

Partie 1

Ci-dessous vous trouverez quelques phrases avec une ou plusieurs ambiguïtés potentielles :

1. Je manque de pièces pour payer le café.
2. Dans mon appartement, il y a une seule grande pièce.
3. J'ai cassé un vase. Il y avait des pièces partout.
4. Je suis allé voir une pièce hier avec une amie et elle était très ennuyeuse.

(a) Identifiez les ambiguïtés possibles, et identifiez le niveau de langage où les ambiguïtés se situent (sémantique, syntaxique ou pragmatique). Donnez l'interprétation la plus plausible en justifiant.

Reponse:

Ambiguïtés possibles et niveaux de langage :

1. "Je manque de pièces pour payer le café."
 - Ambiguïté : "pièces" peut se référer à de la monnaie ou à des morceaux manquants.
 - Niveau de langage : sémantique.
 - Interprétation plausible : Le sens le plus probable est que la personne manque de monnaie pour payer le café.
1. "Dans mon appartement, il y a une seule grande pièce."
 - Ambiguïté : "pièce" peut se référer à une pièce architecturale ou à un objet distinct.
 - Niveau de langage : sémantique.
 - Interprétation plausible : Il y a probablement une seule grande pièce architecturale dans l'appartement.
1. "J'ai cassé un vase. Il y avait des pièces partout."
 - Ambiguïté : "pièces" peut se référer à des morceaux ou à de la monnaie.
 - Niveau de langage : sémantique.
 - Interprétation plausible : Les morceaux du vase étaient éparpillés partout.
1. "Je suis allé voir une pièce hier avec une amie et elle était très ennuyeuse."
 - Ambiguïté : "pièce" peut se référer à une pièce de théâtre ou à une salle.
 - Niveau de langage : sémantique.
 - Interprétation plausible : La personne est probablement allée voir une pièce de théâtre ennuyeuse.

(b) Combien de sens de pièce pouvez vous identifier dans ces phrases ? Donnez un synonyme ou un hyperonyme (précisez) pour chaque sens.

Reponse:

(b) Sens de "pièce" et synonymes/hyperonymes :

- Monnaie : synonyme - "argent", hyperonyme - "numéraire".
- Pièce architecturale : synonyme - "chambre", hyperonyme - "espace".
- Morceau : synonyme - "fragment", hyperonyme - "partie".
- Pièce de théâtre : synonyme - "spectacle", hyperonyme - "représentation".

(c) Si on donne à un modèle de langue pré-entraîné, comme chatGPT, l'entrée Quelle est l'interprétation la plus probable de la phrase il continue comme suit (la phrase donnée est incluse dans la suite) : "Il est difficile de dire quelle est l'interprétation la plus probable du texte "On a skié

dans la soupe ce weekend.” sans plus de contexte. Si le contexte indique que la personne se trouvait dans un endroit où il y avait de la soupe, comme dans une station de ski où il y avait un restaurant proposant de la soupe, alors l’interprétation selon laquelle la personne a skié dans un endroit où il y avait de la soupe est la plus probable. Si, en revanche, le contexte indique que la personne se trouvait dans un environnement où il n’y avait pas de soupe, alors l’interprétation selon laquelle la personne a skié dans une substance ressemblant à de la soupe est plus probable.”

Sa réponse vous paraît-elle exacte ? Qu’est-ce que cela révèle des capacités ou limites de ce genre de modèle ?

Reponse:

Le modele n’as juste pas assez été entrainer sur les differentes experssions française et le jargon en general

Partie 2

L’arrivée du navire humanitaire Ocean Viking dans le port de Marseille en novembre 2022 a suscité de nombreuses réactions sur les réseaux sociaux au sujet de la politique migratoire de la France.

Pour analyser l’impact de ce flux migratoire sur la population et mesurer les stéréotypes que cela peut déclencher (par exemple, “les migrants sont violents”, “les migrants vont profiter des aides sociales”), on se propose de développer un système de détection automatique de messages de haine pour le français¹. On se focalisera sur le réseaux social Twitter.

-I-

On veut construire un classifieur qui est capable de reconnaître automatiquement si un message est haineux ou non.

(a) De quelles sortes de données d’entraînement a-t-on besoin ? Un stéréotype n’est pas forcément haineux et vice-versa.

