# How COVID-19 is Changing Our Language: Detecting Semantic Shift in Twitter Word Embeddings

Yanzhu Guo, Christos Xypolopoulos, Michalis Vazirgiannis

# How COVID-19 is Changing Our Language: Detecting Semantic Shift in Twitter Word Embeddings

Yanzhu Guo[1], Christos Xypolopoulos[1,2], Michalis Vazirgiannis[1,3]

[1] Ecole Polytechnique, LIX
[2] National Technical University of Athens (NTUA)
[3] Athens University of Economics and Business (AUEB)

yanzhu.guo@polytechnique.edu

## Résumé

*Les mots sont des objets malléables, influencés par les événements reflétés dans des textes écrits. Située dans la contexte de COVID-19, notre recherche vise à détecter le changement sémantique dans le langage des médias sociaux provoqué par la crise sanitaire. À l'aide des données de grande taille liées au COVID-19 extraites de Twitter, nous entraînons des modèles de langue pour différentes périodes de temps avant et pendant l'épidémie. Nous comparons ces modèles de langue entraînés avec une série de mesures de similarité afin d'observer la variation sémantique. À l'aide d'une liste de mots saillants provenant de la mise à jour spéciale COVID-19 du dictionnaire anglais Oxford, nous déterminons l'approche de détection la plus adaptée à notre corpus Twitter. En nous appuyant sur cette approche, nous réalisons des études de cas sur un ensemble de mots sélectionnés par détection de sujets et visualisons l'évolution diachronique. Enfin, nous effectuons une analyse exploratoire sur des tweets français et obtenons des perspectives intéressantes sur des enjeux sociaux en France.*

## Mots-clés

*Changement Sémantique, Twitter, COVID-19, Traitement du Langage Naturel.*

## Abstract

*Words are malleable objects, influenced by events reflected in written texts. Situated in the global outbreak of COVID-19, our research aims at detecting semantic shift in social media language triggered by the health crisis. With COVID-19 related big data extracted from Twitter, we train word embeddings models for different time periods before and after the outbreak. We compare these trained embeddings with a series of different dissimilarity metrics in order to observe variation in semantics. Using a list of salient words from the COVID-19 special update of the Oxford English dictionary, we determine the detection approach most adapted to our Twitter corpus at hand. Drawing upon this approach, we carry out case studies on a set of words selected by topic detection and visualize the diachronic evolution. Finally, we perform an exploratory analysis on French tweets and gain interesting insights on societal issues in France.*

## Keywords

*Semantic Shift, Twitter, COVID-19, Natural Language Processing.*

## 1 Introduction

Les mots s'adaptent à l'environnement, leurs sens et leurs significations font l'objet de variations constantes, c'est-à-dire de glissements sémantiques. Comprendre comment ils changent à travers différents contextes et périodes de temps est crucial pour révéler le rôle du langage pendant l'évolution de la société.

Les glissements sémantiques peuvent être la conséquence de changements à long terme en raison de circonstances politiques, sociales, culturelles ou économiques. Un exemple est le mot "gay", dont le sens a changé au cours du 20e siècle, passant de la signification de "gai" et "joyeux" à celle d'"homosexualité" [25]. L'influence d'événements historiques ponctuels est tout aussi importante et peut entraîner des changements spectaculaires en un court laps de temps. Par exemple, l'événement tragique des attentats du 11 septembre 2001 a considérablement modifié l'interprétation générale du mot "terrorisme" [24].

La pandémie de COVID-19 a amené l'ensemble du discours mondial à se concentrer sur un seul sujet, cela s'était rarement produit auparavant. Les gens du monde entier ont dû adapter leur vocabulaire pour pouvoir parler de l'époque extraordinaire que nous vivons tous. Le glissement sémantique qui en résulte est illustré par le contenu généré par les utilisateurs sur les médias sociaux. En fait, l'un des facteurs contribuant à l'ampleur sans précédent du glissement sémantique déclenché par le COVID-19 et à son extraordinaire vitesse de propagation est le fait que la société humaine est connectée numériquement comme jamais auparavant. Les gens utilisent les médias sociaux non seulement pour rapporter les dernières nouvelles, mais aussi pour exprimer leurs opinions et leurs sentiments sur des événements du monde réel. Les utilisateurs montrent un intérêt particulier pour les situations d'urgence telles que COVID-19. Les données collectées sur des plateformes telles que Twitter peuvent constituer des ressources précieuses pour étudier l'effet de la pandémie sur le langage humain. Leur diversité et leur

compréhensibilité sont garanties par la nature inclusive des médias sociaux.

Dans cet article, nous explorons la stabilité sémantique des mots en calculant comment leurs significations, représentées par les embeddings de mots respectifs, ont été influencées par la pandémie. Nous collectons d'abord un corpus de tweets en anglais liés à COVID-19 ainsi qu'un corpus de référence de tweets hétérogènes en anglais postés entre janvier 2019 et décembre 2019. Nous générons des embeddings statiques ainsi que contextuels pour les deux corpus et calculons la stabilité des mots par deux métriques différentes basées sur les embeddings. En utilisant une liste de mots sélectionnés à partir de la mise à jour spéciales COVID-19 du dictionnaire anglais Oxford, nous évaluons les différentes approches et sélectionnons la plus efficace pour effectuer une analyse plus approfondie. Enfin, nous collectons un corpus de tweets en français liés à COVID-19 et réalisons une étude de cas pour la langue française.

# 2 Travaux antérieurs

Dans cette section, nous examinons la littérature connexe qui se divise en deux catégories : la détection du glissement sémantique et l'analyse de Twitter liée à COVID-19.

## 2.1 Détection du glissement sémantique

Alors que les premières études computationnelles du changement sémantique ont débuté par l'analyse des fréquences brutes des mots [11,13] et des cooccurrences [10], des études plus récentes ont principalement fait appel à des embeddings neuronaux de mots. Les approches classiques utilisent généralement des embeddings de mots statiques et peuvent être classées en deux groupes : l'utilisation de mesures de stabilité basées sur la similarité cosinus avec alignement ou l'utilisation de mesures de stabilité basées sur le voisinage sans alignement.

