TER Master 2 Informatique : Topic Modeling with Tweets for Pest Monitoring.

Proposition d’approche par classification semi-supervisée:

“L'apprentissage semi-supervisé est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées.”

- Wikipedia

Objectif:

L’objectif est de créer un modèle NLP basé sur un modèle pré-entraîné capable de déterminer si un message français sur twitter relève du domaine de l’observation d’un agriculteur ou non.
Pour cela j’utiliserais des données étiquetées et non étiquetées.
En se basant sur le principe des GAN un Discriminateur prendra en entrée la “hidden répartition" du modèle NLP pré-entraîné (c-a-d le vecteur de sortie avant la tête du modèle) ou celle d’un Générateur (à tour de rôle).
Pour le moment seul le Discriminateur et le Générateur seront entraînables.

Il y a aura trois cas de prédiction pour le Discriminateur:

- Si la donnée vient de la base labellisé il doit prédire:
    - Le label
    - Que la donnée est réel

- Si la donnée vient de la base NON labellisé il doit prédire:
    - Que la donnée est réel

- Si la donnée vient du Générateur il doit prédire:
    - Le label (cette tâche pourra être optionnel et l'on pourra tester la différence de résultat lors des évaluations quand elle est utilisée ou non)
    - Que la donnée n'est PAS réel (c-a-d générée)

Quant au Générateur, il s'entraîne en tentant de tromper le Discriminateur sur la réalité de la donnée.

Une fois l'architecture entraînée, le modèle final peut être composé. Ce modèle comportera le modèle NPL pré-entraîné (celui utilisé lors de la phase d'entraînement) dont la tête sera remplacée par le Discriminateur qui fera office de classifieur. La prédiction de la réalité des données sera ignorée et on s'intéressera seulement à la prédiction de la classe. (voir annexes)

Il y a plusieurs manières de construire le vecteur de sortie du discriminateur dans le cadre de ce projet:

1ère manière: Deux neurones activés en sigmoid dont chacun sera evalué avec la fonction d'erreur "binary cross entropy", Le premier classera le label (1: observation de l'agriculteur, 0: pas une observation) et le deuxième classera la réalité (1: donnée réel, 0: donnée générée). L'erreur total du Discriminateur sera une moyenne pondérée des erreurs des deux neurones (ce qui permettra lors de la prédiction sur les données non labellisé d'annuler l'erreur lié à la prédiction du label)

2ème manière: 3 neurones activés en softmax, les deux premiers déterminent le label et le 3ème la non réalité de la donnée. Ce cas me semble plus adapté pour une classification non binaire des labels (ce qui n'est pas le cas ici).

Je vais, pour commencer, me concentrer sur la 1ère manière de faire.

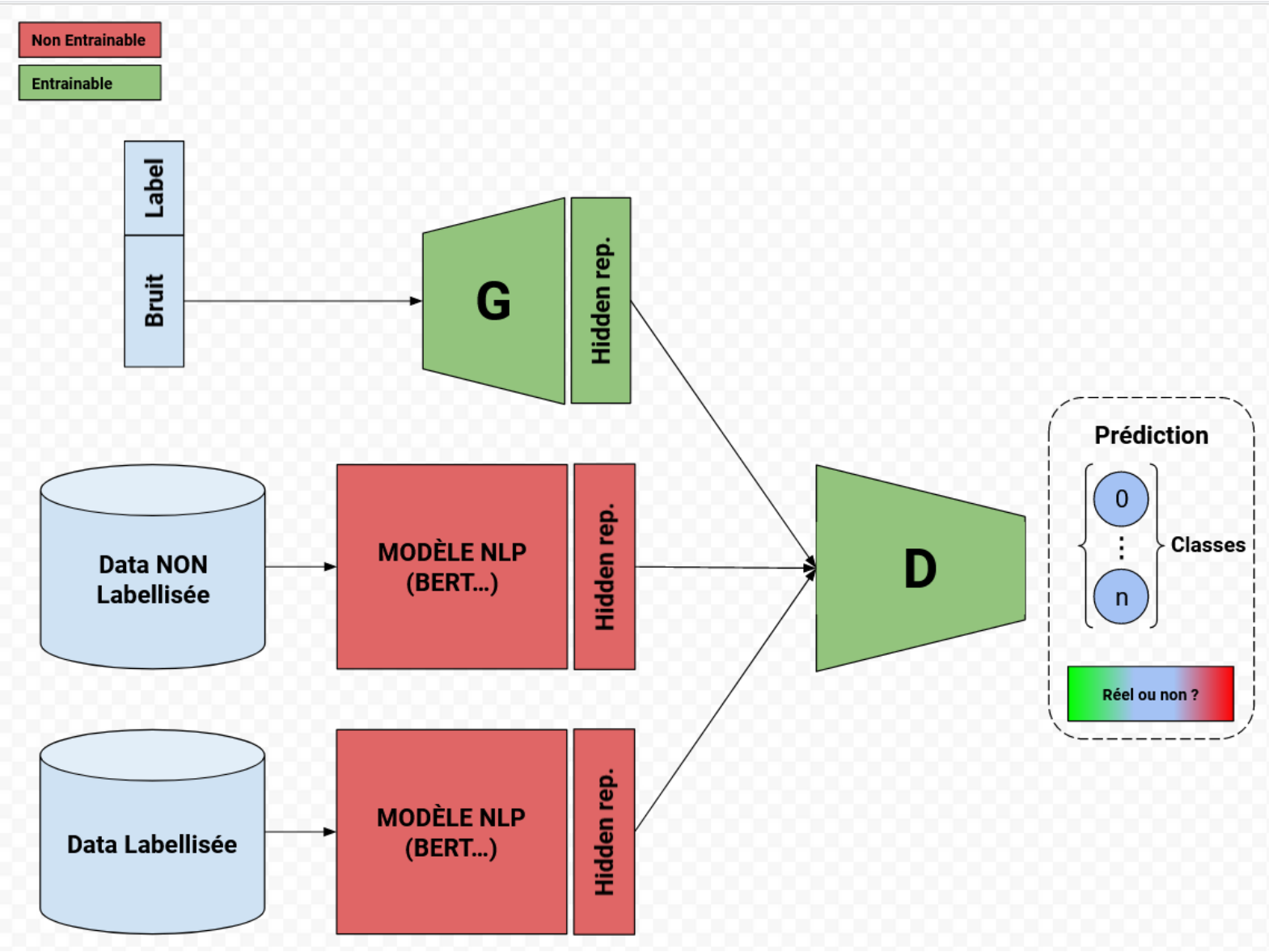
Plan de travail :

- (5h-10h) Etat de l’art sur la classification semi-supervisée, Semi-supervised learning GAN, GAN et NLP
- (2h-5h) Préparation des données
- (15h-20h) Développement de l’architecture (voir annexes)
- (5h-10h) Entraînement de l’architecture (voir annexes)
- (1h) Évaluation de l’architecture
- (10h-20h) Test d'hyper-paramètres / variations de l’architecture / Technique pour améliorer les GAN (extra-label, label bruité, flip labels, dropout…)
- (5-10h) Observer les variations de résultat en fonction de la taille de l'échantillons supervisé

Total: 43-71h

annexes:

Architecture en phase d'entraînement:



Architecture en phase de test / utilisation: