*Actes de JEP-TALN-RECITAL 2024*
*Toulouse, France, 8 au 12 juillet 2024*

# Construction d'une mesure de similarité thématique non supervisée pour les conversations

Amandine Decker[1,2] Maxime Amblard[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) University of Gothenburg, Suède
amandine.decker@loria.fr, maxime.amblard@univ-lorraine.fr

RÉSUMÉ

La structure thématique d'une conversation représente la manière dont l'interaction est organisée à un niveau plus global que le strict enchaînement des interventions. Elle permet de comprendre comment la cohérence est maintenue sur le temps de l'échange. La création d'une mesure de similarité thématique qui donne un score de similarité à deux énoncés du point de vue thématique pourrait nous permettre de produire et d'analyser ces structures. Nous entraînons une mesure non supervisée, basée sur le modèle *BERT* avec prédiction de la phrase suivante, sur des conversations *Reddit*. La structure de *Reddit* nous fournit différents niveaux de proximité de cohérence entre des paires de messages, ce qui nous permet d'entraîner notre modèle avec une fonction de perte basée sur des comparaisons plutôt que sur des valeurs numériques attendues *a priori*. Cette mesure nous permet de trouver des ensembles d'interventions localement cohérents dans nos conversations *Reddit*, mais aussi de mesurer la variabilité en termes de thème tout au long d'une conversation.

ABSTRACT

**Building an Unsupervised Topical Similarity Measure for Conversation.**

The topical structure of a conversation gives insight on the way interaction is organised at a more global level than utterance sequences. It enablles us to understand how coherence is maintained throughout a dialogue. Creating a topical similarity measure which would give a similarity score to two pieces of interaction in terms of topic could enable us to analyse this structure. We train an unsupervised measure based on the *Next Sentence Prediction* BERT model on Reddit conversations. The structure of Reddit provides us different coherence levels between pairs of messages which allows us to train our model with a marginal ranking loss rather than with numerical values. This measure enables us to find locally coherent pieces of interaction in our dataset but also to measure the variability in terms of topic throughout a conversation.

MOTS-CLÉS : topic modelling, apprentissage non supervisé, corpus, dialogue, Reddit, similarité.

KEYWORDS: topic modelling, unsupervised learning, corpus, dialogue, Reddit, similarity.

## 1 Introduction

L'enchaînement des thèmes d'une conversation donne un aperçu de la manière dont l'interaction est organisée à un niveau global. Nous nous intéressons à la production d'une structure permettant d'analyser la dynamique des échanges au delà des énoncés eux-même. Pour parvenir à ces dernières, nous nous intéressons à la cohérence thématique entre messages. La première étape est la segmentation

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 362–375.

en thème qui se définit comme la division d'un discours en morceaux localement cohérents. La plupart du temps, cette segmentation est linéaire, c'est à dire que les morceaux thématiquement cohérents se suivent linéairement, et aucune ou peu de structure entre les morceaux n'est mise en évidence. Si cette approche peut donner de bons résultats sur des textes bien organisés et, dans une certaine mesure, sur des dialogues structurés tels que des comptes rendus de réunions, elle n'est pas aussi adaptée aux conversations informelles. Par conséquent, la production d'une segmentation hiérarchique est essentielle pour modéliser des dialogues spontanés.

Étant donné la complexité de la modélisation hiérarchique des thèmes, en particulier pour ce type de conversations, qui plus est entre plusieurs intervenants, nous proposons de construire une mesure de similarité en thème qui fournit un score de similarité entre deux messages fondée sur leur proximité thématique. L'utilisation de cette mesure permet de trouver des ensembles localement cohérents dans le dialogue, et de manière plus large de mesurer la variabilité thématique tout au long de la conversation. Puisque nous ne pouvons pas donner manuellement un score de similarité à des paires de messages, nous proposons d'utiliser des méthodes non supervisées pour entraîner cette mesure à partir de comparaisons. L'utilisation d'un ensemble de données où les interactions sont déjà organisées de manière hiérarchique nous fournit différents niveaux de cohérence que nous utilisons comme base d'entraînement. Pour cela, nous nous basons sur le média social américain *Reddit* que nous considérons comme un grand ensemble de messages, organisés en structures arborescentes, discutant d'une diversité de sujets et présentant différents formats de relations entre les messages (questions/réponses, discussion informelle, argumentation, etc.)

Nos contributions dans cet article sont de deux ordres : (1) la création d'un corpus à partir d'une extraction des données de *Reddit* et (2) l'entraînement d'une mesure de similarité thématique de manière non supervisée à partir de ces conversations. Nous appliquons également la mesure à notre ensemble de données pour identifier les ensembles de commentaires thématiquement cohérents, et nous procédons à une évaluation de cette cohérence par rapport au sujet initial le long de ces fils de discussion. Nous donnons finalement des perspectives d'utilisation de cette métrique pour construire des représentations du dialogue.

## 2 Background

Les médias sociaux sont largement utilisés dans la recherche sur le discours comme sources de données (Hamilton *et al.*, 2017; Baumgartner *et al.*, 2020; Balouchzahi *et al.*, 2023). Jovanovic & Leeuwen (2018) analysent la structure et le genre des dialogues tenus sur diverses plateformes de médias sociaux, ou encore Misra & Walker (2013) travaillent sur l'identification de l'accord et du désaccord dans les conversations à partir des échanges publics sur les médias sociaux. Diverses études s'appuient sur ces dialogues pour des tâches variées, par exemple sur la santé mentale (Gaur *et al.*, 2018; Turcan & McKeown, 2019; Naseem *et al.*, 2022), ou la reconnaissance des émotions (Balouchzahi *et al.*, 2023). L'activité sur les réseaux sociaux est donc considérée comme représentative de la dynamique de l'interaction dialogique. Si cela reste un usage particulier de la langue, cela nous permet de palier les difficultés de la production de transcription d'échanges réels.

