INTRODUCTION

-          Sanofi chatbot

-          Projet de certification pas applicable à mes missions au sein de sanofi (ibm watson) = je ne peux pas prendre un projet de sanofi pour ma certification

-          Projet d’analyse de sentiment pour analyser les retours de chatbots de sanofi mais le jdd ne permets pas de créer un datasets avec un bon entrainement (donne avis que si pb, langues)

PROBLEMATIQUE

-          Sentiment analysis (analyse de sentiment)

-          Entreprise de nom fictive

-          Besoin d’analyser la réputation de produits de prêt à porter des concurrents pour aiguiller des choix marketing

-           Formulation du besoin

- Stack technique qui répond au besoin (ou dans application)

GESTION DE PROJET (ou à la fin en bilan de projet)

-          Planning

-          Entretiens mensuels avec formateur qui joue le rôle du client

-          Entretiens avec tuteur en entreprise ?

- choix de la méthode de gestion de projet : waterfall/cascade ou cycle en v

- github

- travail en équipe ?

- compte rendus d’avancement

BASE DE DONNEES (Après EDA et nettoyage ?)

-          Tables

-          Relations 🡪 choix des jointures

-          Workbench MySQL

- CMD

- choix de la division dans la base de données

- script de load des données

- code du select avec les join

JEU DE DONNEES ET EDA

-          Description générale : qu’est-ce que c’est et d’où il provient

-          Nombre de features

-          Nombre de lignes

-          Valeurs manquantes (heatmap)

-          Duplicats (1 duplicat)

-          Selection des features

- Faire surtout des graphiques pour cette partie (pandas, matplotlib, seaborn)

- Analyse univariée et

- Analyse des données textuelles

TRAITEMENT DES DONNEES

-          Stopwords

-          Lemmatisation

-          Enlever ponctuation

-          Majuscules-minuscules

-          balance

MODELISATION

- Liste des modèles adaptés au NLP et choix

- Justifier le choix de l’algorithme sélectionné (en amont des performances) : en cohérence avec les données disponibles et le besoin, le but (classification)

- Code pour modeliser le modèle en annexe

-         Tableau recapitualtif des modèles testés

-          Mlflow (versionning)

-          Paramètres du modèle retenu

-          Metriques utilisées pour sélectionner le modèle

-          Performance du modele retenu

-          Gridsearch (optionnel)

-          (feature importance)

APPLICATION ET DEPLOIEMENT

-          Déploiement azure (+ abonnement ?)

-          Protocole de maintenance et de non regression

-          Test unitaires (unittest) (code en annexe) (notamment les tests pour le predict)

- Flask : code backend ? Droits et accès (POC donc ok)? Vérifier le texte en entrée dans l’application ?

- schema fonctionnel de l’application + spécifications fonctionnelle + conformité de l’application au schema

- maquettes front-end

- bdd de l’application : requetes nécessaires, accès aux données, code retourne les données

MONITORING

- Wandb

- alerte de message si seuil du modele trop bas

- quelle métrique choisi pour le monitoring

- des nouvelles données etiquetées arrive et on fait tourner le modèle sur les nouvelles données (5jours) et si performances du modèle trop bas alors alerte qui nous signiifie qu’ont doit réentrainer le modele

- Faire un script de monitoring ?

- partie critique e l’application à monitorer

- journaux sont stockés et organisés dans wandb et détailler les informations dans les journaux de wandb

AMELIORATIONS POSSIBLES

-          Web scrapping sur des sites de vente de prêt-à-porter pour accumuler une nouvelle base de données et étudier les produits les plus populaires

CONCLUSION