Jornada Informativa sobre Inteligencia Artificial aplicada a las Políticas Públicas de Ciencia, Innovación y Emprendimiento

Dentro del Plan TL, la Secretaría de Estado para el Avance Digital (SEAD), en colaboración con la Fundación Española para la Ciencia y la Tecnología, ha organizado una jornada sobre Inteligencia Artificial aplicada a las Políticas Públicas de Ciencia, Innovación y Emprendimiento.

La jornada estaba dividida en tres sesiones, la primera de las cuales denominada: **Postdigitalización de la política científica de innovación y emprendimiento,** tenía por objetivo revisar experiencias internacionales sobre aplicaciones de la inteligencia artificial en el campo de las políticas públicas de ciencia e innovación.

Durante la sesión se ha destacado el potencial y los retos de los nuevos sistemas digitales de información que se usan en la política pública de apoyo a la ciencia y la innovación, se han realizado apuntes de la situación de España en el **contexto internacional** de la I+D+i, destacando el **buen posicionamiento de España en la lista de publicaciones científicas más citadas en el campo de Inteligencia Artificial**.

Se han presentado las iniciativas de la **OCDE** en relación con la digitalización de la política de I+D+i con el proyecto DSIP y la utilidad del PLN para la obtención de indicadores. Asimismo, se han mencionado las principales iniciativas de la **Comisión Europea** para dotarse de nuevos sistemas digitales de información y se ha presentado el trabajo conjunto con la OCDE para extraer buenas prácticas para el diseño de los instrumentos de apoyo a la ciencia y la innovación a partir de la recopilación de las iniciativas nacionales. También se han mostrado las herramientas avanzadas de visualización (STIP Compass) con las que la Comisión sigue la ejecución de su programa marco de ayudas a la I+D+i, Horizonte 2020. Se ha cerrado la sesión explicando un caso concreto de uso de herramientas avanzadas para la política de I+D+i: el seguimiento del desarrollo, la adopción y el impacto de la inteligencia artificial en Europa, presentado por **JRC Sevilla**.

La segunda sesión denominada: **Resultados del proyecto de uso de TL en la política española de ciencia innovación y emprendimiento**, tenía por objetivo revisar la situación actual y perspectivas de evolución del Plan TL y mencionar los avances del uso de las tecnologías del lenguaje en España en el campo de la política científica, prestando especial atención al desarrollo de herramientas propias y a las ventajas de su utilización.

Se empezó señalando que las **tecnologías del lenguaje** (TL) permiten **procesar de manera inteligente grandes colecciones de documentos, cuyo contenido está expresado en lenguaje natural**. Estas tecnologías incluyen el procesamiento del lenguaje natural, la traducción automática y los sistemas conversacionales. Además, las TL son una parte fundamental de la Inteligencia Artificial y tienen el potencial de servir de **apoyo al diseño de estrategias**.

En este sentido, se destacó el hecho de que **el combustible de las TL y de la IA son los datos**, por lo que los resultados de estas nuevas tecnologías están condicionados a la **disponibilidad y calidad** de los datos sobre los que trabajan. El circuito de producción y comunicación de la ciencia genera gran cantidad de datos que obliga a gestionarlos de forma eficiente a través de la implementación de **identificadores únicos y persistentes**, pieza clave para permitir la interoperabilidad entre los sistemas y para mejorar la visibilidad de la información. Internacionalmente, ya se han implementado identificadores únicos y persistentes para las

publicaciones y datos, autor y fuente de financiación. Queda pendiente la creación de un identificador único y permanente para entidades.

Las **TL** complementan otras herramientas estadísticas utilizadas tradicionalmente y permiten el tratamiento de datos no estructurados (colecciones de texto en lenguaje natural). Estas tecnologías permiten **detectar las temáticas principales de los documentos y generar vistas de utilidad**. Se menciona también que la heterogeneidad de las fuentes y los aspectos legales de los datos personales pueden ser los principales retos para su desarrollo.

El **Plan TL** es una iniciativa de la **SEAD** que tiene como objetivo **impulsar en España las tecnologías del lenguaje**, aprovechando estas novedosas capacidades **para favorecer el emprendimiento, impulsar la industria de este sector y mejorar el servicio público**. En el plan se establecen colaboraciones y se desarrollan herramientas mediante convenios entre la SEAD que aporta expertos en IA y entidades que aportan colecciones de documentos y expertos de dominio.

Entre los primeros proyectos del Plan TL, se encuentra la aplicación de las tecnologías del lenguaje a las **políticas públicas (Inteligencia Competitiva)**, en concreto en el ámbito del **I+D+i.** En este proyecto se ha desarrollado la **plataforma propia *"Corpus Viewer"*** (Visor de corpus documentales).