Si *Reddit* n'est pas le média le plus étudié (Kathie Treen & Coan, 2022), sa structure le rend particulièrement pertinent pour notre problématique. En effet, les messages sur *Reddit* sont organisés de manière arborescente, avec un message initial (*post*) sur un sujet donné et des réponses à ce message ainsi que des réponses aux différentes réponses. À partir d'un *post*, des conversations se

mettent en place, formant une extension arborescente en lien avec le thème de départ. Sur cette structure, nous étudions la manière dont le thème évolue à partir du premier message au travers d'une séquence de réponses. Nous souhaitons également comparer l'évolution de différentes séquences issues d'un même message initial. La Section 3 détaille la structure utilisée par *Reddit*.

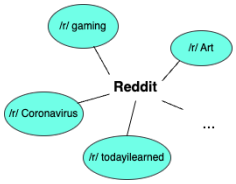
Afin d'étudier l'évolution des sujets dans notre ensemble de données, nous suivons (Wang *et al.*, 2017; Xu *et al.*, 2019; Wang *et al.*, 2020; Xing & Carenini, 2021) et construisons une mesure en nous appuyant sur la tâche de notation de la cohérence de paires de textes. La structure arborescente de *Reddit* nous permet de générer un corpus de paires de commentaires présentant trois niveaux de cohérence thématique. Pour cela, nous affinons le modèle *Next Sentence Prediction* (NSP) de BERT (Devlin *et al.*, 2019) qui utilise la fonction de perte fondée sur le *marginal ranking* pour obtenir une mesure de similarité thématique. Le modèle NSP BERT a l'avantage d'être entraîné sur un large corpus de texte non annoté, et en particulier sur une tâche de jugement cohérence. De plus, les grands modèles de langage (LLM), plus récents et plus puissants, basés sur les séquences, nécessitent des adaptations spécifiques afin d'être utilisés pour une tâche particulière comme l'évaluation de la cohérence. En effet, le *fine tuning* d'un LLM pour qu'il produise des scores de cohérence nécessiterait un ensemble de données incluant ces scores. Étant donné que l'attribution manuelle d'un score de cohérence à des paires de messages est une tâche complexe, s'appuyer sur différents niveaux de cohérence pour former un modèle est une solution de contournement qui nous parait plus prometteuse.

# 3 Données

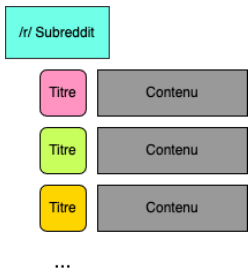
*Reddit* est un réseau social américain où les utilisateurs peuvent faire partie de communautés définies par des sujets particuliers comme un hobby, une culture, ou le format d'un *post* par exemple. La structure de *Reddit* est décrite dans la Figure 1. Dans les communautés, appelées *subreddits* (Figure 1a), un utilisateur peut proposer un *post* qu'il rédige (Figure 1b) et les autres utilisateurs peuvent réagir à ce *post* / interagir avec ce *post* en écrivant à leur tour un message, considéré comme un commentaire (Figure 1c). Un tel commentaire peut être une réponse directe à un *post* ou une réponse à un autre commentaire. La « discussion » ainsi créée à partir d'un *post* est organisée sous forme d'arbre dont la racine est le *post* initial et les noeuds sont les commentaires.

Nous utilisons cette structure pour former des paires de messages auxquelles nous attribuons plus ou moins de similarité thématique, et ce de façon automatique, comme nous l'expliquons dans la section suivante (Section 4). Bien que les messages récupérés ne soient pas à proprement parler du dialogue spontané oral, le format et l'organisation des commentaires nous permettent de considérer une dynamique d'interaction similaire à celle qui apparaît dans des conversations réelles. L'enjeu de la transcription est écarté, nous permettant de nous focaliser sur cette dynamique. Ces conversations sont plus ou moins longues selon le fil et impliquent un nombre variable de participants.

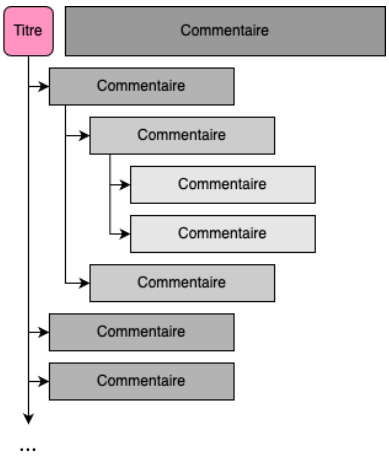
Le corpus est constitué du contenu de cinq thèmes différents (donc cinq *subreddits*, qui seront décrits en annexe dans la version finale) afin d'avoir de la variabilité des contenus et de la forme des messages. L'API *Reddit* nous permet de récupérer seulement une partie des données contenues dans un subreddit. Nous pouvons obtenir jusqu'à cent *threads* (un *post* et l'ensemble de ses commentaires) dans un *subreddit*, ces *threads* pouvant être ordonnés de différentes façons (par exemple en fonction du nombre de votes positifs qu'ils ont récemment reçu, ou en fonction de leurs votes positifs depuis la création du *subreddit*). Afin de récupérer des arbres de discussion de taille substantielle, nous avons décidé de récupérer les cent *posts* avec le plus de votes positifs depuis la création du *subreddit*. Ainsi,




(a) *Reddit* est divisé en *subreddits*.




