con el fenómeno de desinformación y los canales y recursos que se emplean en dicho contexto. El grupo de trabajo en el que se encuadrará este proyecto es de naturaleza multidisciplinar, e incluye a científicos sociales, periodistas e ingenieros en informática y de telecomunicación.

Metodología y plan de trabajo

Para llevar a cabo los objetivos del presente Trabajo de Fin de Máster se efectuarán las siguientes tareas de acuerdo con el cronograma orientativo que figura a continuación:

1. Introducción al concepto y fenomenología de la desinformación, así como a su caracterización mediante técnicas automática de procesamiento de información extraída de fuentes abiertas de datos y redes sociales (~20 h). Aquí se tomarán como principales referencias el estudio del arte y las referencias analizadas en el contexto del proyecto TRESCA[1], y se tomará contacto con la herramienta MsW desarrollada en el contexto de dicho proyecto.
2. Introducción a la programación de modelos de aprendizaje profundo mediante Tensorflow y Keras (~50 h).
3. Búsqueda e identificación de los conjuntos de datos necesarios para el entrenamiento y validación de los modelos de aprendizaje automático (~10h).
4. Diseño e implementación de un módulo de aprendizaje profundo para su integración en la herramienta MsW de TRESCA (~100 h).
    - Definición de requisitos funcionales

[1] https://trescaproject.eu/results/

- Implementación del módulo de acuerdo con la especificación de requisitos

5. Validación del módulo de aprendizaje automático (~80 h).
   - Diseño de un plan de pruebas de las principales funcionalidades de sistema.
   - Ejecución y validación del plan de pruebas de funcionalidades.
6. Redacción de la memoria final (~40 h).

Desde el punto de vista metodológico, la supervisión del TFM comportará los siguientes punto de obligado cumplimiento:
1. Reunión de supervisión cada 15 días. Si se considera que no hay nada que comentar, se aplaza la reunión. Esto es simplemente un mecanismo de gestión para guiar la concreción del proyecto. En la medida que el TFM es una primera aproximación a un proyecto profesional, después de cada reunión tendría que hacer un pequeño acta, un muy breve resumen (no debe suponer más de 5 minutos de trabajo) de los puntos abordados y de las líneas de trabajo acordadas.
2. La correcta gestión del proyecto, por otro lado, requiere un uso eficiente del correo electrónico por parte del alumno y del tutor del TFM. El tutor asume el compromiso de responder el correo en tiempo, eso sí, sujeto a la carga de trabajo que exista en cada momento. Por parte del alumno se asume el compromiso de estar atento al correo y responder en tiempo a las cuestiones que surjan.
3. Al margen de la interacción vía reuniones, será necesario habilitar un repositorio privado (normalmente en Bitbucket o github) para compartir ttanto el código como la documentación del proyecto. Para la compartición del material bibliográfico, se deberá crear un grupo privado en [Mendeley](https://mendeley.com/).
4. La redacción de la memoria del TFM debe realizarse en LaTeX.

Datos del Tutor

David Arroyo Guardeño Científico Titular del Instituto de Tecnologías Físicas y de la Información "Leonardo Torres Quevedo" C/Serrano nº 144 28006 Madrid

Información adicional

https://dargcsic.github.io

david.arroyo@csic.es

Posibles fuentes de datos de memes

1. https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/
2. https://github.com/bharathichezhiyan/Multimodal-Meme-Classification-Identifying-Offensive-Content-in-Image-and-Text
3. https://imatge.upc.edu/web/sites/default/files/pub/xOriol.pdf
4. https://github.com/MIND-Lab/MEME
5. https://snap.stanford.edu/data/memetracker9.html
6. https://www.mongodb.com/blog/post/twitter-memes-dataset-overview-with-pagerank
7. https://competitions.codalab.org/competitions/25337#results

Referencias

[1] C. Ling, I. AbuHilal, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, "Dissecting the meme magic: Understanding indicators of virality in image memes," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–24, 2021.

[2] T. WILSON and K. STARBIRD, "Cross-platform information operations: Mobilizing narratives and building resilience through both 'big'and 'alt'tech," 2021.