

Inégalités d'accès à l'enseignement supérieur : Une contribution grâce à l'analyse statistique textuelle

Contexte et description du sujet

En France, les inégalités d'accès à l'enseignement supérieur en fonction du milieu d'origine sociale sont encore très importantes (Fack and Huillery, 2021; Bryer, 2025). Ainsi, on observe qu'une augmentation de 10 percentiles du revenu des parents est associée à une hausse de 5,8 points de pourcentage dans la proportion d'individus qui entrent dans l'enseignement supérieur (Bonneau and Grobon, 2022). Les inégalités sociales d'accès à l'éducation sont généralement appréhendées en fonction de la catégorie socio-professionnelle des parents, leurs diplômes, le revenu familial ou le lieu d'habitation (OCDE, 2020). Dans ce cadre, de nombreuses études quantitatives nous indiquent que malgré le faible coût monétaire des études supérieures en France, les catégories sociales les plus défavorisées ont un accès plus difficile à l'enseignement supérieur.

Ce projet de thèse vise à compléter les études existantes sur les inégalités sociales d'accès à l'enseignement supérieur grâce à une information novatrice particulièrement riche: le contenu des lettres de motivation que les lycéens déposent sur *parcoursup* lors de leurs vœux d'inscription dans le supérieur. En utilisant les développements récents de la statistique textuelle, ce projet de thèse a pour objectif de quantifier les inégalités sociales que le contenu de ces lettres reflète. En particulier, la thèse se focalisera sur les deux problématiques suivantes:

1. Les lettres de motivation que les lycéens déposent sur *parcoursup* sont-elles socialement différenciées? Quelles différences de thématique, de style mais aussi de forme, et de contenu implicite ces lettres contiennent-elles en fonction de l'origine sociale?
2. Conditionnellement aux caractéristiques observables, l'information véhiculée par les lettres permet-elle de prédire l'inscription et la réussite dans l'enseignement supérieur?

Notre première question de recherche vise à mesurer les différences dans ces lettres en fonction de l'origine sociale. De nombreuses recherches en psychologie, linguistique, marketing et finance ont montré que l'information textuelle est riche et complète l'information quantitative traditionnelle contenue dans les caractéristiques observables individuelles. Ainsi, l'information textuelle permet de saisir des facteurs essentiels de réussite individuelle mais qui sont difficilement observables habituellement comme les traits de personnalité (McAdams, 2001), les états émotionnels (Tausczik and Pennebaker, 2010) ou l'état d'esprit (Stone, Quirk, Gardener, Hutt, Duckworth and D'Mello, 2019). Dans le contexte d'inégalités éducatives, d'autres facteurs psychologiques comme la confiance en soi (Hakimov, Schmacker and Terrier, 2023), l'état d'esprit (Huillery, Bouguen, Charpentier, Algan and Chevallier, 2025) ainsi que l'esprit de persévérance (Duckworth, Peterson, Matthews and Kelly, 2007) peuvent être appréhendés grâce à l'information textuelle saisie par les lycéens. Les lettres de motivation permettent également de saisir des éléments plus tangibles comme la connaissance des filières et des débouchés, ainsi que celle des attendus des formations. Ces éléments sont des facteurs clefs de réussite, mais ils sont aussi très marqués socialement: les individus issus de milieux plus favorisés ont d'avantage accès à l'information et aux expériences formatrices à travers les réseaux familiaux, l'accès aux stages en entreprise, la réalisation de séjours à l'étranger par exemple. Ces éléments mentionnés dans les lettres de motivation sont des signaux directement envoyés aux responsables de formation, mais aussi des indicateurs d'origine sociale. Nous souhaitons évaluer l'importance de ces différences en fonction de l'origine sociale et de la filière de candidature.

La seconde question de notre projet vise à quantifier les conséquences des différences dans les lettres de motivation sur les trajectoires dans le supérieur. En particulier nous cherchons à savoir si

conditionnellement à l'information fournie par les caractéristiques individuelles observables, l'information textuelle permet d'améliorer la prédiction de l'inscription et de la réussite dans le supérieur. Parmi les facteurs précédemment énoncés, quels sont ceux qui sont discriminants pour appréhender l'inscription et la réussite dans le supérieur? Les individus issus de milieux défavorisés peuvent-ils combler un déficit d'expériences formatrices passées par une motivation supplémentaire? Si le contenu des lettres contient une information prédictive additionnelle aux caractéristiques observables, alors les lettres renforcent les inégalités. Dans ce cas, cet exercice de prédiction de la réussite à partir du contenu des lettres est d'autant plus intéressant, que l'on pourrait simuler l'effet d'une politique publique qui supprime la lettre de motivation et voir dans quelle mesure les inégalités d'inscription et de réussite s'atténueraient.

Méthodologie et données

Les thématiques de recherche proposées par ce projet de thèse s'appuient sur les outils récents de l'analyse textuelle (Gentzkow, Kelly and Taddy, 2019; Athey and Imbens, 2019; Ash and Hansen, 2023). On s'intéresse notamment aux techniques récentes de Natural Language Processing (NLP). Ces approches permettent, grâce à de l'apprentissage automatique, de révéler les thèmes principaux d'un très grand nombre de textes. On distingue en particulier deux types d'approches qui peuvent nous être utiles. D'une part, les analyses de *topics modeling* qui mesurent et révèlent des thématiques sous-jacentes à différents textes sans que l'analyste ait au préalable à définir lui-même quelles sont les thématiques communes aux différents textes. D'autre part, les analyses à partir de *dictionary based methods* permettent à partir d'une liste prédéfinie de thèmes, d'en mesurer l'importance dans les différents textes. Bien que répondant à des problématiques différentes, ces deux types d'approches se complètent et pourraient être utilisées dans ce travail doctoral.

En termes de données, les porteurs de projet ont obtenu un accès aux bases de données nationale de Parcoursup ainsi qu'aux données du SIES sur le devenir des étudiants par l'intermédiaire d'un accès au CASD.

Membres impliqués dans le projet :

Nafyssa Affari, doctorante à l'IAE de Nantes

Nicolas Pistolesi, Professeur des universités à IAE de Nantes

Bénédicte Rouland, Maître de conférence à l'IAE de Nantes