(b) Chaque *subreddit* contient des *posts* formés d'un titre et de contenu.




(c) Chaque *post* est organisé sous forme d'arbre avec des commentaires répondant au *post* et des commentaires répondant aux autres commentaires.

FIGURE 1 – Organisation de Reddit.

pour chacun de nos cinq *subreddits*, nous avons récupéré jusqu'à cent *threads* [1] sous la forme d'arbres de discussion. Pour chaque *thread*, nous avons récupéré le titre et le contenu du *post* principal, son identifiant (ID), le nombre de votes positif, le pseudonyme de son auteur et la liste de ses 'enfants', c'est-à-dire l'ensemble des commentaires du *thread*. De même pour les commentaires nous avons récupéré leur contenu, leur ID, le nombre de votes positif, le pseudonyme de l'auteur ainsi que l'ID de leur père dans l'arbre de discussion afin de reconstruire la structure d'arbre. Nous n'utilisons ni le nombre de votes positif ni les pseudonymes des auteurs dans nos expériences mais les avons récupérés dans l'éventualité où ils s'avérerait utiles pour des analyses ultérieures. La licence de la *Data API* de *Reddit* ne nous autorise pas à partager notre dataset sans leur consentement explicite. Nous l'avons demandé mais n'avons pas encore obtenu de réponse (le délai minimum indiqué sur le formulaire est de 14 semaines). Un dataset similaire peut-être obtenu en utilisant notre code [2], bien que certains *threads* du top 100 pourraient avoir changé depuis que nous les avons récupérés. En cas d'accord d'ici publication, l'ensemble des données sera rendu accessible.

# 4 Modèle non supervisé

Notre objectif est donc d'entraîner un modèle pour construire une mesure de similarité thématique pour le dialogue. Comme nous ne pouvons pas attribuer manuellement une valeur de similarité thématique à des paires de messages, nous adaptons le modèle de Hearst (1997) à nos propres données. Ce modèle est entraîné grâce à une fonction de perte *marginal ranking*, ce qui signifie qu'au lieu d'apprendre à prédire un score précis pour une paire de commentaires donnée, il apprend à classer les paires de messages sur la base de leur similarité. Nous nous appuyons sur la structure hiérarchique

1. Certains *threads* nécessitaient plus de requête que l'*API* n'autorisait afin de récupérer tout l'arbre, nous avons donc décidé de les abandonner.
2. https://gitlab.inria.fr/adecker/reddittopicsimilarity.git

de *Reddit* pour construire trois niveaux de comparaisons, comme illustré dans la Figure 2. Nous considérons qu'étant donné un commentaire, un de ses fils est plus proche thématiquement qu'un de ses frères (donc l'oncle du fils), et un de ses neveux est moins proche thématiquement que ses fils et ses frères. Notre modèle utilise l'outil *Next Sentence Prediction* BERT (NSP-BERT) pour apprendre les scores de similarité. Lors de l'entraînement, le modèle reçoit les trois paires en entrée et produit un score de similarité pour chacune d'entre elles. Il est récompensé par la fonction de perte pour chaque score bien ordonné par rapport à l'un des deux autres.

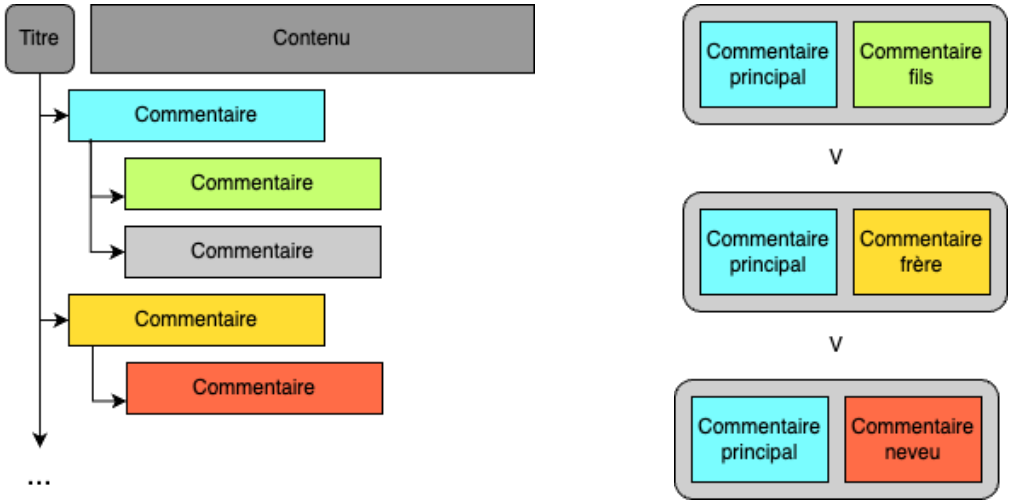



FIGURE 2 – Niveaux de similarité thématique entre les commentaires.

