Analyse de sentiments de tweets dans le contexte de coupe du monde 2022 au Quatar

IA & DATA - YNOV
2022 2022

Contexte. Twitter est un réseau social virtuel où les gens partagent leurs publications et leurs opinions sur la situation actuelle, comme la pandémie de coronavirus. Il est considéré comme la source de données en continu la plus importante pour la recherche sur l'apprentissage automatique en termes d'analyse, de prédiction, d'extraction de connaissances et d'opinions. L'analyse des sentiments est une méthode d'analyse de texte qui a gagné en importance en raison de l'émergence des réseaux sociaux.

Par conséquent, dans ce projet, nous nous proposons de mettre en place un système visant à trouver le modèle d'apprentissage automatique optimal qui obtient les meilleures performances pour la prédiction des sentiments de coronavirus, puis à l'utiliser en temps réel. Deux objectifs principaux :

- Développer un système en temps réel pour prédire le sentiment exprimé dans les tweets liés à la pandémie de coronavirus en utilisant les données de streaming Twitter.

- Collecter des données de tweets sur le coronavirus à l'aide des hashtags # faisant réencense à cette dernière coupe du monde 2022), puis recycler les données des tweets.

- Suivre en temps réel et à travers un graphique, les sentiments des utilisateurs (Bonnus)

Nous envisageons donc une analyse selon une architecture à deux niveaux :

- Couche froide - Entrainement du modèle de prédiction sur des données froides. Les données sont donc collectées au préalable et stockées dans un espace de stockage. Cassandra peut servir pour la persistance des données. Par exemple, application de l'ensemble de données des tweets historiques collectées au cours des derniers mois.
-
- Couche chaude - Le modèle conçu est ensuite utilisé pour prédire les nouveaux tweets arrivant en temps réel. Plusieurs algorithmes d'apprentissages peuvent être explorés et se trouvent dans le package MLliB. Nous pouvons cibler une classification des sentiments en négatif/positif.

Le prétraitement des tweets

Le prétraitement des données est essentiel dans tout système d'analyse basé sur les réseaux sociaux (c'est-à-dire l'analyse des sentiments des données Twitter en continu) car il a un impact direct sur l'efficacité de l'analyse des sentiments en raison de la complexité des données. Twitter est considéré comme l'une des données les plus bruyantes car il est composé de nombreux liens, hashtags, symboles spéciaux, emojis, etc. Par conséquent, les données Twitter collectées doivent être prétraitées à l'aide des étapes suivantes : suppression du bruit, tokénisation, normalisation et racinisation, qui sont décrites ci-dessous :

Suppression du bruit. Dans cette phase, les données inutiles sont supprimées selon les étapes suivantes :

1- La mise en minuscules est la forme la plus efficace de prétraitement du texte, qui garantit la corrélation au sein de l'ensemble de caractéristiques. Par exemple, CovId et CovID doivent être convertis en COVID.
2- Suppression des URL. Dans cette étape, nous supprimons les liens non pertinents intégrés dans les messages Twitter.
3- Suppression des symboles spéciaux. Dans cette étape, nous supprimons les symboles spéciaux comme les ponctuations.
4- Suppression du hashtag. Le hashtag de Twitter est utilisé pour indexer des mots-clés ou des sujets sur Twitter, écrits avec le symbole #.
5- Suppression des mots vides (stop word). Les mots vides sont des mots insignifiants dans une langue et inutiles dans l'analyse du sentiment, qui sont utilisés pour la structure grammaticale de la langue. Nous filtrons ces mots d'arrêt, y compris les articles, les conjonctions, les prépositions, certains pronoms et les mots courants tels que « the », « about », « by », etc.

Tokénisation. Dans le prétraitement, la tokenisation consiste à décomposer les longues chaînes d'un texte en tokens (c'est-à-dire en petits morceaux). Ces tokens peuvent être des paragraphes qui peuvent être divisés en phrases plus courtes, qui peuvent à leur tour être divisées en mots. Par exemple, considérons cette phrase avant la tokénisation "je suis de toulouse" et après la tokénisation « je », « suis », « de », « toulouse ».

Normalisation. L'étape de normalisation du prétraitement consiste à transformer un texte en une forme standard pour augmenter l'uniformité du prétraitement du texte. Elle comprend la conversion de tout le texte en majuscules ou en minuscules.

Racinisation. Après l'étape de tokénisation, l'étape suivante est la racinisation. Cette étape consiste à ramener les mots dans leur forme originale (c'est-à-dire la forme racine pour réduire le nombre de types de mots ou de classes dans les données). Par exemple, les termes « Eating», "Eaten" et "Eater" seront réduits au mot "Eat".

