DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **09 juin 2023**

Nom de famille et prénom de l'auteur : **Monsieur GACI Yacine**

Titre de la thèse : *« Sur la subjectivité, les préjugés et l'équité dans l'apprentissage des modèles linguistiques »*

Résumé

Avec la croissance stupéfiante des modèles linguistiques au cours des dernières années, la technologie du langage est en train de prendre le contrôle des procédures les plus influentes de la société moderne, comme le recrutement, l'enseignement, les affaires, la législation et les systèmes juridiques. Par exemple, au lieu d'engager un travailleur humain lent pour étudier des centaines de CV dans le cadre d'une offre d'emploi, un analyseur automatique de CV peut le faire en quelques minutes. Au lieu de perdre du temps et de l'argent dans des poursuites et des procès juridiques coûteux, les modèles linguistiques peuvent analyser les preuves et construire une argumentation adéquate pour les défendeurs au tribunal. Le succès récent des modèles de langage est dû à deux facteurs majeurs : (i) leur taille massive atteignant des centaines de milliards de paramètres comme GPT3 ou ChatGPT, et (ii) l'idée intelligente de les préentraîner sur des corpus textuels colossaux avec très peu d'annotation et de curation. Bien que le pré-entraînement sur des ensembles de données non étiquetés ait facilité l'adoption du langage humain par les modèles, il leur a également permis d'absorber facilement les croyances subjectives néfastes contenues dans ces corpus. En effet, de plus en plus de recherches signalent que les modèles de langage ont hérité d'une grande partie des préjugés sociaux et des stéréotypes humains contenus dans les ensembles de données. Par conséquent, les modèles de langage courent le risque de prendre le parti des candidats masculins dans les offres d'emploi (en raison du stéréotype qui présente les hommes comme plus compétents et plus habiles que les femmes) ; de discriminer les personnes de couleur dans les tribunaux (en raison du stéréotype qui présente les Noirs comme des partisans du crime et de la violence) ; sans parler du risque de propager ces stéréotypes aux enfants lorsque les modèles de langage sont utilisés dans des contextes d'enseignement. Dans cette thèse, nous cherchons à caractériser et à mesurer les préjugés sociaux encodés dans les modèles de langage, et à quantifier les dommages causés par la discrimination lorsque ces

modèles sont utilisés dans des applications en aval. De plus, nous proposons trois nouvelles méthodes pour réduire la quantité de biais des modèles de langage : BiasMeter, ADV-Debias et AttenD qui opèrent respectivement sur les données, les représentations vectorielles de texte et le mécanisme d'attention.
Contrairement aux stéréotypes, la subjectivité peut parfois être bénéfique aux modèles de langage. Par exemple, un agent conversationnel orienté tâche peut utiliser les attributs subjectifs des énoncés de l'utilisateur pour permettre une recherche subjective. De même, la subjectivité peut améliorer l'extraction d'opinions et d'émotions à partir de commentaires en ligne. Des recherches antérieures ont montré que le fait de ne pas modéliser explicitement la subjectivité dans les technologies linguistiques orientées vers l'utilisateur, telles que les chatbots et la recherche, entraîne l'insatisfaction des utilisateurs. Dans cette thèse, nous nous concentrons sur la recherche et la similarité textuelle, et proposons des méthodes pour les augmenter avec la subjectivité. Que ce soit pour la subjectivité désirable (attributs subjectifs) ou indésirable (biais, stéréotypes et préjugés), nous fournissons une évaluation et une validation approfondies des techniques proposées.
Mots clés:
Modèles Linguistiques, Subjectivité, Préjugés Sociaux, Stéréotypes, Equité, Débiaisage, Traitement Automatique des Langues (TAL), Apprentissage Profond.