La structure de notre modèle est décrite dans Figure 3. Étant donné un commentaire $C$, nous récupérons un de ses fils, un de ses frères et un de ses neveux. Les quatre messages sont segmentés et les trois paires ($C$, fils), ($C$, frère) et ($C$, neveu) sont créées. Ces paires sont transmises à NSP-BERT dont la sortie est ensuite transmise à un perceptron multicouche. Nous utilisons la fonction d'activation linéaire rectifiée (ReLU) et un *dropout* de 10% entre les couches. La sortie finale est un score, ramené entre 0 et 1 avec une fonction sigmoïde, pour chacune des trois paires [3]

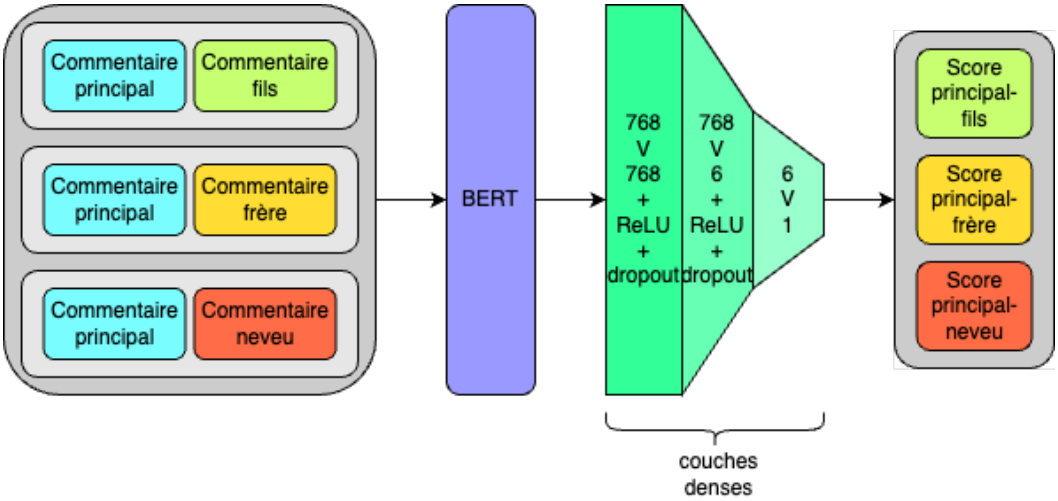



FIGURE 3 – Description du modèle.

3. L'ensemble des hyperparamètres est donné en Annexe B.

# 5 Expériences

Nous avons entraîné des modèles distincts pour chacun des cinq *subreddits* que nous avons récupérés. À chaque fois, l'ensemble des données est divisé en un ensemble d'entraînement et un ensemble de test dans les proportions 80%/20%. Afin de maintenir un équilibre concernant la taille des fils de discussion, nous les sélectionnons pour l'ensemble de test en les classant par taille, c'est-à-dire le nombre de commentaires qu'ils contiennent, et conservons un fil de discussion sur cinq. Pour la phase d'entraînement, nous extrayons également 10% de l'ensemble pour la validation.

## 5.1 Résultats

Comme expliqué ci-dessus, les données que nous utilisons pour former et évaluer notre modèle sont constituées de triplets de paires de messages (voir l'entrée du modèle sur la Figure 3). Pour construire ces triplets, nous sélectionnons au hasard un commentaire dans un fil de discussion et récupérons un de ses fils, un de ses frères et un de ses neveux. La plupart des commentaires ont plusieurs fils, frères et neveux. Par conséquent, pour maximiser la variabilité des données, nous ne générons pas tous les tuples possibles (commentaire, fils, frère, neveu), mais seulement l'un d'entre eux. Malgré cela, la quantité de commentaires dans notre ensemble de données nous permet de produire plus de triplets que nécessaire pour entraîner le modèle. Nous avons essayé trois tailles d'ensembles de données (1000, 7500 et 15000-triplets) pour l'entraînement afin de déterminer une quantité appropriée, en tenant compte du fait qu'un ensemble de données plus important nécessite plus de temps et de ressources pour entraîner notre modèle. Les meilleurs résultats ont été obtenus avec l'ensemble de données de 15000 triplets, mais l'amélioration par rapport à l'ensemble de données de 7500 triplets n'était pas suffisamment importante pour justifier l'utilisation d'un encore plus grand ensemble de données. Par conséquent, les résultats présentés dans la suite sont ceux que nous avons obtenus avec l'ensemble de données 15000-triplets.

Les résultats de nos expériences sont décrits dans le Tableau 1. Nous avons formé trois modèles par ensemble de données (c'est-à-dire 15 modèles au total). Nous les avons évalués sur tous les ensembles de données pour voir s'ils sont performants sur les messages d'autres *subreddits* qui diffèrent par leur forme (Choi *et al.*, 2015). Les résultats sont la moyenne et l'écart type des trois modèles. Le modèle utilisé est indiqué par le nom de la colonne et l'ensemble de données d'évaluation par celui de la ligne. Les résultats pour lesquels le modèle et l'ensemble de données sont congruents apparaissent sur la diagonale. Comme on pouvait s'y attendre, les résultats sont meilleurs dans ce cas (voir les résultats en gras dans Tableau 1), mais les résultats de l'application d'un modèle à un ensemble d'évaluation différent sont du même ordre (en restant inférieur).

La métrique obtenue est particulièrement sensible à la proximité thématique. Si elle varie entre 0 et 1, elle prend très fréquemment des valeurs proches des extremums se comportant de fait comme un bon classifieur thématique. Pour la suite, nous introduisons un seuil pour cette métrique que nous fixons à $0,5$ qui est une valeur classique et en adéquation avec nos observations.

