# DÉTECTION DE MESSAGES ABUSIFS AU MOYEN DE RÉSEAUX CONVERSATIONNELS

8ème Conférence MARAMI

Modèles et analyse des réseaux : Approches mathématiques et informatiques 19 octobre 2017 – La Bochelle

Étienne Papégnies<sup>1,2</sup>, Richard Dufour<sup>1</sup>, **Vincent Labatut**<sup>1</sup> & Georges Linarès<sup>1</sup>





prénom.nom@univ-avignon.fr

1 : LIA EA 4128 – Université d'Avignon et des Pays de Vaucluse

2 : Nectar de Code, Barbentane

Détection de messages abusifs au moyen de réseaux conversationnels

ABUSIFS AU MOYEN DE RÉSEAUX CONVERSATIONNELS

Modèles et analyse des réseaux : Approches mu 19 octobre 2017 – La Rochelle

Vincent Labatut <sup>1</sup> & Georges Linarès

1 : LIA EA 4129 – Université d'Avignon et des Pays de Vaucius 2 : Nectar de Code, Barbentane



- 1. Contexte
- 2. Approches existantes
- 3. Méthode proposée
- 4. Résultats
- 5. Conclusions & perspectives

conversationnels
Plan
Overview -Overview

OVERVIEW

### COMMUNAUTÉS EN LIGNE & MODÉRATION

- Communautés en ligne
  - Médium important : niveau d'utilisation, impact sociétal, enjeu économique
  - Utilisateurs généralement anonymes



COMMUNAUTÉS EN LIGNE & MODÉRATION

Bla bla blab la

### COMMUNAUTÉS EN LIGNE & MODÉRATION

- O Communautés en ligne
  - Médium important : niveau d'utilisation, impact sociétal, enjeu économique
  - Utilisateurs généralement anonymes
- Comportement abusif
  - Violation des règles d'utilisation
  - o Conséquences : baisse de qualité, poursuites légales

Détection de messages abusifs au moyen de réseaux conversationnels

—Contexte

Médium important : niveau d'utilisation, impact sociétal, enjeu doncerrique
 Utilisatures généralement anonymes
Comportement abusil
 Violation des règles d'utilisation

Violation des regles d'utilisation
 Conséquences : baisse de qualité, poursuites légals

COMMUNAUTÉS EN LIGNE & MODÉRATION

Communautés en ligne & modération

Bla bla blab la

201

### COMMUNAUTÉS EN LIGNE & MODÉRATION

- Communautés en ligne
  - Médium important : niveau d'utilisation, impact sociétal, enjeu économique
  - Utilisateurs généralement anonymes
- Comportement abusif
  - Violation des règles d'utilisation
  - Conséquences : baisse de qualité, poursuites légales
- Modération
  - Détection de cas d'abus et application de sanctions aux utilisateurs concernés
  - Tâche généralement manuelle : coût élevé

Détection de messages abusifs au moyen de réseaux 201

-Communautés en ligne & modération

Conséquences : baisse de qualité, poursuites légales Détection de cas d'abus et application de sanctions au Tâche pénéralement manuelle : coût éleve

Utilisateurs généralement anonymes Violation des règles d'utilisation

COMMUNAUTÉS EN LIGNE & MODÉRATION

Bla bla blab la

### **AUTOMATISATION DE LA MODÉRATION**

#### Automatisation

- Simple assistance : attirer l'attention des modérateurs sur des cas d'abus potentiels
- Modération complète : détection des abus et application des sanctions
- Problème non-trivial (ex. Google Perspective API)

cas d'abus potentiels Problème non-trivial (ex. Google Perspective API)

AUTOMATISATION DE LA MODÉRATION

Simple assistance : attirer l'attention des modérateurs sur de

-Automatisation de la modération

201

### **AUTOMATISATION DE LA MODÉRATION**

#### Automatisation

- Simple assistance : attirer l'attention des modérateurs sur des cas d'abus potentiels
- Modération complète : détection des abus et application des sanctions
- Problème non-trivial (ex. Google Perspective API)
- Travail présenté
  - Détection de cas d'abus = problème de classification binaire
  - Modélisation des conversations par des graphes
  - Application aux données issues du MMORPG SpaceOrigin

Détection de messages abusifs au moyen de réseaux

20.

Automatisation de la modération

AUTOMATISATION DE LA MODÉRATION

- Simple assistance : attirer l'attention des modérateurs sur de cas d'abus potentiels

- Application aux données issues du MMORPG SpaceOr

### DÉTECTION D'ABUS

- Approches utilisant le contenu [Spe97, CZZX12, DRL11, CS15]
  - Dictionnaires de mots abusifs
  - Règles prédéfinies
  - Approches à bases de *n*-grammes de mots
  - Modèles sac-de-mots (tf-idf)

201 -Détection d'abus

Approches à bases de n-grammes de mots

Dictionnaires de mota abusifa

DÉTECTION D'ABUS

### DÉTECTION D'ABUS

- Approches utilisant le contenu [Spe97, CZZX12, DRL11, CS15]
  - Dictionnaires de mots abusifs
  - Règles prédéfinies
  - Approches à bases de *n*-grammes de mots
  - Modèles sac-de-mots (tf-idf)
- Approches contextuelles [YXH+09, CDNML15, BS15, GDFMGM16]
  - Contenu des messages voisins
  - Modèles d'utilisateurs (langage, comportement)
  - Interactions hors-discussion (ex. abonnements)

Détection de messages abusifs au moyen de réseaux conversationnels

Approphes existantes

-Détection d'abus

DÉTECTION D'ABUS

Dictionnaires de mots abusils
 Régles prédéfinies
 Approches à bases de n-grammes de mots