Corpus Viewer es una plataforma genérica que, haciendo uso de **tecnologías del lenguaje y otras técnicas de IA**, puede ser explotada con prácticamente **cualquier colección de documentos** de texto. El despliegue actual de la plataforma aloja principalmente corpus relacionados con el I+D (proyectos de I+D, publicaciones científicas y patentes). Los primeros **usuarios** han sido: la **SEAD**, la **FECYT** y la **SEUIDI**. Se emplea en el **diseño de políticas**, como la Estrategia de Inteligencia Artificial, así como en la **gestión de convocatorias de ayudas** como apoyo en la evaluación y seguimiento.

En la Jornada se ha hecho una **demostración detallada de la Plataforma Corpus Viewer** que incluye entre sus utilidades: **visor de documentos por temáticas, evolución temporal, caracterización y clasificación automática de documentos, detección de copias y fraudes, cuadros de mando personalizados tipo BI, grafos,** etc. Esta **plataforma** ha sido desarrollada gracias a la colaboración de varios **grupos de investigación universitarios y empresas, coordinados por la SEAD**.

Está basada en componentes de **código abierto, con escalabilidad completa**, orientada a microservicios, e independiente del hardware, con despliegue basado en **contenedores *docker*** y orquestación con ***Kubernetes***. Tiene el **despliegue automatizado** con ***Ansible.***

La tercera sesión: **Aplicaciones futuras de las tecnologías del lenguaje en las políticas públicas**, tenía por objetivo plantear los **próximos pasos de la herramienta Corpus Viewer** y el potencial de aplicar este tipo de herramientas basadas en TL, no sólo en la política de I+D+i, sino también **en otras políticas públicas**.

Como ejemplo de **aplicación de estas técnicas a otras políticas públicas**, encontramos el caso presentado por el ONTSI, en el que a través de la recogida de datos de Internet y mediante la aplicación de tecnologías del lenguaje, se ha abordado el piloto **Internet as Data Source – IaD**, en base a dos objetivos: Analizar la oferta y la **demanda de profesionales del sector TIC** en nuestro país y detectar las empresas españolas que realizan **comercio electrónico**. Tanto las empresas como los ciudadanos dejan numeroso 'rastro digital' en Internet. Mediante la recolección y explotación de dicha información es posible describir numerosos fenómenos socio-económicos casi en tiempo real. IaD permite identificar datos e indicadores que se pueden obtener directamente de Internet, describiendo nuevos hábitos y usos que no están cubiertos por las metodologías tradicionales o exigen un enorme esfuerzo en dedicación de recursos

económicos y humanos que lo hacen inviable. De cara a la definición de políticas futuras de la economía y sociedad digital, el uso de IaD se conforma como una alternativa posible para disponer de datos sobre los usos de Internet. Utilizar IaD puede proporcionar una visión rápida sobre fenómenos nuevos sobre los que las técnicas tradicionales tienen dificultad de medir. Pueden mejorar la calidad de las estadísticas, sobre todo cuando se combinan con las metodologías tradicionales. Además, puede ser una forma de reducir la carga de trabajo sobre las unidades informantes, ya sean empresas o individuos.

Se presentaron también en esta sesión los últimos desarrollos realizados y se resaltó como la **plataforma Corpus Viewer está en constante evolución** y prevé seguir incorporando nuevas funcionalidades (i.e. perfilado de agentes y modelado de redes de colaboración) cuyos desarrollos están finalizando actualmente.

En relación a los **próximos pasos de Corpus Viewer**, se resaltó la gran **necesidad de datos** (**nuevos corpus** i.e. corpus de empleo, corpus de contratación pública, corpus completos de concesiones Plan Nacional I+D…) para el éxito de estas herramientas que permiten a los responsables de Políticas Públicas basar sus decisiones en datos. También se resaltó como punto de mejora, el realizar **adaptaciones** de la herramienta **a** los distintos **dominios**, ya que en las tecnologías del lenguaje cuando se realizan adaptaciones al dominio específico de los textos, la mejora de calidad de los resultados es muy significativa. Asimismo, se destacó la importancia de definir una **metodología común** y de establecer mecanismos de **validación científica**, proponiéndose para ello el uso de las **Campañas de Evaluación**. Por último, se indicó la apuesta decidida del Plan TL por avanzar hacia un **modelo distribuido** de **federación de servicios** que permita que los datos residan en el organismo correspondiente y la realización de consultas distribuidas con agregación de resultados: Repositorios locales con modelos globales compartidos.

Al concluir la Jornada, se ha **invitado a los organismos a solicitar un usuario** para poder acceder directamente al demostrador online que ha publicado el Plan TL. Desde el Plan TL se quiere impulsar la **reutilización de los desarrollos de las AAPP** de forma que se haga un **uso eficiente y racional de los recursos públicos**. Por ello, se informa que todos los recursos desarrollados en el Plan están disponibles a través de su página web y se invita a los organismos a utilizar la el **demostrador online de Corpus Viewer**, para que puedan valorar los resultados que podría generar la herramienta para sus propias fuentes de datos y a contactar con los responsables del Plan TL en caso de querer adoptar esta plataforma. Por otro lado, se recuerda a las empresas e investigadores con experiencia en TL que pueden darse de alta en el catálogo de la web para dar a conocer sus productos y servicios en este sector emergente.