Les approches nécessitant un alignement sont généralement séparées en deux étapes : d'abord générer des modèles d'intégration de mots distincts pour chaque corpus indépendamment, puis utiliser des approches mathématiques pour les aligner dans le même espace latent sous-jacent. La qualité de l'alignement est déterminante pour les résultats des comparaisons. Kulkarni et al. [15] ont calculé la transformation linéaire optimale entre l'espace d'embeddings de base et l'espace d'embeddings cible en résolvant un problème de moindres carrés de k voisins les plus proches. Hamilton et al. [9] ont également utilisé des transformations linéaires pour l'alignement mais n'ont considéré que les transformations orthogonales. Zhang et al. [29] ont réalisé l'alignement de manière similaire, en ajoutant l'utilisation de mots d'ancrage, dont la signification est censée rester stable entre les deux espaces d'embeddings.

Les approches basées sur le voisinage reposent sur l'hypothèse que les mots ayant des significations significativement divergentes d'un corpus à l'autre sont censés avoir un contexte différent et donc un ensemble de voisins différent dans chaque corpus. Hamilton et al. [8] mesurent les changements des plus proches voisins d'un mot dans le but de capturer les changements drastiques dans le sens principal. Azarbonyad et al. [1] construisent un graphe pour chaque corpus avec les mots comme nœuds et les similarités entre eux comme arêtes. Ils calculent les similarités entre les voisinages d'un même mot dans différents graphes par des mesures de similarité basées sur le graphe. Gonen et al. [7] analysent directement l'espace de vocabulaire partagé des mots dans différents corpus, en considérant simplement les k plus proches voisins dans chaque corpus et en calculant la taille de l'intersection des deux listes. Cette méthode s'est avérée simple, interprétable et robuste.

Le succès récent des représentations de mots contextualisées telles que BERT [4] et ELMo [23] a ouvert de nouvelles portes aux chercheurs en détection de changements sémantiques. Hu et al. [12] construisent un embedding de sens distingué pour chaque sens d'un mot en s'appuyant sur des embeddings contextualisés profonds. En faisant correspondre les emeddings de chaque tranche de temps aux embeddings de sens construits, ils sont capables de suivre l'évolution du sens de chaque mot cible dans le temps. Giulianelli et al. [6] considèrent que chaque apparition d'un mot représente un sens différent. Ils déterminent comment les sens des mots varient dans le temps en affinant de manière incrémentée [14] sur des tranches de temps successives, puis en effectuant un regroupement K-means. Montariol et al. [20] améliorent l'évolutivité de l'approche précédente en fusionnant dynamiquement les clusters pendant la génération d'embeddings des mots. Martinc et al. [18] représentent directement le sens global d'un mot dans un corpus donné comme la moyenne de ses embeddings contextualisés et étudient leur évolution. Malgré la popularité croissante de l'utilisation d'embeddings contextualisés dans cette branche de la recherche, la récente tâche SemEval sur la détection non supervisée des changements sémantiques lexicaux [26] a montré que les méthodes d'embeddings statiques obtenaient les meilleures performances moyennes dans tous les corpus.

## 2.2 Analyse de Twitter liée à COVID-19

De nombreux articles s'intéressent à la réaction des utilisateurs de Twitter et d'autres médias sociaux face à cette crise sanitaire. Chen et al. [2] ont publié le premier dataset public de données Twitter sur le COVID-19. Cinelli et al. [3] ont traité de la diffusion de fausses informations concernant le COVID-19 sur Twitter. Ziems et al. [30] ont révélé l'origine et le mode de diffusion des comportements racistes en ligne pendant l'épidémie de COVID-19. Lopez et al. [17] ont permis de comprendre les réactions des gens aux politiques de COVID-19 en exploitant un ensemble de données Twitter multilingues. Müller et al. [21] ont publié COVID-Twitter-BERT, un modèle basé sur un transformateur pré-entraîné, avec un large éventail d'applications dans les tâches de NLP liées à COVID-19.

En ce qui concerne les glissements sémantiques, Tahmasbi et al. [28] ont entraîné des modèles Word2Vec hebdomadaires à partir de données Twitter collectées après l'épidémie et ont observé des glissements vers l'apparition d'injures plus sinophobes. Cependant, à notre connaissance, il n'y a pas eu d'étude systématique ou complète sur les glissements

sémantiques du langage des médias sociaux induits par le COVID-19.

# 3 Jeux de données et évaluation

Afin de suivre l'évolution des usages des mots, nous collectons deux corpus à grande échelle de tweets en anglais : un corpus de référence pré-COVID-19 avec les tweets publiés avant l'épidémie ainsi qu'un corpus lié à COVID-19 avec les tweets mentionnant explicitement la pandémie.

**Corpus de référence pré-COVID-19** Pour construire un corpus de référence qui remonte à l'époque pré-COVID-19, nous téléchargeons le flux Twitter général saisi par l'équipe d'archivage : [1], contenant des tweets diffusés de janvier 2019 à décembre 2019. Après avoir sélectionné uniquement les tweets en anglais et supprimé les tweets dédupliqués avec l'outil open-source runiq [2], nous obtenons un corpus de $127M$ tweets uniques.

**Corpus lié à COVID-19** Les tweets utilisés pour construire notre corpus relatif à COVID-19 datent de mars 2020 à août 2020. Dans ce cas, nos filtres se concentrent sur les tweets qui incluent les hashtags "covid19" et "coronavirus". Grâce à l'API publique de streaming de Twitter, nous avons extrait les tweets en anglais marqués par les deux hashtags ci-dessus. Après les mêmes étapes de filtrage linguistique et de déduplication que pour le corpus précédent, nous obtenons un corpus de $53M$ de tweets uniques.

**Évaluation** Afin d'évaluer nos différentes approches de détection du glissement sémantique, nous sélectionnons une liste de mots dont le glissement sémantique lié à COVID-19 a été confirmé par des lexicographes experts du dictionnaire anglais Oxford. Le dictionnaire anglais Oxford [3] a récemment publié plusieurs mises à jour spéciales COVID-19 dédiées à la langue de COVID-19. Nous sélectionnons tous les mots avec de nouveaux sens ajoutés et tous les mots (sauf les mots d'arrêt) qui appartiennent à de nouvelles entrées de phrases composées. Notre étude se concentre sur le glissement sémantique des mots existants et laisse le sujet de l'innovation lexicale et des néologismes pour un travail futur. La liste complète des 37 mots se trouve en annexe A.
Comme démontré dans des travaux antérieurs [8] [7], les méthodes automatiques de détection des glissements sémantiques ne peuvent que fournir une liste candidate de mots susceptibles de subir des glissements sémantiques mais ne peuvent en aucun cas rendre compte des garanties. Les cas d'utilisation appropriés de notre cadre de détection consistent notamment à offrir aux lexicographes un premier aperçu avec une liste de mots sur lesquels mener une enquête plus approfondie, ainsi qu'à sensibiliser les fonctionnaires aux récits sociaux en cours afin de communiquer avec le grand public de manière adaptée et efficace. Par conséquent, nous choisissons le score de recall de la détection des glissements sémantiques comme métrique d'évaluation, conformément à la motivation sous-jacente des scénarios d'utilisation.