## 5.2 Localiser les sous-*clusters* cohérents

Afin d'évaluer la qualité de la mesure obtenue, nous analysons la structure thématique des fils de discussion de *Reddit*. Nous avons donc appliqué notre métrique entraînée pour localiser des séquences

| Data \ Model | Art | Coronavirus | gaming | OnePiece | sports |
|---|---|---|---|---|---|
| **Art** | **65.80 ± 0.17** | 62.48 ± 0.68 | 63.72 ± 0.33 | 63.40 ± 0.51 | 63.71 ± 0.38 |
| **Coronavirus** | 66.01 ± 0.46 | **73.13 ± 1.42** | 67.62 ± 0.34 | 67.11 ± 0.70 | 67.42 ± 0.69 |
| **gaming** | 65.14 ± 0.59 | 64.42 ± 0.57 | **66.65 ± 0.26** | 65.04 ± 0.70 | 65.41 ± 0.39 |
| **OnePiece** | 66.01 ± 0.45 | 65.01 ± 0.24 | 66.06 ± 0.29 | **69.95 ± 0.62** | 65.77 ± 0.58 |
| **sports** | 64.91 ± 0.43 | 63.90 ± 0.57 | 65.07 ± 0.06 | 64.66 ± 0.45 | **66.43 ± 0.42** |

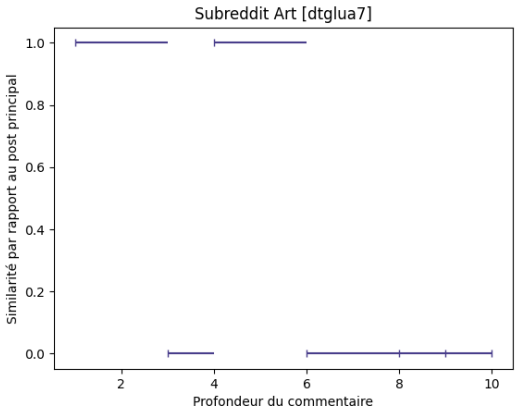
TABLE 1 – Résultats du transfert (moyenne ± écart type pour les trois versions de chaque modèle). *le gras indique les meilleurs modèles pour un ensemble d'évaluation.*

cohérentes de messages à l'intérieur d'un fil de discussion. Nous avons utilisé l'algorithme suivant :

— Dans une séquence donnée de commentaires, (sans regarder la structure horizontale des arbres de conversation,) nous calculons la similarité de chaque message avec son message suivant ;
— Chaque fois que la similarité est inférieure à un seuil $(0, 5)$, nous considérons que le message suivant est le début d'un nouveau sujet et qu'il contient le sujet dans son contenu ;
— Nous calculons la similarité entre le *post* initial et le premier commentaire de chaque nouveau *cluster* afin d'analyser l'évolution de la conversation.

Nous avons d'abord extrait toutes les séquences de commentaires de dix messages ou plus afin d'avoir des séquences assez longues pour observer des variations thématiques. La Figure 4 est une représentation de ce processus. Sur la Figure 4a, chaque ligne horizontale représente un *cluster* cohérent de commentaires et leur position verticale indique la similarité avec le *post* principal. Leur longueur dépend du nombre de messages. La Figure 4b reprend des exemples de commentaires ainsi que le *post* initial de la conversation. Nous voyons que le deuxième *cluster* est réduit au commentaire de profondeur 3 et est considéré comme thématiquement différent du *post* principal (*"They Don't Even Taste That Good Anymore (I), oil on canvas, 24x30""*). Cela a du sens car le commentaire ne contient que le mot *"Arrow"*. En revanche le *cluster* suivant est à nouveau proche du sujet d'origine, ce que nous confirmons avec le commentaire de profondeur 4 : *"It has to be good in the first place to stop being good, though."*. D'autres exemples seront donnés en annexe dans la version finale.

Le Tableau 2 présente différentes statistiques sur les séquences de messages et les *clusters*. Nous pouvons voir que le *subreddit Coronavirus* a peu de séquences que nous avons considérées comme suffisamment longues pour le regroupement. Tous les *subreddits* ont en moyenne un nombre similaire de messages par séquence, mais le *subreddit gaming* a quelques séquences très longues (au plus 275 messages) et les *subreddits Art* et *sports* ont également des séquences assez longues (environ 50 messages). En ce qui concerne l'évolution des thèmes, nous constatons qu'environ un tiers des premiers commentaires suivant un *post* sont proches de celui-ci, mais que près de 90% des têtes des




(a) Longueur et similarité au *post* initial des *clusters* identifiés.

**Post principal :** They Don't Even Taste That Good Anymore (I), oil on canvas, 24x30"
**Commentaire de profondeur 3 :** Arrow.
**Commentaire de profondeur 4 :** It has to be good in the first place to stop being good, though.
**Commentaire de profondeur 9 :** Fine, as long as you have a beard, like a proper dwarf.

(b) Commentaires en tête des *clusters* thématiques identifiés.

FIGURE 4 – Exemple de représentation d'une séquence de messages

derniers *clusters* sont thématiquement éloignées des *posts* d'origine. Cela semble soutenir l'idée que la conversation évolue naturellement au fur-et-à-mesure. Nous avons également calculé le nombre de fois où la similarité entre le *post* initial et la tête d'un *cluster* de commentaires est suffisamment différente de celle entre le *post* initial et la tête du *cluster* suivant. Ainsi nous mesurons la présence d'un changement thématique pour revenir au thème de départ ou s'en éloigner. Il y a en moyenne entre zéro et un changement de thème, ce qui nous indique que la conversation est soit considérée comme hors sujet par notre modèle dès le début, soit commence par être sur le sujet et évolue en s'éloignant du thème sans y revenir. Toutefois, ces structures ne sont pas les plus intéressantes. Les données contiennent également des séquences de commentaires comme ceux présentés dans la Figure 4 où un groupe proche du sujet principal est situé entre des groupes hors sujet. La structure de ces dialogues nous intéresse particulièrement car nous voulons suivre comment un sujet évolue dans une conversation. Pour mieux comprendre la structure thématique de nos séquences de messages, nous avons procédé à la modélisation globale de la conversation.

