# Summary

Salvatore Greco

October 15, 2024

Natural Language Processing (NLP) is an interdisciplinary field of research that combines linguistics, computer science, and artificial intelligence, aiming to enable computers to understand and interpret human language.

In the last decade, NLP has experienced remarkable advancements, achieving performance levels previously unimaginable, with models even surpassing human performance in certain tasks. These advancements are driven by the development of innovative architectures like transformer-based models, as well as the increasing availability of computational resources and large datasets. However, despite their accuracy and capabilities, deep learning-based NLP models still face several challenges that limit their trustworthiness and responsible use in real-world applications.

Our research aims to face some of these challenges by designing innovative techniques for inspecting deep learning NLP classifiers to increase their transparency, fairness, and robustness. Specifically, we focus on three major challenges in NLP classifiers: (1) Explaining their predictions; (2) Identifying and mitigating bias; (3) Detecting performance degradation over time due to the presence of concept and data drift. To tackle these challenges, we design and propose three frameworks.

Firstly, we propose T-EBAnO, a novel explanation framework designed to identify and measure the importance of words in the predictions made by NLP classifiers. This framework provides both local explanations for individual predictions and global explanations for overall model behavior, thereby enhancing the transparency and interpretability of NLP classifiers.

Secondly, we propose NLPGuard, a framework aimed at mitigating the use of words related to protected attributes by NLP classifiers while maintaining their predictive capabilities. NLPGuard ensures fairness by reducing reliance on protected attributes, identified automatically through Large Language Model (LLM) prompts, thereby making the approach both automatable and adaptable over time.

Thirdly, we propose DriftLens, a novel drift detection framework designed to monitor the robustness over time of deep learning classifiers handling unstructured data, including NLP. This unsupervised methodology efficiently detects drift without requiring labeled data, making it suitable for real-time applications and enabling continuous monitoring of model performance in real-world production environments. Additionally, it identifies the labels most affected by drift, facilitating detailed characterization and explanation of the changes.

Our research aims to significantly advance the field of NLP by tackling crucial challenges related to transparency, fairness, and robustness in NLP systems. Through the development of innovative techniques for explaining NLP model predictions, mitigating bias, and detecting concept drift, our work aims to enhance the trustworthiness and responsible deployment of NLP technologies across various real-world applications. This effort not only enhances the performance and robustness of NLP systems but also addresses broader ethical considerations in artificial intelligence, fostering inclusivity, fairness, and accountability in algorithmic decision-making.