1. https ://archive.org/details/ twitterstream
2. https ://github.com/whitfin/runiq
3. https ://www.oed.com

# 4 Méthodologie

Comme l'a montré la tâche SemEval sur la détection non supervisée des changements sémantiques lexicaux [26], aucun modèle d'embeddings ni aucune métrique de stabilité ne permet d'obtenir des résultats optimaux pour tous les corpus. Les performances des différentes approches dépendent fortement des caractéristiques spécifiques des corpus analysés. Par conséquent, nous combinons les modèles d'intégration de mots les plus avancés avec les métriques de stabilité les plus largement appliquées, dans l'espoir de découvrir l'approche la plus adaptée à nos corpus.

## 4.1 Modèles d'embeddings

Nous expérimentons des modèles d'embeddings de mots statiques et contextualisés, en prenant word2vec [19] et BERT [4] comme représentants respectifs.

**Word2vec** Nous utilisons l'implémentation open source de Word2Vec dans le package gensim [4] pour générer des embeddings de mots. Nous supprimons les mots qui apparaissent moins de 10 fois et appliquons l'architecture Skipgram, en choisissant une taille de fenêtre de 4 et une dimensionnalité de 300. Nous entraînons deux modèles word2vec distincts, indépendamment l'un de l'autre, pour le corpus de référence pré-COVID-19 et le corpus connexe COVID-19. Le pipeline de prétraitement des tweets est détaillé dans l'annexe B.

**BERT** Pour la modélisation du langage avec BERT, nous utilisons BERTweet [22], le premier modèle de langage à grande échelle pré-entraîné pour les Tweets anglais. BERTweet est entraîné sur l'architecture RoBERTa [16], en utilisant la même configuration de modèle que BERT-base. Le corpus original utilisé pour le pré-entraînement de BERTweet est constitué de $850M$ de tweets anglais, contenant $845M$ de tweets diffusés de janvier 2012 à août 2019 et $5M$ de tweets liés à la pandémie COVID-19. Nous utilisons le *bertweet-covid19-base-uncased* [5] version du modèle disponible sur Hugging Face. Cette version est le résultat d'un pré-entraînement supplémentaire du modèle original sur un corpus de $23M$ de Tweets anglais COVID-19 pour $40$ époques. Ce modèle répond bien à notre objectif car il a été entraîné de manière extensive sur les tweets pré-COVID-19 et les tweets liés à COVID-19, ce qui lui permet d'incorporer des sens de mots contextualisés pour les périodes pré- et post-pandémique.

## 4.2 Mesures de stabilité

Nous étudions à la fois les mesures de stabilité basées sur la similarité cosinus et celles basées sur le voisinage, qui s'avèrent plus performantes dans différents contextes : [8].

### 4.2.1 Mesure basé su la similarité cosinus

La similarité en cosinus est une méthode populaire en NLP pour estimer la similarité de deux vecteurs de mots. Cependant, pour les modèles **word2vec**, le problème de l'alignement est un élément clé de la comparaison d'embeddings de mots indépendants formés sur des corpus différents. En

4. https ://radimrehurek.com/gensim/
5. https ://huggingface.co/vinai/bertweet-covid19-base-uncased

| | Word2vecCos | Word2vecNN | BertCos | BertNN |
|---|---|---|---|---|
| recall | 0.78 | 0.73 | 0.19 | 0.22 |

TABLE 1 – Valeurs de recall de la détection du glissement sémantique pour les mots cibles.

| Approche | Les 5 premiers mots détectés |
|---|---|
| Word2vecCos | cerb, vtm, ggd, pums, adria |
| Word2vecNN | redzone, cerb, wha, corona, ceba |
| BertCos | unionism, fisa, bandage, carona, hoses |
| BertNN | drywall, flintstone, spfl, corny, trav |

TABLE 2 – Les 5 premiers mots détectés par chaque approche.

raison de l'invariance rotationnelle des fonctions de coût dans l'algorithme d'entraînement word2vec, les embeddings appris séparément sont placées dans des espaces latents différents. Cela n'affecte pas les similarités cosinus par paire au sein d'un même espace d'embeddings mais entrave la comparaison d'un même mot entre deux modèles différentes. La solution la plus recherchée est la méthode d'alignement des espaces vectoriels. En suivant Hamilton et al. [9], nous utilisons des Procrustes orthogonaux [27] pour aligner les embeddings de mots appris sur les mêmes axes de coordonnées. Nous faisons l'hypothèse simplificatrice que les espaces sont équivalents sous une rotation orthogonale. Plus précisément, nous définissons $W_{precovid} \in R^{d \times |V|}$ comme la matrice d'embeddings dans le modèle de référence pré-COVID-19 et $W_{covid} \in R^{d \times |V|}$ comme la matrice d'embeddings dans le modèle lié à COVID-19. Nous alignons $W_{covid}$ sur $W_{precovid}$ tout en préservant les similarités cosinus en optimisant :

$$R = argmin_{Q^T Q = I} \|W_{covid} Q - W_{precovid}\|_F$$

Nous résolvons ce problème d'optimisation par une application de la décomposition en valeurs singulières et obtenons la meilleure transformation rotationnelle orthogonale $R$ entre les deux espaces d'embeddings. Après avoir projeté les embeddings du modèle COVID-19 dans l'espace du modèle de référence avec $R$, nous pouvons calculer en toute sécurité les similarités en cosinus de tous les mots du vocabulaire partagé. La similarité en cosinus sert de mesure de la stabilité sémantique : plus la similarité en cosinus entre les embeddings d'un mot dans les deux espaces est élevée, plus sa stabilité est grande.

