NLP & TEXT PROCESSING

STOP AU CYBER-HARCELEMENT

ASPECT-TARGET SENTIMENT CLASSIFICATION FOR CYBERBULLYING DETECTION

NASSAR Ibrahim
TRANG Thomas

Sommaire

Détection du cyber-harcèlement

Quel est le contexte ?

Que peut-on qualifier de cyber-harcèlement ?

Pour que ce soit considéré comme tel, le message ne doît pas être uniquement négatif, il doit être visé

CERTAINES FORMES DE CYBER-HARCÈLEMENT

- **Insultes, menaces moqueries**

  Méthodes d'intimidation visées sur les réseaux sociaux ou forums

- **Rumeurs**

  Divulgations de données personnelles

- **Usurpation d'identité**

  Piratage de comptes et prise de contrôle de l'identité digitale d'un individu

20%

DES 8-18 ANS DISENT AVOIR ÉTÉ CONFRONTÉS À UNE SITUATION DE CYBER-HARCÈLEMENT

46%

PART DES INTERNAUTES AYANT ÉTÉ TÉMOINS DE PROPOS DIFFAMATOIRES SUR LES RÉSEAUX SOCIAUX

Qu'existe-il déjà dans la lutte contre le cyber-harcèlement ?

FONCTIONNALITÉS

Signaler les posts
Bloquer les utilisateurs

CONTACTER ET DÉNONCER

Numéros gratuits et addreses mail à disposition

EN IA

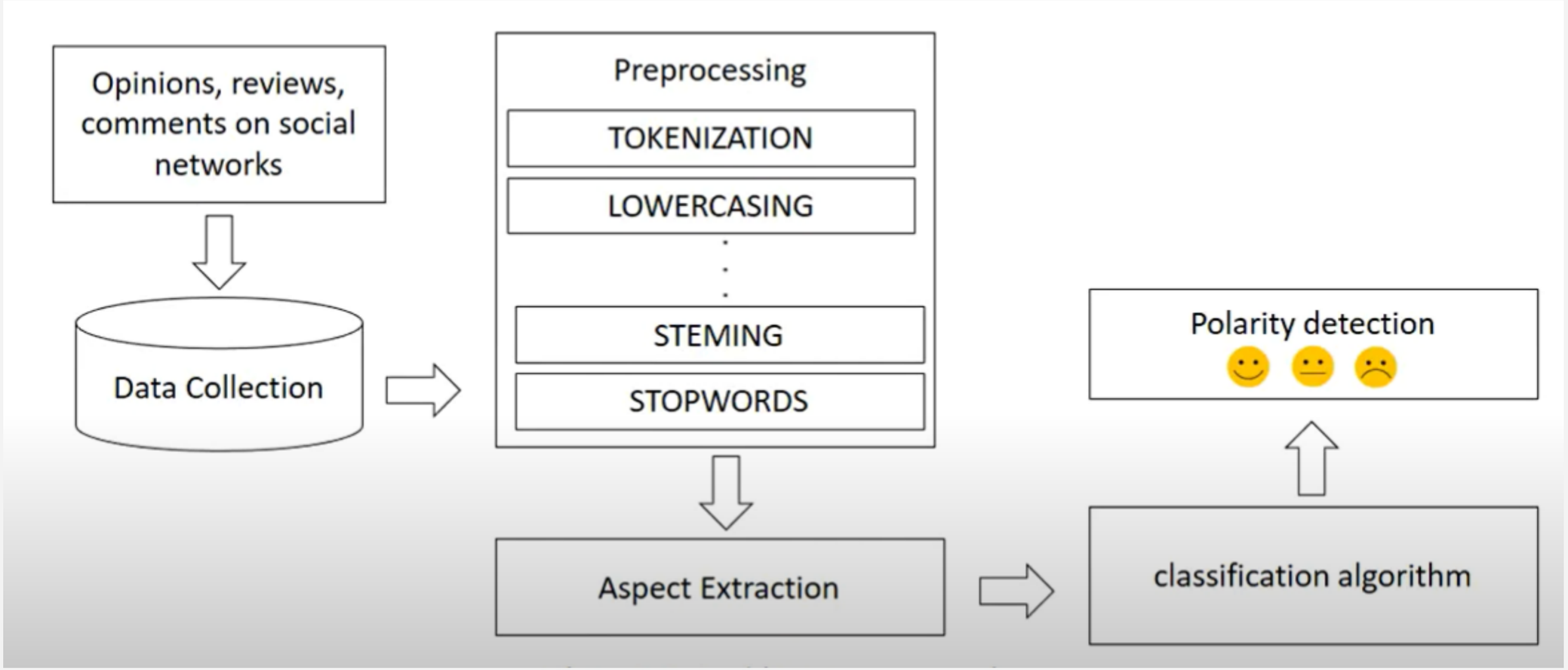
SENTIMENT ANALYSIS

En NLP, une technique pour déterminer la polarité d'un message

MODÈLES PRÉ-ENTRAINÉS

Aspect Ciblé de la classification de sentiments

Quelle est la méthode utilisée ?



Named Entity Recognition (NER)

Qu'est ce que c'est ?

- Étape importante dans l'extraction d'informations
- Détection et classification grammaticale des mots dans une phrase
- Catégorisable selon le sens du mot, par exemple : prénom, entreprise, ville, date, taille ...

Quelle est la méthode utilisée ?

Le modèle BERT-ADA de classification pré-entraîné

- **BERT** : Bidirectional Encoder Representation From Transformers, modèle de langage crée par Google en 2018
- Spécialisé dans la classification de texte
- Similaire aux RNN : Extraction des caractéristiques récurrentes séquentielles

- **BERT-ADA** : Modèle adapté à l'aspect target sentiment classification pré-entraîné
- Fine tuning nécessaire pour calibrer le modèle

PERFORMANCE DE BERT-ADA FACE AUX AUTRES MODÈLES

Table 3: Quantitative analysis of BERT-ADA and baselines

Method	NER	Accuracy	Precision	Recall	F1 Score
BERT-base	–	0.608	0.888	0.432	0.582
DistilBERT	–	0.600	0.892	0.414	0.565
twitter-roBERTa-base	–	0.825	0.944	0.768	0.847
BERTweet	–	0.784	0.848	0.763	0.817
BERT-ADA	No	0.662	0.627	0.882	0.733
BERT-ADA (fine-tuned)	No	0.920	0.868	0.882	0.940
BERT-ADA	Yes	0.694	0.695	0.917	0.791
BERT-ADA (fine-tuned)	Yes	**0.952**	**0.936**	**0.944**	**0.940**

Fine-tuning

HATE SPEECH TWITTER

Tuning du BERT-ADA
sur un corpus Twitter de commentaires haineux labélisés

MEILLEURE PERFORMANCE

Adaptation du modèle au corpus du discours de haine

LE SEUL CAPABLE DE RECONNAÎTRE L'IMPLICIT

Exemples :
"Kobe was mad as hell" ,
"savage"

LES PERSPECTIVES

- **Relecture automatique**

  Des modèles qui relisent avant publication afin de limiter le cyber-harcèlement à la source

---

- **Score de bonne conduite**

  Score pour chaque utilisateur
  suivant son comportement