# Resumen

Hoy en día, la sociedad tiene acceso y posibilidad de contribuir a grandes cantidades de contenidos presentes en Internet, como redes sociales, periódicos online, foros, blogs o plataformas de contenido multimedia. Todo este tipo de medios han tenido, durante los últimos años, un impacto abrumador en el día a día de individuos y organizaciones, siendo actualmente medios predominantes para compartir, debatir y analizar contenidos online. Por este motivo, resulta de interés trabajar sobre este tipo de plataformas, desde diferentes puntos de vista, bajo el paraguas del Procesamiento del Lenguaje Natural. En esta tesis nos centramos en dos áreas amplias dentro de este campo, aplicadas al análisis de contenido en línea: análisis de texto en redes sociales y resumen automático. En paralelo, las redes neuronales también son un tema central de esta tesis, donde toda la experimentación se ha realizado utilizando enfoques de aprendizaje profundo, principalmente basados en mecanismos de atención. Además, trabajamos mayoritariamente con el idioma español, por ser un idioma poco explorado y de gran interés para los proyectos de investigación en los que participamos.

Por un lado, para el análisis de texto en redes sociales, nos enfocamos en tareas de análisis afectivo, incluyendo análisis de sentimientos y detección de emociones, junto con el análisis de la ironía. En este sentido, se presenta un enfoque basado en Transformer Encoders, que consiste en contextualizar *word embeddings* pre-entrenados con tweets en español, para abordar tareas de análisis de sentimiento y detección de ironía. También proponemos el uso de métricas de evaluación como funciones de pérdida, con el fin de entrenar redes neuronales, para reducir el impacto del desequilibrio de clases en tareas *multi-class* y *multi-label* de detección de emociones. Adicionalmente, se presenta una especialización de BERT tanto para el idioma español como para el dominio de Twitter, que tiene en cuenta la coherencia entre tweets en conversaciones de Twitter. El desempeño de todos estos enfoques ha sido probado con diferentes corpus, a partir de varios *benchmarks* de referencia, mostrando resultados muy competitivos en todas las tareas abordadas.

Por otro lado, nos centramos en el resumen extractivo de artículos periodísticos y de programas televisivos de debate. Con respecto al resumen de artículos, se presenta un marco teórico para el resumen extractivo, basado en redes jerárquicas siamesas con mecanismos de atención. También presentamos dos instancias de este marco: *Siamese Hierarchical*

*Attention Networks* y *Siamese Hierarchical Transformer Encoders*. Estos sistemas han sido evaluados en los corpora CNN/DailyMail y NewsRoom, obteniendo resultados competitivos en comparación con otros enfoques extractivos coetáneos. Con respecto a los programas de debate, se ha propuesto una tarea que consiste en resumir las intervenciones transcritas de los ponentes, sobre un tema determinado, en el programa "La Noche en 24 Horas". Además, se propone un corpus de artículos periodísticos, recogidos de varios periódicos españoles en línea, con el fin de estudiar la transferibilidad de los enfoques propuestos, entre artículos e intervenciones de los participantes en los debates. Este enfoque muestra mejores resultados que otras técnicas extractivas, junto con una transferibilidad de dominio muy prometedora.