Comme pour les modèles **BERT**, la représentation de différents sens sémantiques est assurée par sa nature contextuelle. Nous n'avons pas besoin d'entraîner des modèles indépendants sur des corpus différents, il n'y a donc pas de problème d'alignement. Cependant, étant donné que l'architecture BERT consiste en 12 couches d'encodeurs retournant des sorties distinctes pour chaque instance de séquences d'entrée, il n'est pas simple d'obtenir un vecteur d'embedding unique pour chaque mot donné. Avec une approche similaire à celle de Martinc et al. [18], nous concaténons tous les tweets d'un même corpus et les séparons en séquences de 128 tokens. En introduisant ces séquences dans le modèle par des batches de 32 séquences, nous générons des embeddings de séquences en additionnant les quatre dernières couches de sortie de l'encodeur. Nous séparons ensuite les embeddings de séquence en 128 sous-parties, chacune correspondant à l'un des 128 tokens de la séquence d'entrée. Maintenant, chaque token a un vecteur d'embedding différent pour chaque instance de contexte dans laquelle il est apparu. Pour chaque token du vocabulaire du corpus, nous prenons la moyenne de tous ses vecteurs d'intégration comme représentation globale de sa signification sémantique dans le corpus donné.

#### 4.2.2 Mesure basé sur le voisinage

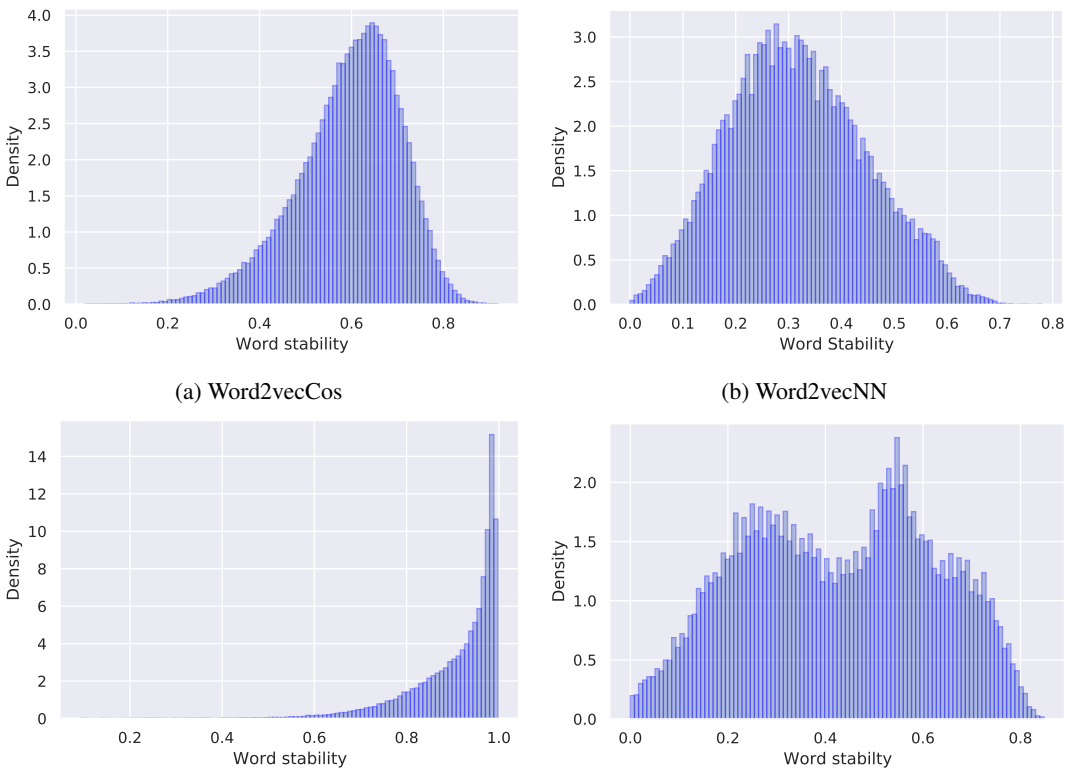
Plutôt que de tenter de projeter deux espaces d'embeddings dans un espace partagé, Gonen et al. [7] proposent de travailler dans l'espace de vocabulaire partagé. L'intuition sous-jacente est que les mots faisant l'objet d'un glissement sémantique sont susceptibles d'être interchangeables avec différents ensembles de mots, et donc d'avoir des voisins différents dans les deux espaces d'embeddings. Ceci donne lieu à un algorithme simple et efficace : chaque mot dans un corpus est représenté comme l'ensemble de ses k plus proches voisins (NN). La stabilité sémantique d'un mot à travers les corpus est simplement déterminée en considérant la taille de l'intersection des deux ensembles :

$$simNN^k(w) = |NN^k_{precovid}(w) \cap NN^k_{covid}(w)|$$

où $NN^k_i(w)$ est l'ensemble des k plus proches voisins du mot w dans l'espace $i$.

Pour calculer la liste des plus proches voisins, nous ne considérons que les mots du corpus dont la fréquence est supérieure au percentile $30\%$ afin de filtrer le bruit présent dans les données générées par les utilisateurs de médias sociaux. À la suite de Gonen et al. [7], k est choisi à 1000 afin de retenir plus d'informations globales.

Bien que cette méthode ait été initialement proposée pour les embeddings statique de mots, elle peut être naturellement étendue aux embeddings contextuels en calculant les deux ensembles de k plus proches voisins dans le même espace d'embeddings contextuels au lieu des deux espaces d'embeddings statiques indépendants.




(a) Word2vecCos (b) Word2vecNN

(c) BertCos (d) BertNN

FIGURE 1 – Distribution du score de stabilité sémantique obtenu par chaque approche.

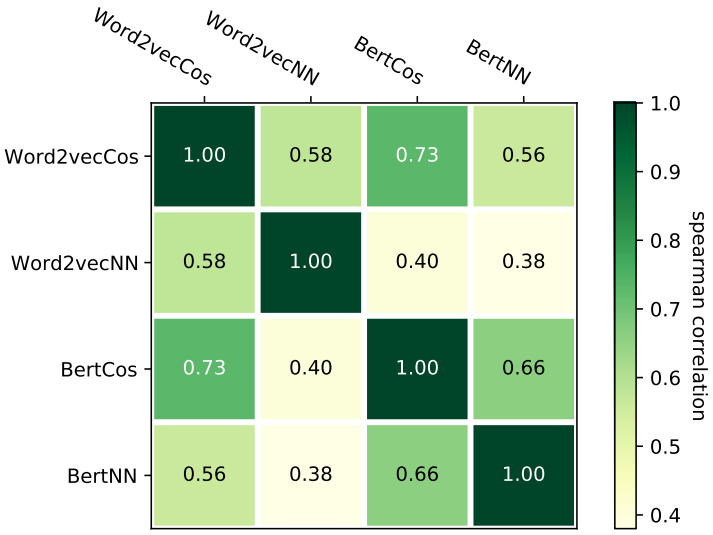



FIGURE 2 – Heatmap de la corrélation de Spearman entre les scores de stabilité des mots générés par différentes approches.