Reponse:

- *Corpus annoté : Des messages provenant de Twitter avec des annotations indiquant s’ils sont haineux ou non.*
- *Caractéristiques linguistiques : Informations textuelles comme les mots utilisés, la syntaxe, la ponctuation, etc.*

(b) Détaillez brièvement la démarche à suivre pour collecter ces données d’entraînement ? Détaillez en particulier les précautions à prendre lors de cette collecte afin d’éviter les biais d’apprentissage.

Reponse:

- *Sélection du corpus : Choisissez un échantillon représentatif de messages Twitter en français.*

Annotation : Engagez des annotateurs humains pour marquer les messages comme haineux ou non.

- *Diversité des sources : Assurez-vous que le corpus provienne de différentes sources pour éviter les biais.*

(c) Détaillez brièvement la démarche à suivre pour construire ces données d’entraînement ? Comment mesurer leur qualité ?

Reponse:

- *Traitement des données : Nettoyez les messages, enlevez les informations non pertinentes, normalisez le texte, etc.*
- *Représentation des caractéristiques : Convertissez les messages en vecteurs numériques, utilisant des techniques telles que TF-IDF ou embeddings.*
- *Choix d'algorithme : Sélectionnez un algorithme de classification adapté, comme les réseaux de neurones, les machines à vecteurs de support, etc.*
- *Entraînement du modèle : Utilisez le corpus annoté pour entraîner le modèle.*
- *Évaluation de la qualité : Utilisez un ensemble de validation pour mesurer la performance du modèle. Assurez-vous que le modèle ne présente pas de biais.*

-II-

Après les phases de collecte et de construction, on se retrouve avec corpus de 10K tweets avec une proportion de messages haineux de 20%.

(a) Pensez-vous utile d'équilibrer les données. Si oui, expliquez comment procéder à cette équilibrage. Si non, justifiez.

Reponse:

Oui, il peut être utile d'équilibrer les données pour éviter un biais du modèle vers la classe majoritaire. Procédure : Sous-échantillonnage de la classe majoritaire ou suréchantillonnage de la classe minoritaire. Assurez-vous de maintenir une représentation adéquate des deux classes.

(b) Pensez-vous utile de pré-traiter les données. Si oui, quelles sont les pré-traitements qui vous semblent les plus pertinents. Si non, justifiez.

Reponse:

Oui, le pré-traitement peut améliorer la qualité des données et la performance du modèle. Pré-traitements pertinents :

- *Tokenization : Divisez les tweets en mots ou sous-mots.*
- *Suppression des stopwords : Éliminez les mots courants qui ne portent pas de sens.*
- *Lemmatisation : Ramenez les mots à leur forme de base.*
- *Suppression des caractères spéciaux et des liens : Pour simplifier le texte.*

(c) On souhaite rendre les données disponibles à la communauté dans un repo type GitHub ou Zenodo. Expliquez comment doit-on procéder et les précaution à prendre.

Reponse:

Procédure :

- *Créez un répertoire public : Sur GitHub ou Zenodo.*
- *Documentation détaillée : Incluez des informations sur la collecte, l'annotation et le pré-traitement des données.*
- *Licence appropriée : Choisissez une licence adaptée aux conditions de partage et d'utilisation.*

- *Évitez les données sensibles : Éliminez les informations personnelles et sensibles pour respecter la vie privée.*
- *Instructions d'utilisation : Fournissez des instructions claires sur la manière d'utiliser les données.*

-III-

On se propose maintenant d'entraîner trois classifieurs binaires sur nos données : (M1) Un premier classifieur baseline bag of words, (M2) un second classifieur à base de traits, (M3) un modèle neuronal.

(a) Expliquez comment fonctionne la baseline (M1).

Reponse:

Bag of Words (BoW) :

- *Le modèle BoW représente chaque document (tweet) comme un sac de mots, ignorant l'ordre des mots.*
- *Création d'un vocabulaire unique à partir de tous les mots dans le corpus.*
- *Chaque document est représenté par un vecteur où chaque élément correspond à la fréquence d'un mot dans le document.*
- *Un modèle simple, rapide à entraîner, mais perd l'ordre et la structure du langage.*

(b) Expliquez comment fonctionne le modèle (M2). Quels traits vous semblent les plus pertinents ? Quel algorithme d'apprentissage utiliseriez-vous ?