La labélisation des tweets

L'analyse des sentiments identifie les émotions ou les attitudes de l'auteur (c'est-à-dire le pseudo/utilisateur de Twitter), que ces émotions/attitudes soient positives, négatives ou neutres. Nous utilisons TextBlob, une bibliothèque Python, pour effectuer une analyse des sentiments sur les données recueillies sur Twitter. Comme déterminé par le TextBlob, celui-ci utilise le modèle naïf bayésien pour la classification, et il renvoie deux propriétés en sortie, à savoir la polarité et la subjectivité La polarité se situe entre [-1,1], -1,0 et 1 définissent respectivement des sentiments négatifs neutre et positif. La subjectivité se situe entre [0,1]. La subjectivité quantifie la quantité d'opinions personnelles et d'informations factuelles contenues dans le texte. Une subjectivité plus élevée signifie que le texte contient des opinions personnelles plutôt que des informations factuelles. Nous utilisons les polarités de sortie des tweets pour étiqueter l'ensemble de données (tweets) collectées à intégrer dans les modèles d'apprentissage automatique.

Extraction de caractéristiques

L'extraction de caractéristiques est l'un des défis de l'analyse des données textuelles, en raison de l'apprentissage à partir de données de grande dimensionnalité. Il est préférable d'utiliser certaines méthodes d'extraction de caractéristiques pour convertir le texte en une matrice (ou un vecteur) de caractéristiques. Par conséquent, nous avons appliqué deux des méthodes d'extraction de caractéristiques les plus populaires sur les données de tweet collectées historiquement, à savoir, n-gramme et TF-IDF.

La modélisation n-gramme est une méthode populaire de sélection et d'analyse des caractéristiques largement utilisée dans l'exploration de texte et le traitement du langage naturel. Selon l'analyse des données textuelles, le n-gramme est utilisé pour calculer une séquence contiguë de mots de longueur n dans une fenêtre donnée. Dans ce travail, nous utilisons la méthode n-gramme, incluant n=1 à n=4 (c'est-à-dire unigramme, bigramme, trigramme et quadragramme) pour représenter le contexte des données Twitter.

La fréquence des termes - fréquence inverse des documents (TF-IDF) est une méthode célèbre utilisée pour évaluer le niveau d'importance d'un mot dans un document utilisé dans la recherche d'informations et le traitement du langage naturel. L'objectif de la méthode TF-IDF est de calculer la fréquence des mots dans le texte d'un corpus massif de documents. La méthode TF-IDF utilise le niveau de fréquence relative à travers le corpus de documents de référence, ce qui peut être considéré comme un grand mérite.

Modèles d'apprentissage automatique

Les modèles d'apprentissage automatique que nous ciblons sont les suivants : SVM, KNN et RandomForest.

Avant de faire appel à ces méthodes, nous divisons les données en deux parties appelées A et B contenant respectivement 90% et 10% de données.

Lancer 10 fois ces trois méthodes et à chaque expérience générer à partir un sous-ensemble d'entraînement ayant 91% de données de A et un sous-ensemble de validation ayant 9% de données de A. L'ensemble d'entraînement est utilisé pour optimiser et entraîner les modèles d'apprentissage automatique, tandis que l'ensemble de validation est utilisé pour évaluer les modèles d'apprentissage automatique durant les 10 expériences.

Après les 10 expériences par exemple, garder uniquement le modèle le plus efficace de chacune des trois méthodes d'apprentissage automatique. Enfin, évaluer chacun des trois modèles sur les données de l'ensemble B.

Veillez à bien vérifier que dans les deux sous-ensembles de A et l'ensemble B les tweets négatifs, Positifs et neutres sont significativement bien présents pour assurer par un apprentissage de qualité.

Pour mesurer l'efficacité des modèles de l'apprentissage automatique, vous êtes invité à consulter ce lien pour choisir au moins l'une des métriques d'évaluations, de préférence deux ou trois : https://spark.apache.org/docs/latest/mllib-evaluation-metrics.html#binary-classification

Optionnel (M2 uniquement) :
Nous envisageons de suivre en temps réel, les avis des clients sur les produits et les catégories les plus utilisés pendant la coupe du monde.

Le socle technique :

- ✓ Obligatoire : Kafka – Spark straiming, SGBD NoSQL,
- ✓ Optionnel : Nifi pour utiliser des fonctionnalisées ETL avant injection dans le SGBD