### 4.3 Approches étudiées

Dans la suite de ce document, nous ferons référence aux approches étudiées avec les noms suivants :

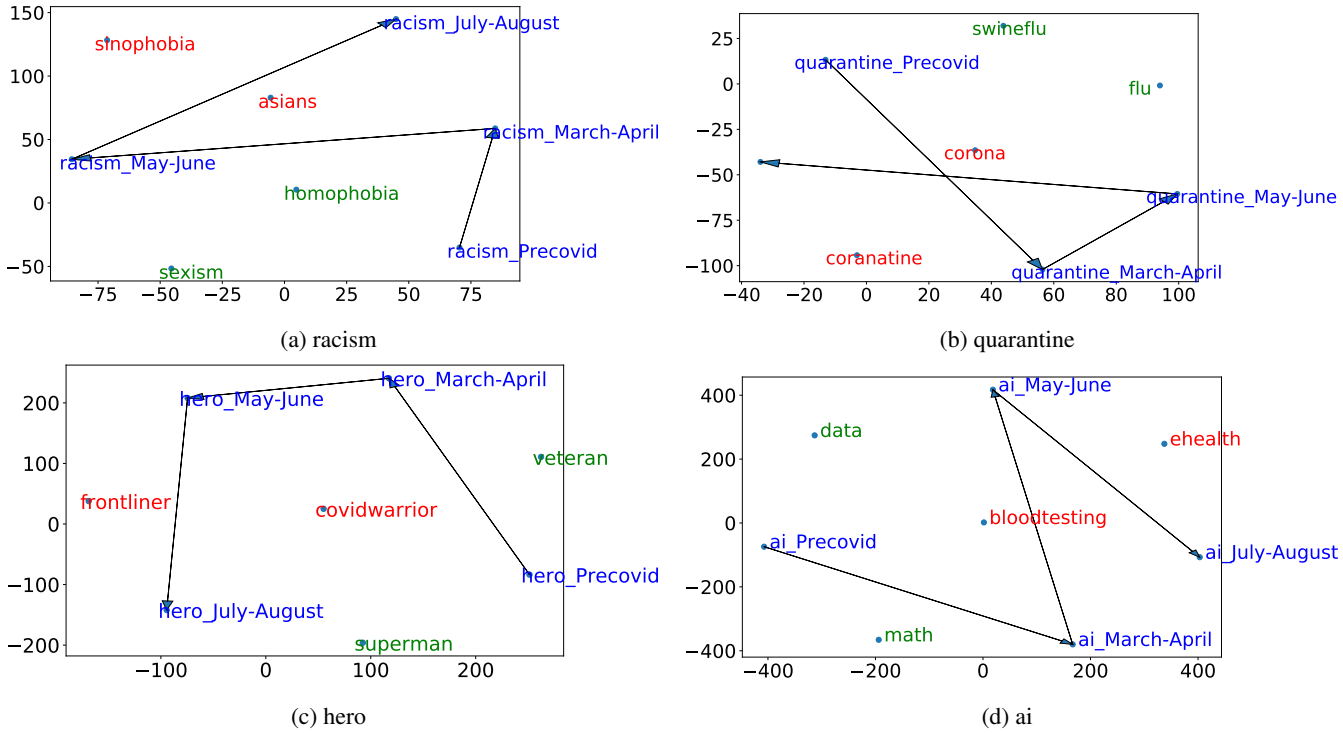
- **Word2vecCos** : combinaison d'embeddings word2vec avec la similarité cosinus.
- **Word2vecNN** : combinaison d'embeddings word2vec avec la similarité des plus proches voisins.
- **BertCos** : combinaison d'embeddings BERT avec la similarité en cosinus.
- **BertNN** : combinaison d'embeddings BERT avec la similarité des plus proches voisins.

## 5 Résultats

Nous calculons les scores de stabilité d'une liste de 37 mots avec un glissement sémantique confirmé, comme mentionné dans 3. Le centre d'intérêt n'est pas les valeurs absolues des

| Mot | S'éloigner | S'approcher |
|---|---|---|
| racism | sexism, homophobia | asians, sinophobia |
| hero | veteran, superman | frontliner, covidwarrior |
| quarantine | swineflu, flu | coranatine, corona |
| ai | math, data | ehealth, bloodtesting |

TABLE 3 – Trajectoires des mots étudiés.




