Modération intelligente des discours dans l'écosystème web sénégalais

Contextes, problématique, enjeux et defis

Dr Edouard Ngor SARR

UFR SES - Dépt Eco-Ges

Laboratoire LI3 & Check4Decision

Université Assane SECK de Ziguinchor-UASZ

Le 06/07/2024

A propos

- Véracité des information
  - Fact checking : Lutte contre les fake news
  - *Edouard Ngor Sarr. Contribution aux mécanismes d'automatisation du fact-checking pour un journalisme augmenté. Web. Université de Thiès (Sénégal), 2019. Français. ⟨NNT : ⟩. ⟨tel-02938383v1⟩*
    - Web crawling
    - Web scraping d'articles et de commentaires
    - Algorithmes pour l'automatisation du Fact-Checking
    - Tagging des faits
    - Détection des Fake news
    - …
- **Aujourd'hui**
  - **Problématiques liées au Web Scraping et web crawling**
  - Machine learning sur les question de GESTION & ECONOMIE
  - Analyse des données pour IoT avec Abel Diatta
  - **Problématiques de la qualité des données sur le Web au Sénégal**
    - Amélioration de la Modération

Pourquoi modérer les contenus ?

- La multiplication des **outils de mise en relation**
  - Réseaux sociaux, sites de rencontre, Forum de discussions …
  - Chaines TV en ligne (sur YouTube)
  - *Une **démocratisation** de la prise de parole souvent sous l'anonymat*
- Une rapide **augmentation**
  - Nombre d'utilisateurs
  - Quantité de données produites et stockées
- La montée d'un **tissu social délétère** marqué des entraves fréquentes aux règles sociales et aux conditions d'utilisation des plateformes
- **Conséquences**
  - **Prolifération discours inappropriés**
    - Appels à la violence
    - Insultes
    - Propos racistes et/ou haineux
    - ….
  - **Polluer les discussions**
  - **Rendre ces espaces en ligne très propices aux abus**

Modération intelligente des discours dans l'écosystème web sénégalais

Pourquoi modérer les contenus ?

@abdourahmaneka2517 il y a 1 an

Rey lén [redacted]domeram bii

Traduire en français

18 **Répondre**

13 réponses

@piispandiaye4749 il y a 1 an

Dolén dagg[redacted]nguéni xalaat

5 **Répondre**

Cheikh Ahmed Cissé insUlte et menace tous les Oustaz "sen d@ta y@ye"

116 vues • il y a 9 mois

@francoisndiaye8439 il y a 1 an

Dakkkkaaar lei[redacted]s femme bi nei nako sac bi sah damako wara dieul dieuleul deh ndah yeelo ngako vrai nga 😂 😂 😂

@idyba2605 il y a 1 an

Galsen 🇸🇳 yakkou na taassarr 😭 dagg[redacted] ndèye [redacted] nguène chalumeau[redacted]

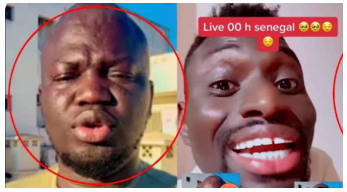
@damembow6374 il y a 1 an

Tallllll lénn [redacted]

Traduire en français

@zalendiaye4789 il y a 1 an

D[redacted]lene ndeyam



@souleyendiaye7660 il y a 1 an

Loléndi khaar pour ciment rapide [redacted]ii 😭 😂

Traduire en français

@daarayseexsaalihu il y a 1 an

Rayleen ko, lu ngeen di xaar

Traduire en français

Comment gérer la situation ?

- **Deux niveaux**
  - **Au niveau Etatique**
    - **Travers des projets de lois**
      - Sanctionnent les responsables des actes
        - Offense au chef de l'Etat
        - Procès sur attente aux mœurs
      - Imputent la responsabilité des propos tenus au responsable des plateformes en ligne.
        - Sanctions lourdes au plateforme
  - **Au niveau des plateformes**
    - *Face à leur incapacité à gérer la situation*
      - Désactive les commentaires
      - Recours
        - ***Recrutement de modérateurs***
        - ***Délégation de l'activités à cabinets de modération***

- **Objectif :** Limitait les abus

C'est quoi la Modération ?

- **Pratique Consistant :**
  - Evaluer et filtrer les contenu diffusés par les internautes sur les plateformes en ligne afin de trouver un juste équilibre entre la protection contre les abus et la liberté d'expression des internautes
  - Mettre en place un mécanisme de gouvernance qui structure la participation à une communauté en ligne pour faciliter la coopération et prévenir les abus
- **Objectifs**
  - Quelles publications et quels utilisateurs sont autorisés à rester en ligne ?
  - Quelles publications et quels utilisateurs sont à supprimer ou à suspendre ?
  - De quelle manière les publications autorisées sont affichées ?
  - Quelles actions accompagnent les suppressions de contenu ?

- **Technique**
  - Phase d'Evaluation : Vérifier si le contenu est en phase avec les regles
  - Phase de Classification : Mettre le commentaire dans
  - Phase de Prise de décision : Choisir l'action adéquate

Modération: Formes

- **Selon l'instant de l'action**
  - Une modération priori
  - Une modération posteriori.

- **Selon l'objectif**
  - Modérer pour veiller au respect des règles de la plateforme
  - Modérer pour veiller pour lutter contre les abus. C'est d'ailleurs la forme la plus courante de la modération des contenus.

