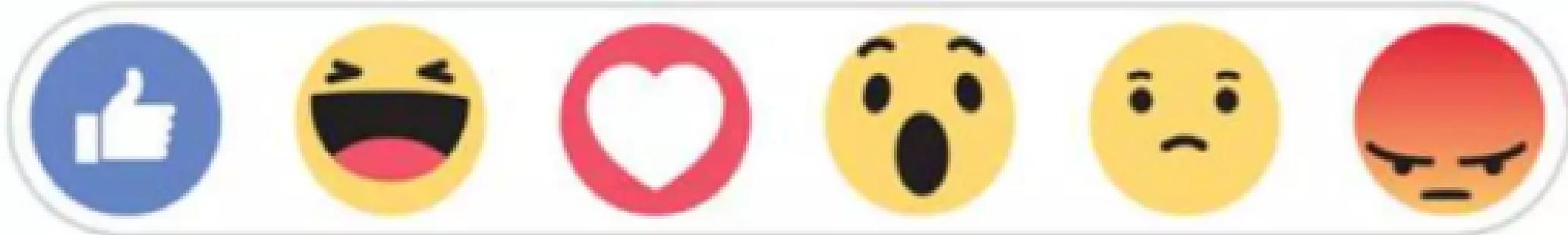


Présentation du text mining

Filière DATA SCIENCE & IoT

Sentiment Analysis



Présenté par :

BENANI GHIZLANE
LACHAM Fadwa

Professeur :

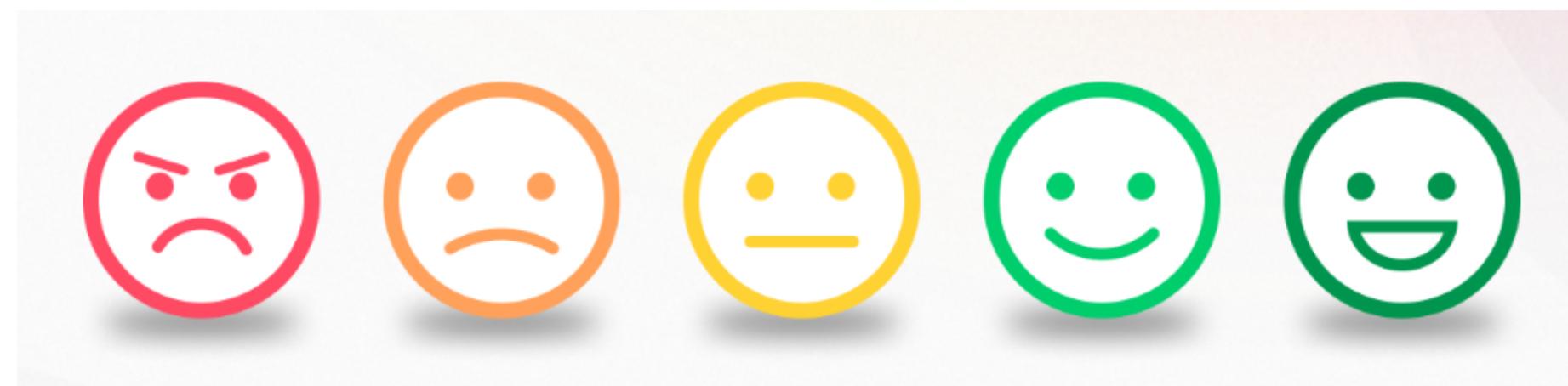
Pr. Taoufik RACHAD

PLAN DE PRÉSENTATION:

- 1-Définition
- 2-Applications
- 3-VADER
- 4-RoBERTa
- 5-Atelier

L'analyse des sentiments

- L'analyse des sentiments désigne le processus d'identification et de catégorisation, de manière informatique, des opinions exprimées dans un texte. Ce procédé vise principalement à déterminer si l'attitude de l'auteur à l'égard d'un sujet, d'un produit, etc., est positive, négative ou neutre.



Applications

- Surveillance des réseaux sociaux
- Suivi de la marque
- Voix du client (VoC)
- Service client
- Analyse des produits
- Recherche et analyse de marché



Quelle est l'utilité du NLP dans l'analyse des sentiments ?