## 5.3 Hiérarchie de clusters

Pour aller plus loin dans notre analyse, nous avons regroupé les séquences cohérentes dans un fils de discussion. Pour cela, nous avons appliqué récursivement la méthode décrite précédemment mais sur les têtes des *clusters* au lieu de le faire sur chaque commentaire. L'algorithme est donc le suivant :

— Identifier les séquences cohérentes dans un fil avec l'algorithme de la Section 5.2 ;
— Calculer la similarité du premier message de chaque groupe avec celui du groupe suivant ;
— Si la similarité est inférieure à un seuil, le deuxième groupe est considéré comme le début d'un nouveau thème ;
— Application de la même stratégie pour le nouveau groupement jusqu'à stabilisation.

La Figure 5 est une illustration de ce processus. Nous pouvons mettre en évidence une hiérarchie entre les *clusters* thématiques. En effet, nous localisons d'abord les sujets avec une granularité élevée et chaque couche supplémentaire de regroupement est liée à des sujets avec une granularité plus générale, ce qui constitue une première étape vers une modélisation hiérarchique des thèmes.

| Subreddit | Nb de séquences ($\geq$ 10 msgs) | Nb max | Nb moyen | Sim. initiale (<0,1 / >0,9) | Sim. finale (<0,1 / >0,9) | Nb de changements brutaux |
|---|---|---|---|---|---|---|
| **Art** | 1556 | 51 | 11,88 (±2,96) | 63% / 34% | 87% / 11% | 1,07 (±1,67) |
| **Coronavirus** | 58 | 18 | 11,62 (±1,95) | 69% / 28% | 84% / 12% | 0,55 (±0,90) |
| **gaming** | 9409 | 275 | 11,88 (±5,44) | 64% / 34% | 88% / 10% | 0,70 (±1,14) |
| **OnePiece** | 908 | 28 | 11,84 (±2,48) | 61% / 37% | 91% / 07% | 0,54 (±1,01) |
| **sports** | 5930 | 52 | 11,71 (±2,58) | 46% / 52% | 85% / 14% | 0,75 (±1,14) |

TABLE 2 – Statistiques sur les séquences de messages et les *clusters* thématiques formés.

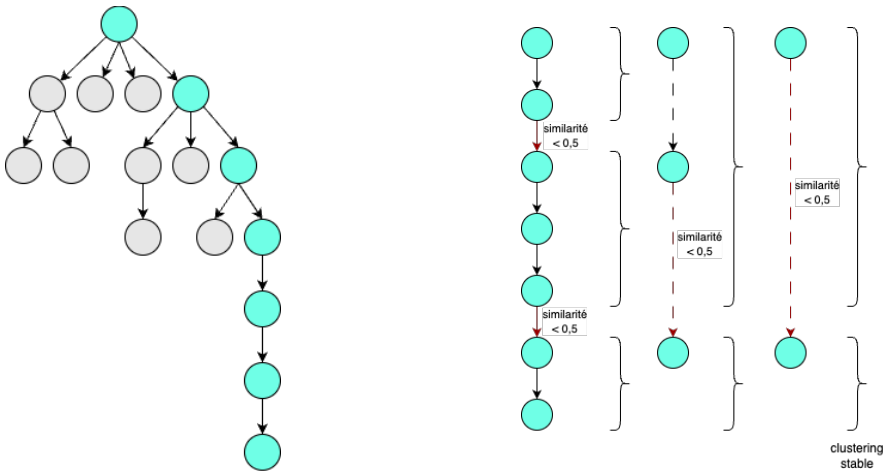
En suivant la méthode décrite précédemment, nous avons travaillé avec les séquences d'au moins 10 messages et calculé les différents niveaux de *clusters*. différentes statistiques sont reprises dans le Tableau 3. En ce qui concerne le nombre de niveaux de *clustering*, tous les *subreddits* nécessitent en moyenne trois niveaux. Le *cluster* de niveau le plus élevé contient en moyenne environ quatre *clusters* avec un écart-type élevé pour tous les *subreddits*, ce qui montre plus de variabilité sur cette question.

| Subreddit | Nb moyen de niveaux de *clustering* | Nb moyen de *clusters* finaux |
|---|---|---|
| **Art** | 3,03 (±1,34) | 4,59 (±2,97) |
| **Coronavirus** | 2,97 (±1,04) | 3,86 (±2,54) |
| **gaming** | 3,09 (±1,90) | 4,41 (±3,23) |
| **OnePiece** | 3,15 (±1,30) | 4,72 (±2,90) |
| **sports** | 3,14 (±1,33) | 4,30 (±2,91) |

TABLE 3 – Statistiques sur les hiérarchies de *clusters* formées.

# 6 Conclusion et Perspectives

Dans cet article, nous avons présenté une mesure de similarité thématique non supervisée. Nous avons montré qu'elle nous permet d'identifier des *clusters* de messages thématiquement cohérents au sein d'une conversation. Ce résultat correspond à la tâche de segmentation thématique linéaire.




(a) Extraction d'une séquence de commentaires.

(b) Regroupement selon la cohérence thématique à plusieurs niveaux.

FIGURE 5 – Processus de *clustering* hiérarchique selon les thèmes dans un fil de discussion.

Mais cette mesure nous permet également de construire une structure à plusieurs niveaux qui est une représentation hiérarchique des *clusters* thématiquement cohérents.