(a) racism

(b) quarantine

(c) hero

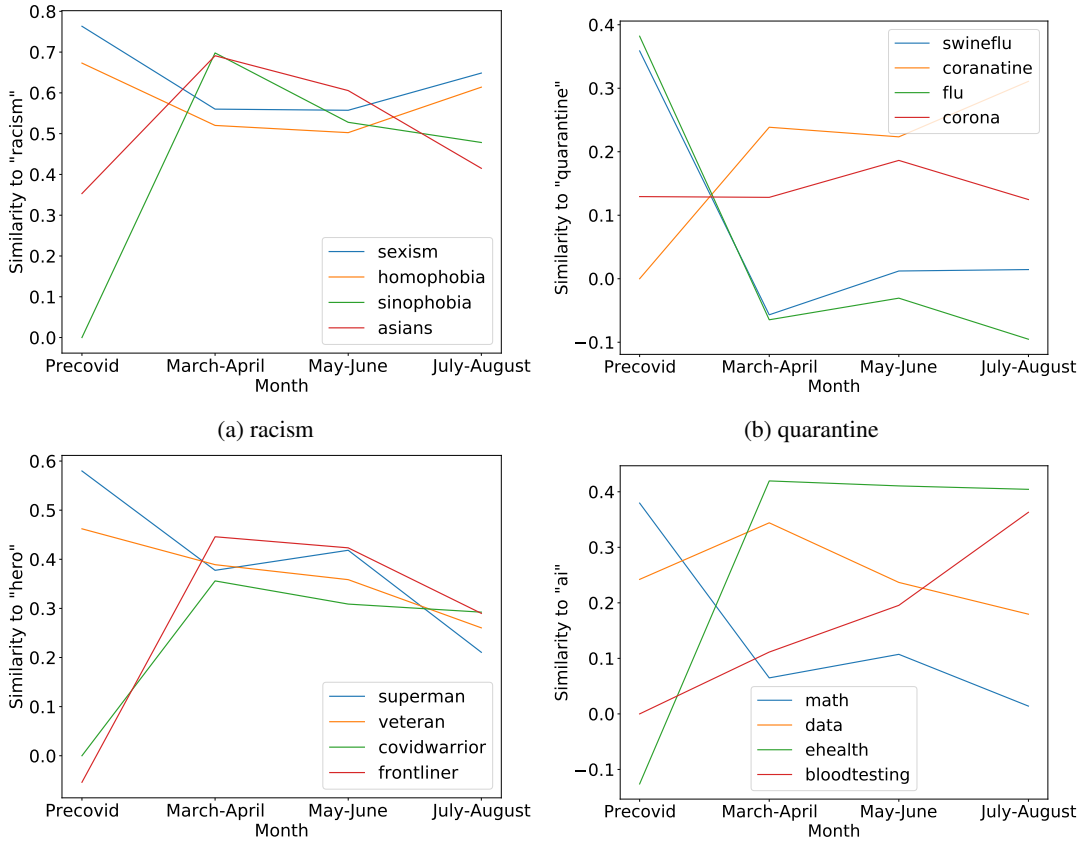
(d) ai

FIGURE 3 – Visualisation t-SNE des glissements sémantiques diachroniques dans les tweets anglais.

scores de stabilité eux-mêmes mais leurs positions relatives dans la distribution des scores de stabilité de l'ensemble du vocabulaire du corpus. Pour les mêmes raisons que dans 4.2.2, nous filtrons le bruit de la distribution en ignorant les mots à faible fréquence. Enfin, nous calculons le percentile du score de stabilité du mot cible dans l'ensemble de la distribution. En suivant la solution globale la plus performante de SemEval [26], nous définissons les mots dont le classement est inférieur au percentile 25% comme étant sémantiquement décalés. Les scores de recall et les distributions des scores de stabilité des mots cibles pour chaque approche sont respectivement présentés dans le tableau 1 et la figure 1. Le tableau 2 présente le top-5 des mots détectés par chaque approche. La corrélation de Spearman entre les scores générés par les différentes approches est présentée dans la figure 2.

Il est évident que les embeddings de mots statiques word2vec sont nettement plus performants que les embeddings de mots contextualisés BERT. Une explication possible est la caractéristique anisotrope des embeddings contextualisés. Par une série d'analyses géométriques, Ethayarajh [5] montre que dans toutes les couches de BERT, les représentations de tous les mots occupent un cône étroit dans l'espace au lieu d'être distribuées partout. Cette observation se manifeste également dans la figure 1 : les valeurs de similarité en cosinus entre les embeddings BERT moyens dans différents corpus sont extrêmement concentrées vers 1, la valeur la plus élevée. Avec tous les vecteurs serrés dans un cône étroit et la similarité entre tous les mots extrêmement élevée, cette méthode est plus sensible au bruit, qui est couramment présent dans les données des médias sociaux. Nous argumentons que les embeddings word2vec sont plus adaptés à nos corpus.

Il convient de souligner que la corrélation de Spearman (illustrée dans la figure 2) entre les classements générés par word2vec et BERT est étonnamment élevée étant donné l'écart significatif entre leurs scores de recall. Nous en déduisons que BERT peut effectivement générer des classements de décalage sémantique raisonnables pour les mots sémantiquement significatifs, mais que leur sensibilité au bruit fait qu'une grande partie des mots sémantiquement insignifiants sont classés dans la partie supérieure du décalage sémantique. Ces mots bruyants peuvent être le résultat d'une orthographe non standardisée ou d'une utilisation intensive




(a) racism

(b) quarantine

(c) hero

(d) ai

FIGURE 4 – Changements de similarité entre les mots sémantiquement décalés et leurs voisins.

d'abréviations et d'argots.

En ce qui concerne la comparaison entre les mesures de stabilité basées sur la similarité cosinus et les mesures de stabilité du plus proche voisin, leurs performances en termes de score de recall sont équivalentes, tandis que la corrélation entre leurs classements générés est relativement faible. Cela indique qu'elles sont toutes deux efficaces dans la tâche de détection des glissements sémantiques et que l'une peut compléter l'autre en détectant différents ensembles de mots.

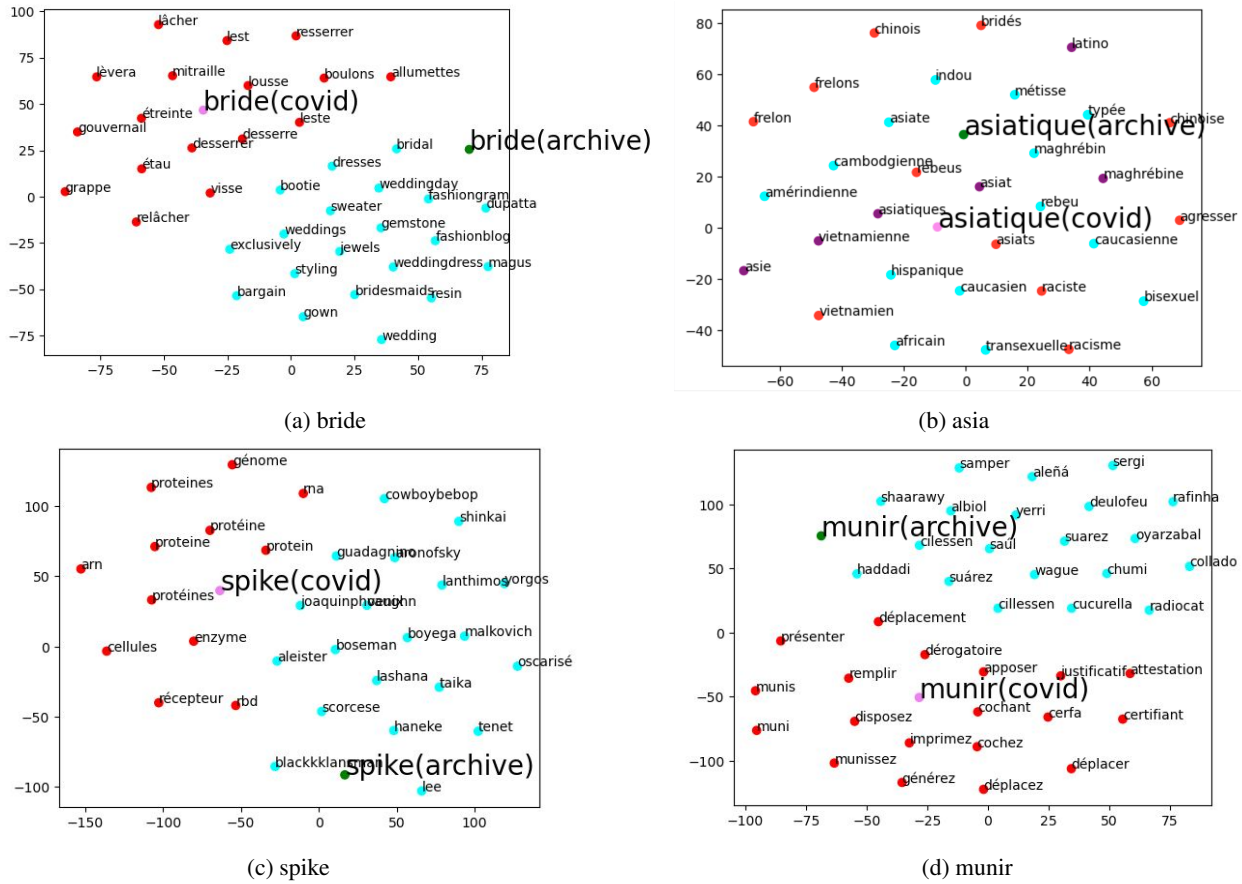
La liste des cinq premiers mots détectés dans le tableau 2 confirme notre analyse. Les mots détectés à l'aide de word2vec sont plus significatifs que ceux détectés à l'aide de BERT, tandis que la métrique de similarité cosinus et la métrique des plus proches voisins se complètent. Les principaux mots détectés par word2vec sont principalement des acronymes d'organisations ou d'événements liés à COVID-19. Par exemple, "Cerb" est l'acronyme de "Canada Emergency Response Benefit", tandis que "adria" est le nom d'un tournoi de tennis où plusieurs joueurs ont été testés positifs. Les mots détectés avec BERT comprennent plus de bruit dû aux fautes d'orthographe et aux abréviations, comme "carona" qui est une forme erronée de "corona" et "trav" qui est le diminutif de "traveling".