Reponse: Traits avec un algorithme d'apprentissage :

- *Utilise des caractéristiques spécifiques (traits) des données, comme la longueur du tweet, la présence de certains mots clés, etc.*
- *Traits pertinents dépendent du contexte, mais peuvent inclure la polarité des mots, la présence de hashtags, etc.*
- *Algorithme d'apprentissage : Peut être un algorithme de classification traditionnel comme les machines à vecteurs de support (SVM) ou un modèle ensembliste comme Random Forest.*

(c) Quelle architecture neuronale utiliseriez-vous pour battre (M1) et (M2) ? Expliquez les entrées de votre modèle (M3), les sorties qu'il doit prédire ainsi que le fonctionnement de cette architecture. Précisez la fonction loss.

Reponse:

Modèle neuronal :

- *Architecture : Utiliser un réseau de neurones récurrents (RNN) ou un modèle de transformer pour capturer les dépendances séquentielles.*
- *Entrées : Séquence de mots dans le tweet.*
- *Sorties : Probabilité que le tweet soit haineux ou non (classification binaire).*
- *Fonction loss : Cross-entropy loss.*

(d) Quelle métrique d'évaluation est la plus adaptée pour évaluer vos modèles ? Justifiez.

Reponse:

Métrique d'évaluation : F1-score peut être une métrique adaptée, car elle prend en compte à la fois la précision et le rappel, ce qui est crucial pour les tâches de classification déséquilibrée comme celle-ci.

(e) L'analyse des performances du modèle (M2) et celui proposé en (M3) montre que les scores diffèrent de 0.03% en faveur de (M3). Peut-on dire que le modèle (M3) est le plus performant ? Justifiez.

Reponse:

Comparaison entre (M2) et (M3) : Une différence de 0.03% en faveur de (M3) peut ne pas être significative. Il est nécessaire de considérer d'autres métriques et de faire des tests statistiques pour déterminer si la différence est statistiquement significative.

(f) La dernière étape est l'évaluation qualitative de vos modèles. Expliquez en quoi consiste cette évaluation.

Reponse:

Évaluation qualitative des modèles : Implique une analyse approfondie des erreurs de classification, une évaluation manuelle des prédictions, et peut-être des études de cas pour comprendre comment les modèles se comportent dans des scénarios spécifiques.

(g) Si on voulait une idée des informations qu'utilise le modèle (M3) pour prendre sa décision, quelle(s) méthode(s) pourrait-on utiliser ? Quel type d'information donneraient-elles ?

Reponse:

Interprétation du modèle (M3) : Méthodes telles que la Grad-CAM (Class Activation Mapping) peuvent être utilisées pour visualiser quelles parties du tweet le modèle considère comme importantes pour prendre sa décision. Cela permettrait de comprendre quelles informations spécifiques sont utilisées par le modèle pour prédire la haine

-IV-

On veut améliorer les performances de ce classifieur de sorte qu'il soit également capable de mieux appréhender le contexte d'énonciation des messages, c'est à dire prendre en compte des informations linguistiques pertinentes mais non présentes dans le message à classer. Par exemple la classification des messages "On doit tous les renvoyer chez eux", "On doit tous les renvoyer chez eux", "On doit tous les renvoyer chez eux" est rendu difficile par l'absence de contexte.

(a) Le classifieur proposé en (II) est-il adapté ? Justifiez votre réponse.

Reponse:

Adaptabilité du classifieur en (II) :

- *Le classifieur en (II) pourrait avoir des limitations pour appréhender le contexte d'énonciation, car il se base principalement sur les caractéristiques linguistiques des messages individuels sans tenir compte du contexte plus large.*
- *Le modèle neuronal (M3) peut capter certaines dépendances séquentielles, mais il peut ne pas être suffisamment robuste pour traiter des contextes plus complexes.*

(b) Proposez une solution pour adapter le classifieur initial pour prendre en compte un contexte d'énonciation donné. Attention, il n'est pas demandé une solution pour TOUS les contextes d'énonciation. Fixer un type de contexte et proposer une solution par rapport à ce contexte.