- L'analyse des sentiments (ou *opinion mining*) utilise le NLP pour identifier et extraire les opinions dans un texte.
- Elle combine le NLP et l'apprentissage automatique pour attribuer des scores de sentiments aux entités et sujets d'une phrase.
- Le NLP, branche de l'IA, permet aux ordinateurs de comprendre et manipuler le langage humain.

Les Approches d'Analyse des Sentiment

VADER

Une méthode lexicale simple

RoBERTa

Un modèle préentraîné basé sur
Transformer

VADER : Valence Aware Dictionary and Sentiment Reasoner

C'est un modèle de lexique supervisé conçu pour évaluer la polarité des textes.

neg

proportion de
texte négatif.

neu

proportion de
texte neutre.

pos

proportion de
texte positif

coumpound

un score agrégé
de la polarité
globale.

Facteurs influençant l'analyse de sentiment dans VADER

- **Amplificateurs :**

Les mots comme "very" augmentent la polarité du sentiment.

"I am *very happy*" → Le sentiment devient plus positif.

- **Négation :**

Les mots de négation comme "not" inversent ou réduisent la polarité du sentiment.

"I am *not happy*" → Le sentiment devient négatif.

- **Emphase (Ponctuation et Majuscules) :**

L'utilisation de ponctuation ou de majuscule accentue l'intensité du sentiment.

"I AM SO EXCITED!!!" → Le sentiment est amplifié.

Score compound

$$\text{compound} = \frac{\text{sum of sentiment scores}}{\sqrt{\text{sum of sentiment scores}^2 + \alpha}}$$

- Il synthétise tous les scores individuels en un unique indicateur de polarité.
- Facile à interpréter :

Positif : score compound > 0.

Négatif : score compound < 0.

Neutre : score compound proche de 0 (environ entre -0.05 et +0.05).

Cela fait de **compound** un outil puissant pour résumer le sentiment d'un texte tout en tenant compte de ses nuances.

alpha est généralement choisi comme une petite valeur positive pour empêcher une division par zéro lors de la normalisation et pour contrôler l'effet d'amplification de scores sentimentaux extrêmes, fixée par défaut à 15

Exemple : Calcul du compound pour "I am so happy!"

1/ Analyse des mots :

I : neutre (score = 0).

am : neutre (score = 0).

so : amplificateur.

happy : positif (score lexique = +3.2).

2/ Application des ajustements :

so : augmente l'intensité de happy, amplifiant son score de polarité.

3/ Agrégation et normalisation :

$$\text{compound} = \frac{4.5}{\sqrt{4.5^2 + \alpha}}$$

Ce qui donne un score proche de **0.6468**.

Pourquoi RoBERTa pour l'Analyse des Sentiments ?

Robustly optimized BERT approach

RoBERTa :

- Pré-entraîné sur une grande quantité de données textuelles.
- Excellente performance pour les tâches de classification de texte.

Avantages :

- Précision élevée.
- Capable de gérer un vocabulaire complexe et des nuances linguistiques.

Pipeline de Traitement

Étape 1 : Prétraitement des textes

- Conversion des critiques en vecteurs numériques grâce au **tokenizer** de RoBERTa
"RobertaTokenizer"

Étape 2 : Analyse par RoBERTa

- Le modèle génère des scores pour chaque sentiment (positif, neutre, négatif).

Étape 3 : Post-traitement

- RoBERTa génère des scores pour chaque catégorie de sentiment (positif, neutre, négatif), et un calcul **softmax** est effectué pour normaliser ces scores et obtenir des **probabilités**.
- En fonction des scores obtenus pour chaque catégorie, RoBERTa détermine si le texte est globalement positif, négatif ou neutre.

Limites et Améliorations

- **Limites :**

- Modèle RoBERTa généraliste, non spécialisé pour les critiques de produits.
- Performances sensibles à la qualité des données d'entrée.

- **Améliorations :**

- Ajustement du modèle via fine-tuning avec des données spécifiques.
- Nettoyage et prétraitement approfondis des textes (ex. suppression des fautes).

**MERCI POUR
VOTRE
ATTENTION**