# 6 Analyse diachronique

Nous utilisons des modèles word2vec dans ce qui suit car ils sont plus adaptés à nos corpus. Étant donné que la pandémie évolue rapidement, nous construisons des embeddings word2vec mensuels pour chacune des paires de mois *mars-avril, mai-juin et juillet-août*.

Nous choisissons un ensemble de mots clés pour analyser l'évolution diachronique (*c.-à.-d.*, comment la sémantique fluctue dans le temps). Nous n'utilisons pas directement les mots ayant les scores de stabilité les plus bas renvoyés par l'algorithme car nous voulons un ensemble de mots étroitement liés à des questions sociales d'actualité. À cette fin, nous effectuons une détection des sujets parmi les tweets liés à COVID-19, puis nous réalisons des études de cas sur les mots clés des sujets détectés. Une autre raison pour laquelle nous procédons à la détection des thèmes est que les mots présentant le plus grand changement sémantique global ne sont pas forcément ceux qui présentent les fluctuations mensuelles les plus importantes. Nous choisissons finalement quatre mots clés : **"racisme"**, **"quarantaine"**, **"héros"** et **"ai"**.

Nous alignons les trois modèles word2vec mensuels liés à COVID-19 avec le modèle word2vec de référence en utilisant la méthode mentionnée dans 4.2.1. Nous visualisons les trajectoires des mots clés pour mieux comprendre leur évolution dans le temps. Le tableau 3 résume les glissements




(a) bride

(b) asia

(c) spike

(d) munir

FIGURE 5 – Visualisation t-SNE des glissements sémantiques dans les tweets français.

sémantiques de l'ensemble des mots clés tandis que la figure 3 montre leurs trajectoires. Nous traçons la projection t-SNE bidimensionnelle des mots clés dans chacun des modèles alignés. Nous traçons également certains des mots environnants des mots clés dans les modèles alignés. La figure 4 illustre les changements de similarité cosinus entre le mot d'intérêt et les mots environnants.

Dans tous les cas, les illustrations de la trajectoire et la variation de la similarité cosinus démontrent que les mots d'intérêt ont changé de sens de manière significative après l'apparition de COVID-19 et montrent également une évolution sur des périodes mensuelles. Par exemple, nous voyons le mot "racisme" s'éloigner d'autres concepts généraux de discrimination et se rapprocher de mots exprimant explicitement la haine envers la communauté chinoise/asiatique. Cela coïncide avec le phénomène anti-asiatique mondial observable depuis le tout début de la pandémie. Il est intéressant de constater sur la figure 4 que la similitude entre les concepts de racisme et d'anti-Asiatique a atteint un pic en mars-avril et a commencé à diminuer légèrement en mai-juin et juillet-août, ce qui indique que les gens retrouvent lentement leur rationalité à mesure que le stade de COVID-19 progresse.

# 7 Étude de cas des tweets français

En suivant les mêmes procédures que pour l'anglais, nous avons construit un ensemble de données de référence pré-COVID-19 composé de $34M$ de tweets français et un ensemble de données lié au COVID-19 composé de $19M$ de tweets français. En utilisant l'approche de détection optimale décrite dans les sections précédentes, nous effectuons une analyse exploratoire pour la langue française. Dans la figure 5, nous montrons les résultats de la visualisation t-SNE pour quatre mots-clés. Nous représentons les mots voisins dans les espaces d'intégration respectifs avant et après l'épidémie. Les points bleus représentent les mots voisins de l'espace avant l'épidémie et les points rouges de l'espace après.

La figure 5 :(a) montre un exemple typique de glissement sémantique. Avant la pandémie, le mot "bride" dans les tweets signifiait principalement le mot anglais "bride", une femme qui est sur le point de se marier. Il se trouvait dans la même zone de l'espace d'embeddings que des mots tels que "weddingday", "bridesmaids", "jewels", etc. Cependant, après la pandémie, le mot est devenu plus étroitement associé aux politiques restrictives gouvernementales, ce qui signifie "contrôle renforcé". Il se rapproche de mots tels que "lâcher" et "resserrer" dans l'espace d'embeddings.