Reponse:

Adaptation pour prendre en compte le contexte d'énonciation :

- *Contexte choisi : Supposons que le contexte d'énonciation pertinent soit l'auteur du message.*
- *Solution : Ajouter une couche d'attention au modèle neuronal (M3) qui se concentre sur des parties spécifiques du profil de l'auteur, comme la bio, le nombre de followers, la fréquence des tweets, etc.*
- *Entrées du modèle (M3) : Séquence de mots dans le tweet et informations du profil de l'auteur.*
- *Fonction loss : La même fonction cross-entropy loss, mais avec une pondération appropriée pour les deux parties (tweet et profil).*

Partie 3

On va s'inspirer du texte suivant pour un problème d'extraction d'information, où on veut repérer des noms de modèles d'ordinateur dans des documents. Après avoir servi pendant la Seconde Guerre mondiale, Grace Hopper est devenue chargée de recherche à Harvard où elle a travaillé sur les ordinateurs Mark II et Mark III avant de passer dans le secteur privé. 2Dans ces exemples, et désignent des liens externes vers des images, articles de presse, etc.

En 1952, alors qu'elle supervisait la programmation de l'ordinateur UNIVAC chez Remington Rand, elle et son équipe ont créé le premier compilateur, un programme qui convertit les instructions du langage en code, en langage informatique, afin que les informations puissent être lues et exécutées par un ordinateur. On attribue également à Hopper la popularisation des termes «bug informatique» et «débogage». Ces termes sont nés d'un incident survenu pendant son séjour à Harvard, lorsqu'on a découvert qu'une mite avait court-circuité l'ordinateur d'Harvard Mark II.

(a) Dans le cadre d'une approche non supervisée, quel(s) patron(s) lexical(ux) d'extraction pourrait-on utiliser ? Evaluer la précision et le rappel des patrons proposés sur le texte exemple. Pouvez-vous trouver un patron avec un rappel de 1 sur le texte choisi ?

Reponse:

Approche non supervisée - Patrons lexicaux :

- *Patrons possibles :*
 - *“ordinateur [MODELE]”*
 - *“[MODELE] Mark [NUMERO]”*
- *Évaluation de la précision et du rappel dépend de la qualité des patrons choisis.*
- *Un exemple de patron avec un rappel de 1 sur le texte choisi pourrait être “[MODELE]”.*

(b) Comment pourrait-on trouver de nouveaux patrons ?

Reponse:

Recherche de nouveaux patrons : Utiliser des méthodes comme l'extraction de motifs fréquents, l'analyse de co-occurrence, ou l'utilisation de techniques basées sur des embeddings pour identifier des séquences récurrentes de mots liées aux modèles d'ordinateur.

(c) Si on veut développer un modèle supervisé pour le même problème, quel type d'architecture neuronale pourrait-on utiliser ? Préciser ce que serait une instance du problème, les entrées correspondant et les sorties que doit prédire le modèle, en prenant comme exemple une phrase (ou partie de phrase) pertinente du texte précédent.

Reponse:

Modèle supervisé Architecture neuronale :

- *Instance du problème : Prédire si une séquence de mots contient le nom d'un modèle d'ordinateur.*
- *Entrées : Des séquences de mots (phrases ou parties de phrases) extraites du texte.*
- *Sorties : Classification binaire indiquant si le nom d'un modèle d'ordinateur est présent ou non.*

(d) Si on donne la phrase suivante en entrée de BERT, « En 1952, alors que Grace Hopper supervisait la programmation de l'ordinateur UNIVAC chez Remington Rand, [MASK] et son équipe ont créé le premier compilateur. » Les trois mots les plus probables sont "Hopper" (98%), "lui" (0.4%), "elle" (0.1%). Pouvez vous expliquer pourquoi ?

Reponse:

BERT et les trois mots les plus probables :

- *BERT utilise le contexte environnant pour remplir le masque. Dans le contexte donné, la phrase mentionne "Grace Hopper" supervisant la programmation, ce qui indique que le mot masqué est probablement un pronom personnel lié à Grace Hopper.*
- *"Hopper" est le choix le plus probable car il correspond au contexte et au genre féminin. "Lui" et "elle" sont moins probables car ils peuvent ne pas convenir au contexte immédiat. La forte probabilité pour "Hopper" reflète la capacité de BERT à capturer les relations sémantiques et syntaxiques dans le contexte.*