Notre mesure peut être utilisée pour une segmentation thématique précise puisque les morceaux de conversation seraient organisés logiquement en plus d'être séparés en fonction de leur sujet. À l'avenir, nous voulons enrichir notre représentation en étendant la représentation avec des relations rhétoriques de coordination et subordination entre *clusters*. Nous produirions une structure semblable à celles de la *Segmented Discourse Representation Theory* (SDRT, Lascarides & Asher (2007)) qui se focaliserait d'abord sur les liens thématiques. La comparaison de ces structures avec celles produites par des analyseurs de la SDRT (Li *et al.*, 2023) pourrait nous aider à comprendre l'influence des thèmes sur les représentations SDRT, et donner du sens aux changements de thème non adéquats.

Le regroupement que nous avons effectué était axé sur les séquences verticales de messages, en ce sens que nous avons extrait des séquences de messages plutôt que des sous-arbres complets de la conversation. La dimension horizontale doit également être explorée pour comparer l'évolution des sujets dans des fils de discussions parallèles et, par exemple, voir dans quelle mesure les conversations ont tendance à converger ou à diverger, et si ce comportement varie en fonction du *subreddit*. Cette application permettrait de mettre en avant dans l'ensemble d'un fil de conversation l'importance d'un thème donné, même si celui-ci apparaît peu dans chaque conversation. La récurrence transverse d'un thème permet d'identifier l'apparition de nouveaux sujets. Nous proposons également une structure d'analyse des échanges qui permet de mettre en avant des stratégies d'échange d'informations non explicite à la lecture des données. L'ajout d'un plus grand nombre de *subreddits* à notre ensemble de données nous permettrait d'effectuer des analyses sur une plus grande diversité de domaines, ce qui renforcerait la robustesse des résultats. Nous envisageons aussi d'entraîner de manière non supervisée et itérative la mesure pour gagner en précision. La sélection des messages pour l'entraînement ne serait pas laissée au hasard mais serait informée par la version de la métrique à l'itération précédente.

Enfin, contrairement aux modèles de classification binaires qui permettent seulement d'indiquer si deux segments sont thématiquement cohérents ou pas, notre modèle donne un score de proximité thématique. Cela pourrait permettre de repérer différents types de changements tels que des changements graduels qui ne sont pas repérables par classification binaire.

# Références


BALOUCHZAHI F., BUTT S., SIDOROV G. & GELBUKH A. (2023). Reddit : Regret detection and domain identification from text. *Expert Systems with Applications*, **225**, 120099. DOI : https ://doi.org/10.1016/j.eswa.2023.120099.

BAUMGARTNER J., ZANNETTOU S., KEEGAN B., SQUIRE M. & BLACKBURN J. (2020). The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, **14**(1), 830–839. DOI : 10.1609/icwsm.v14i1.7347.

CHOI D., HAN J., CHUNG T., AHN Y.-Y., CHUN B.-G. & KWON T. T. (2015). Characterizing conversation patterns in reddit : From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, COSN '15, p. 233–243, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2817946.2817959.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GAUR M., KURSUNCU U., ALAMBO A., SHETH A., DANIULAITYTE R., THIRUNARAYAN K. & PATHAK J. (2018). "let me tell you about your mental health !" contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, p. 753–762.

HAMILTON W. L., YING R. & LESKOVEC J. (2017). Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 1025–1035, Red Hook, NY, USA : Curran Associates Inc.

HEARST M. A. (1997). Text tiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1), 33–64.

JOVANOVIC D. & LEEUWEN T. V. (2018). Multimodal dialogue on social media. *Social Semiotics*, **28**(5), 683–699. DOI : 10.1080/10350330.2018.1504732.

KATHIE TREEN, HYWEL WILLIAMS S. O. & COAN T. G. (2022). Discussion of climate change on reddit : Polarized discourse or deliberative debate ? *Environmental Communication*, **16**(5), 680–698. DOI : 10.1080/17524032.2022.2050776.

LASCARIDES A. & ASHER N. (2007). Segmented discourse representation theory : Dynamic semantics with discourse structure. In *Computing meaning*, p. 87–124. Springer.

LI C., HUBER P., XIAO W., AMBLARD M., BRAUD C. & CARENINI G. (2023). Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In A. VLACHOS & I. AUGENSTEIN, Éds., *Findings of the Association for Computational Linguistics : EACL 2023*, p. 2562–2579, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-eacl.194.

MISRA A. & WALKER M. (2013). Topic independent identification of agreement and disagreement in social media dialogue. In M. ESKENAZI, M. STRUBE, B. DI EUGENIO & J. D. WILLIAMS, Éds., *Proceedings of the SIGDIAL 2013 Conference*, p. 41–50, Metz, France : Association for Computational Linguistics.

NASEEM U., DUNN A. G., KIM J. & KHUSHI M. (2022). Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, p. 2563–2572.

TURCAN E. & MCKEOWN K. (2019). Dreaddit : A Reddit dataset for stress analysis in social media. In E. HOLDERNESS, A. JIMENO YEPES, A. LAVELLI, A.-L. MINARD, J. PUSTEJOVSKY & F. RINALDI, Éds., *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, p. 97–107, Hong Kong : Association for Computational Linguistics. DOI : 10.18653/v1/D19-6213.

WANG L., LI S., LV Y. & WANG H. (2017). Learning to rank semantic coherence for topic segmentation. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1340–1344, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1139.

WANG W., HOI S. C. & JOTY S. (2020). Response selection for multi-party conversations with dynamic topic tracking. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6581–6591, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.533.

