Etat d’avancement de la thèse intitulé :
Constitution de fils de discussion cohérents à partir de conversations issues d’outils professionnels de communication et de collaboration.

La thèse porte comme son intitulé l’indique sur la constitution de fils de discussion cohérents et a débuté le 18 mars 2020 sur l’année académique 2019-2020.
Le démêlage de conversation est la problématique inhérente de cette thèse et qui depuis une décénnie est approchée avec différents outils et technologies dont les systèmes à base de règles, les approches statistiques et plus récemment les modèles d’intelligence artificielle. Cette problématique est très fréquemment étudiée avec des ressources qui sont qualifiées de : comminucations médiées par ordinateur qui sont en général des contenus de forum, de plateforme de tchat et d’emails. De part notre exploration de l’état de l’art nos travaux font partir des premiers qui abordent la problématique de démêlage de conversation sur les emails.

Chez Orange Innovation où j’effectue ma thèse, les emails sont la source de données que nous avons choisis d’exploiter pour la résolution de notre problématique parce qu’au début de ma thèse, était l’outil principal de communication mais qui est entrain d’être progressivement remplacé par Teams. La solution à mettre en place dans le cadre de la thèse sera appliquer sur les emails d’Orange qui sont en langue française. L’une des premières difficultés de la thèse a été l’inexistance de corpus d’emails en français disponible sur internet et qui nous aurait aidé à mieux avancer dans nos recherches.

Les premiers mois de la thèse ont porté sur un état de l’art sur la problématique de démêlage de conversation, ses variantes et les différents corpus utilisées pour approcher cette problématique. Par la suite, nous nous sommes intéressés à la constitution d’un corpus Orange qui a été assez long à cause des différentes contraintes auxquelles nous avons faits face notamment le respect du RGPD et dele secret de la correpondance. Nous avons adressé des demandes de consentement à un certain nombre de collaborateurs chez Orange afin de pouvoir exploiter leur emails. Nous avons ainsi pu obtenir 120 accords de consentement qui ont permi de collecter environ 12k emails via un outil développé dans le cadre de la thèse sur le poste d’environ 5 collaborateurs.

Dans un soucis de pouvoir partager ce corpusdans le monde de la recherche nous avons publié un premier papier méthodologique inttitué “Building A Corporate Corpus For Threads Constitution” sur la constitution et la pseudo-anonymisation d’un corpus d’emails en entreprise.

Suite à la constituion de ce corpus, nous avons effectué des analyses de ce corpus et sa petite taille nous a poussé vers d’autres corpus mais cette fois en anglais et sur des problématiques sous-ajenctes comme la reconnaissance en actes de dialogue, les détections de structures discursives. Wikidisc-2013-fr, MRDA, BC3, Reddit et Enron sont ces corpus sur lesquelles nous avons et sommes entrain d’effectuer diverses expérimentations afin de pouvoir approcher notre problématique de démêlage de conversation.