- **Selon la mise en œuvre**
  - Modération **manuelle** : par des modérateurs humains (des individus possédant des « droits » de publication, de suppression, d'édition, sur les commentaires postés par les lecteurs)
  - Modération **automatisée** à l'aide de programme informatique
  - Modération **intelligente** à l'aide d'algorithmes d'apprentissage

Modération: Problématique & Défis

- **En quoi et comment les technologies innovantes (IA, ML, TAL, …) peuvent-elles contribuer à une modération plus intelligente des discours dans l'écosystéme ?**
  - Comment identifier un mots ou un propos toxique dans ces situations ?
    - *La toxicité étant définie comme tout ce qui est grossier, irrespectueux ou susceptible de pousser quelqu'un à quitter une discussion, à passer à l'acte de violence envers une personne ou une communauté.*
  - Comment classer automatiquement les commentaires ?

- **Défis sont de faire Face**
  - A la rapidité des flux des discussion en ligne
  - Au volume de données mis en cause
  - A une Insuffisance
    - Documentation et des recherches dans nos langues locales
    - Datasets bien étiqueté de propos ou mots toxiques dans ces langue
  - A la diversité
    - Des langues et des dialectes utilisés
    - Des variantes des mots dans l'écriture et dans le sens
  - Utilisation massive d'abréviations dans les discussions

Modération intelligente : Approches

- **Modération par reconnaissance de schémas**
  - Des algorithmes sont entraînés à reconnaître des schémas et des mots clés spécifiques associés à un comportement problématique et à les classer en différentes catégories simple (valeurs booléennes) ou complexe (plus de deux catégories).
- **Modération par Analyse du sentiment**
  - Des algorithmes sont utilisés pour analyser le sentiment général d'un commentaire permettant d'identifier les commentaires positifs, neutres ou négatifs. Cela peut aider à repérer les commentaires offensants ou provocateurs.
- **Modération contextuelle**
  - Des algorithmes sont utilisées pour comprendre le contexte dans lequel un commentaire est fait par la prise en compte du ton de la discussion, des échanges antérieurs entre les utilisateurs, du contenu de l'article ou de la publication initiale. Cette compréhension du contexte peut contribuer à une modération plus précise et pertinente.
- **Modération prédictive :**
  - Des algorithmes sont entraînés à prédire la probabilité qu'un commentaire soit problématique ou enfreigne les règles, en se basant sur l'historique de l'utilisateur, le contexte ou l'origine. Cela permet de détecter en amont les commentaires potentiellement préjudiciables avant même qu'ils ne soient publiés.
- **Modération Mixte :** Plus fréquent dans la litterature

Modération intelligente : Etat de l'art

- Etat de l'art
  - Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, *12*(1), 129.
  - Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2023, July). SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)* (pp. 868-895). IEEE.
- Détection
  - Jhaver, S., Chen, Q. Z., Knauss, D., & Zhang, A. X. (2022, April). Designing word filter tools for creator-led comment moderation. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1-21).
  - Pradel, F., Zilinsky, J., Kosmidis, S., & Theocharis, Y. (2024). Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, 1-18.
- Modération
  - Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, *73*(10), 1365-1386.
  - Basé sur OPEN AI
    - Li, L., Fan, L., Atreja, S., & Hemphill, L. (2024). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, *18*(2), 1-36.
  - Defis
  - Young, G. K. (2022). How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law*, *31*(1), 1-16.

Modération intelligente : Etat de l'art

- **Travaux inspirants**
  - John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. **Deep Learning** for User Comment Moderation. In Proceedings of the First Workshop on Abusive Language Online, pages 25–35, Vancouver, BC, Canada. Association for Computational Linguistics.
  - Horta Ribeiro, M. (2024). *Content Moderation in Online Platforms* (No. 10387). EPFL.
  - Sam'an, M., & Imaddudin, K. (2024). **Hybrid deep learning model** for YouTube spam comment detection. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(3), 3313-3319.
  - *Damir Korencic, Ipek Baris, Eugenia Fernandez, Katarina Leuschel, and Eva Sánchez Salido. 2021. To Block or not to Block: Experiments* ***with Machine Learning*** *for News Comment Moderation. In Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 127–133, Online. Association for Computational Linguistics.*
  - *Mémoire de master 2022, DIALLO Abdoul Karim, Université 8 Mai 1945 – Guelma, Détection automatique des contenus offensifs* ***en Bambara*** *sur les réseaux sociaux https://dspace.univ-guelma.dz/jspui/bitstream/123456789/14808/1/DIALLO_ABDOUL%20KARIM_F1.pdf*

Nos travaux en cours

1. Propos Toxiques
   - Mise en place et étiquetage manuelle de mots et propos toxiques
   - +20 Etudiants UASZ, UCAO et UNCHK
     - Repartis en plusieurs Groupes : Insultes, Racistes, Violences, Abréviations
     - 10 000 expressions et mots toxiques
2. Identification automatique de propos toxiques en langue locales
   - Apprentissage Supervisé (En cours)
   - Deep Learning
3. Acquisition des données de Test
   - Crawling Scraping large spectre de commentaires par AI (En cours-UNCHK)

4. Modération des contenus avec Opinion Mining
   - L. Faty *et al*., "SenOpinion: A New Lexicon for Opinion Tagging in Senegalese News Comments," *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, Seville, Spain, 2020, pp. 1-6, doi: 10.23919/CISTI49556.2020.9140887.
   - L. Faty, M. Ndiaye, E. N. Sarr and O. Sall, "OpinionScraper: A News Comments Extraction Tool for Opinion Mining," *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Paris, France, 2020, pp. 1-5, doi: 10.1109/SNAMS52053.2020.9336576.

Modération intelligente des discours dans l'écosystème web sénégalais

Contextes, problématique, enjeux et defis

Dr Edouard Ngor SARR

UFR SES

Laboratoire LI3

Université Assane SECK de Ziguinchor-UASZ

Le 06/07/2024