XING L. & CARENINI G. (2021). Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In H. LI, G.-A. LEVOW, Z. YU, C. GUPTA, B. SISMAN, S. CAI, D. VANDYKE, N. DETHLEFS, Y. WU & J. J. LI, Éds., *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 167–177, Singapore and Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.sigdial-1.18.

XU P., SAGHIR H., KANG J. S., LONG T., BOSE A. J., CAO Y. & CHEUNG J. C. K. (2019). A cross-domain transferable neural coherence model. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 678–687, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1067.

# A Description des subreddits utilisés dans notre dataset

Les descriptions ci-dessous sont toutes basées sur la description présente sur la page d'accueil des subreddits.

- Art : Une communauté où les utilisateurs parlent d'art et d'artistes, la description de la communauté insiste sur le fait de discuter d'art de façon mature et substantielle ;
- Coronavirus : Une communauté pour parler de la Covid-19, la description de la communauté demande des *posts* et des discussions de haute qualité ;
- gaming : Une communauté pour parler de jeux vidéos ;
- OnePiece : Une communauté pour discuter de toutes choses liées au manga One Piece de Eiichiro Oda et l'adaptation en anime ;
- sports : Une communauté pour parler de l'actualité sportive et de différentes ligues dans le monde comme la NBA.

# B Hyperparamètres du model

— *Architecture* :
  — NSP-BERT
  — *Multi-layer Perceptron* ($768 \rightarrow 768$, $768 \rightarrow 6$, $6 \rightarrow 1$)
  — ReLU
  — *Dropout* 10% ;
— *Optimizer* : Adam ;
— *Learning rate* : 2e-5 ;
— *Epsilon* : 1e-8 ;
— *Batch size* : 16 ;
— *Epochs* : 5 ;
— *Output* : fonction sigmoïde (valeur entre 0 et1).

# C Entraînement des modèles sur différentes tailles d'ensemble de données

TABLE 4 – Résultats (en %) pour différents nombres de triplets lors de l'entraînement

| **Subreddit** | **1 000 triplets** | **7 500 triplets** | **15 000 triplets** |
|---|---|---|---|
| **Art** | 59,25 ± 6,24 | 64,89 ± 0,23 | 65,80 ± 0,17 |
| **Coronavirus** | 67,42 ± 0,66 | 72,26 ± 0,66 | 73,13 ± 1,42 |
| **gaming** | 64,40 ± 0,04 | 65,34 ± 0,13 | 66,65 ± 0,26 |
| **OnePiece** | 65,39 ± 0,28 | 69,08 ± 0,43 | 69,95 ± 0,62 |
| **sports** | 64,22 ± 0,32 | 64,98 ± 0,31 | 66,43 ± 0,42 |

# D Exemples de représentations de séquences de messages

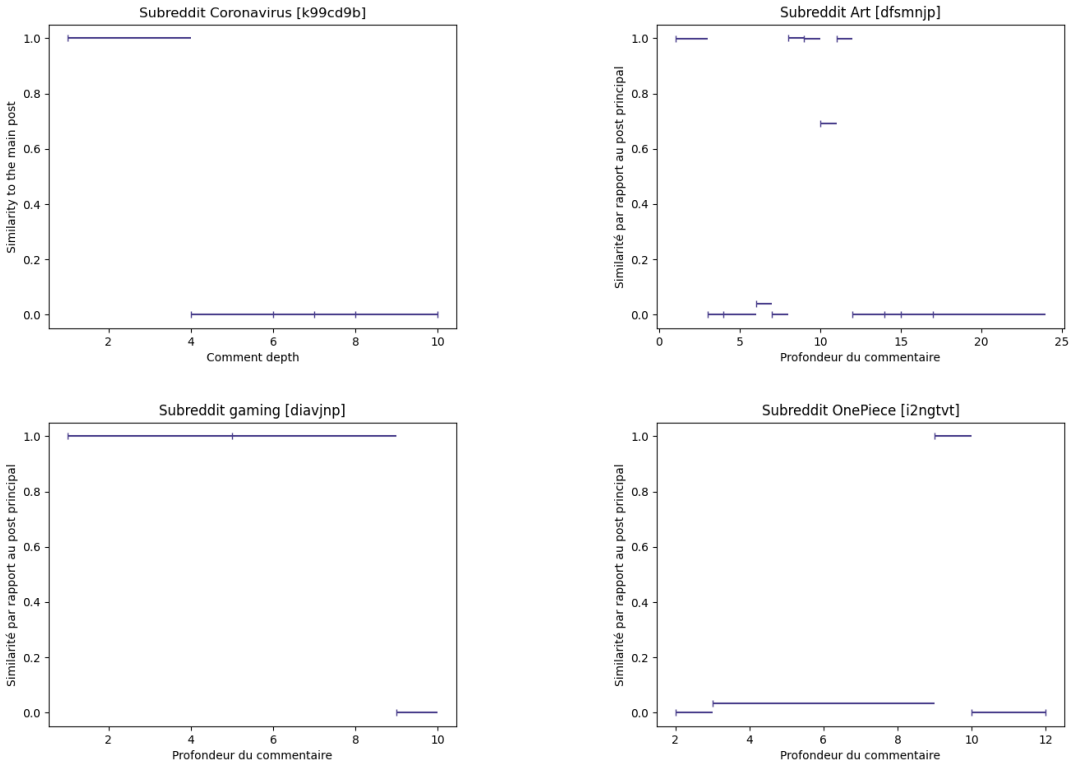



FIGURE 6 – Différentes structures thématiques.