Comme le montre la figure 5 :(b), nous constatons que le mot "asiatique" était plus proche des noms d'autres groupes

ethniques avant l'épidémie, tels que "hispanique", "maghrébine" et "caucasienne". Nous faisons l'observation troublante qu'après l'épidémie, il s'est rapproché de mots à caractère violent tels que "racisme" et "agresser". Cette observation est cohérente avec la montée de l'agressivité envers les communautés asiatiques déclenchée par la pandémie de COVID-19. De tels résultats analytiques peuvent alerter le gouvernement et les décideurs politiques sur les enjeux sociaux. En effet, la société française a récemment été témoin d'un certain nombre d'incidents violents graves à l'encontre des Asiatiques, notamment la proposition largement retweetée de "poignarder tous les Asiatiques que vous rencontrez dans la rue".

# 8 Conclusion

Dans ce projet, nous avons effectué une analyse sémantique comparative sur des modèles d'embeddings de mots formés à partir de tweets postés avant ou pendant la pandémie mondiale COVID-19. Une telle période constitue un bon point de référence pour étudier les changements sémantiques induits par des événements d'urgence dans une société en crise.

Nos principales contributions sont les suivantes :

(1) Nous avons montré que la pandémie COVID-19 a introduit des changements sémantiques notables dans le langage Twitter, en obtenant un ensemble de mots potentiellement décalés. Ces résultats peuvent éclairer les lexicographes, les décideurs politiques ainsi que le grand public.

(2) Nous avons réalisé une étude complète sur l'adaptabilité de différentes approches de détection des glissements sémantiques aux corpus de médias sociaux. Nous sommes parvenus à la conclusion que les embeddings word2vec sont plus efficaces que BERT, tandis que les métriques de similarité cosinus et de plus proches voisins peuvent se compléter. Si les méthodes d'intégration et les mesures de stabilité que nous avons choisies sont parmi les plus représentatives, certaines d'entre elles n'ont jamais été combinées dans des recherches antérieures. Néanmoins, notre étude est la première à aborder le problème de la conformité des méthodes aux données des médias sociaux.

(3) Nous avons démontré que les fluctuations sont continues de mars à août en visualisant les trajectoires d'un ensemble de mots-clés liés à COVID-19. Les résultats reflètent de manière vivante les questions sociétales en cours, attestant de la valeur de la détection des glissements sémantiques en sociologie.

(4) Nous avons réalisé des études de cas pour la langue française, ce qui nous a permis d'obtenir des informations intéressantes sur les questions sociétales et de combler les lacunes de ce type d'analyse en français.

(5) Nous avons construit une version anglaise et une version française des jeux de données Twitter lié à COVID-19. Chacun de ces jeux de données est accompagné d'un modèle d'embeddings de mots. Ces ressources sont utiles pour les recherches futures en matière d'analyse de Twitter.


# Acknowledgements

Nous remercions la Chaire ANR AML/HELAS et le projet ANR XTCOVIF pour leur soutien de cette recherche.


# Références


[1] Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. Words are malleable : Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518, 2017.

[2] Emily Chen, Kristina Lerman, and Emilio Ferrara. Covid-19 : The first public coronavirus twitter dataset. *arXiv preprint arXiv :2003.07372*, 2020.

[3] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific Reports*, 10(1) :1–10, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[5] Kawin Ethayarajh. How contextual are contextualized word representations ? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, 2019.

[6] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, 2020.

[7] Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, 2020.

[8] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift ? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access, 2016.

[9] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1489–1501, 2016.

[10] Gerhard Heyer, Florian Holz, and Sven Teresniak. Change of topics over time-tracking topics by their change of meaning. *KDIR*, 9 :223–228, 2009.

[11] Martin Hilpert and Stefan Th Gries. Assessing frequency changes in multistage diachronic corpora : Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4) :385–401, 2009.

[12] Renfen Hu, Shen Li, and Shichen Liang. Diachronic sense modeling with deep contextualized word embeddings : An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, 2019.

[13] Patrick Juola. The time course of language change. *Computers and the Humanities*, 37(1) :77–96, 2003.

[14] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, 2014.

[15] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*, 2019.

[17] Christian E Lopez, Malolan Vasu, and Caleb Gallemore. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset. *arXiv preprint arXiv :2003.10359*, 2020.

[18] Matej Martinc, Petra Kralj Novak, and Senja Pollak. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4811–4819, 2020.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv :1310.4546*, 2013.

[20] Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4642–4652, 2021.

[21] Martin Müller, Marcel Salathé, and Per E Kummervold. Covid-twitter-bert : A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv :2005.07503*, 2020.

[22] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet : A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 9–14, 2020.

[23] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv :1802.05365*, 2018.

[24] Stephen D Reese and Seth C Lewis. Framing the war on terror : The internalization of policy in the us press. *Journalism*, 10(6) :777–797, 2009.

[25] Justyna A. Robinson. A gay paper : why should sociolinguistics bother with semantics ? : Can sociolinguistic methods shed light on semantic variation and change in reference to the adjective gay ? *English Today*, 28(4) :38–54, 2012.

[26] Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. Semeval-2020 task 1 : Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, 2020.

[27] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1) :1–10, 1966.

[28] Fatemeh Tahmasbi, Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. "go eat a bat, chang !" : On the emergence of sinophobic behavior on web communities in the face of covid-19. In *Proceedings of the Web Conference 2021*, pages 1122–1133, 2021.

[29] Yating Zhang, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka. The past is not a foreign country : Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10) :2793–2807, 2016.

[30] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus : Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv :2005.12423*, 2020.

# A Liste des mots avec un glissement sémantique confirmé de l'Oxford English Dictionary

covid, coronavirus, corona, cv, infodemic, isolate, quarantine, distancing, comorbid, tracing, frontline, mers, zoom, elbow, bump, wfh, ppe, flatten, curve, cfr, spread, transmission, cytokine, facemask, mask, surgical, covering, reproductive, spike, shield, hydroxychloroquine, dexamethasone, bubble, stimulus, cpap, handgel, essential

# B Pipeline de prétraitement des Tweet

Chaque tweet utilisé dans le processus de formation de nos modèles word2vec est prétraité avec les étapes suivantes :

- **Conversion en minuscules** : Toutes les lettres de chaque tweet sont converties en minuscules.
- **Suppression des URL et des mentions** : Nous supprimons toutes les mentions et URL apparaissant dans les tweets.
- **Normalisation des symboles emoji et des caractères d'émoticônes** : Nous remplaçons tous les symboles emoji et les émoticônes par leur correspondance dans une liste de représentations normalisées.
- **Suppression des tweets courts** : Les tweets extrêmement courts, c'est-à-dire les tweets comportant moins de 10 tokens au total, ne contiennent pas de contenu significatif dans la plupart des cas et sont donc supprimés.