Modèles sac-de-mots (f'-idf')
 Approches contextuelles (1001)
 Approches contextuelles (1001)

Approches contextuelles [1001\*03, CDNML15, BS15, GDI Contenu des messages volsins Modèles d'utilisateurs (langage, comportement)

Modèles d'utilisateurs (langage, comportement Interactions hors-discussion (ex. abonnements)

201

#### **DÉTECTION D'ABUS**

- Approches utilisant le contenu [Spe97, CZZX12, DRL11, CS15]
  - Dictionnaires de mots abusifs
  - Règles prédéfinies
  - Approches à bases de *n*-grammes de mots
  - Modèles sac-de-mots (*tf-idf*)
- Approches contextuelles [YXH+09, CDNML15, BS15, GDFMGM16]
  - Contenu des messages voisins
  - Modèles d'utilisateurs (langage, comportement)
  - Interactions hors-discussion (ex. abonnements)
- SLSP'17 [PLDL17]
  - Prétraitement spécifique (ex. codes hexadéc ou binaires)
  - Features morphologiques : décomptes de caractères, taux de compression...
  - Features linguistiques : tf-idf, décomptes de mots, d'entités nommées, scores de sentiment...
  - Features comportementales : amplitude de la réponse, modèle de language utilisateur...

Détection de messages abusifs au moyen de réseaux Approches existantes

-Détection d'abus

DÉTECTION D'ABUS

Dictionnaires de mots abusifa

- Approphes à bases de n-grammes de mot Modèles sac-de-mots (tf-idf)
- Approches contextuelles (170110
- Contenu des messages voisins
- Modèles d'utilisateurs (langage, comportement Interactions hors-discussion (ex. abonnements
- Features morphologiques : décomptes de caractères, taux de
- Features linguistiques : If-irlf décommes de mots d'entité
- Features comportementales : amplitude de la réponse, modèl

201

O Données : logs de chats



- O Données : logs de chats
- Objectifs
  - Visualisation des interactions
  - Identification de classes d'utilisateurs, de rôles

201

-Extraction de réseaux conversationnels

- Données : logs de chats
- Objectifs
  - Visualisation des interactions
  - Identification de classes d'utilisateurs, de rôles
- O Difficulté : identité du/des destinataires des messages
  - Mention explicite (plus ou moins flexible) [Mut04, TMZ14, SR14]
  - Connexion à tous les destinataires potentiels [TMZ14]
  - Règles prédéfinies pour identifier les destinataires [Mut04, TMZ14]
  - Proximité et densité temporelle des messages [Mut04]
  - Analyse du contenu (ex. thèmes dans conversations entrelacées) [TMR10]

Détection de messages abusifs au moyen de réseaux conversationnels

Approches existantes

20.

Extraction de réseaux conversationnels

EXTRACTION DE RÉSEAUX CONVERSATIONNELS

Données : logs de chats
 Objectifs

- Visualisation des interactions
   Identification de classes d'utilisateurs, de rôles
- Difficulté : identité du/des destinataires des messages
- Connection à tous les destinataires potentiels [TM214]

  Rântes nefetificies mur intentifier les réstinataires
- Règles prédéfinies pour identifier les destinataires [Mut04, TMZ14]
- Proximité et densité temporelle des messages [Mut
   Analyse du contenu (ex. thèmes dans conversations

conversationnels

Méthode propo └─Méthode proposée Méthode proposée

1. Extraction du réseau conversationnel

MÉTHODE PROPOSÉE

### MÉTHODE PROPOSÉE

- 1. Extraction du réseau conversationnel
  - Graphe pondéré non-orienté
  - Relatif à un message ciblé
  - Couvre une certaine *période de contexte*
  - Nœuds : utilisateurs intervenant pendant la période
  - Liens : échange de message entre deux utilisateurs
  - Poids : intensité de l'interaction

201 -Méthode proposée

MÉTHODE PROPOSÉE

1. Extraction du réseau conversationne Graphe pondéré non-orienté

Relatif à un message ciblé

Couvre une certaine période de context

Liens : échange de message entre deux utilisateurs

### MÉTHODE PROPOSÉE

- 1. Extraction du réseau conversationnel
  - Graphe pondéré non-orienté
  - Relatif à un message ciblé
  - Couvre une certaine période de contexte
  - Nœuds : utilisateurs intervenant pendant la période
  - Liens : échange de message entre deux utilisateurs
  - Poids : intensité de l'interaction
- 2. Calcul des mesures topologiques

Détection de messages abusifs au moyen de réseaux
conversationnels
Méthode proposée

Méthode proposée

MÉTHODE PROPOSÉE

Extraction du réseau conversationnel
 Graphe pondéré non-orienté

Graphe pondere non-onente
 Relatif à un message ciblé
 Couvre une certaine période de context

Nœuds : utilisateurs intervenant pendant la période
 Liens : échange de message entre deux utilisateurs
 Poids : intersité de l'interaction

Calcul des mesures topologiques

### MÉTHODE PROPOSÉE

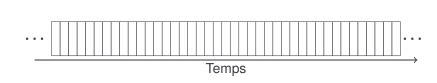
- Extraction du réseau conversationnel
  - Graphe pondéré non-orienté
  - Relatif à un message ciblé
  - Couvre une certaine *période de contexte*
  - Nœuds : utilisateurs intervenant pendant la période
  - Liens : échange de message entre deux utilisateurs
  - Poids : intensité de l'interaction
- 2. Calcul des mesures topologiques
- 3. Entrainement d'un SVM

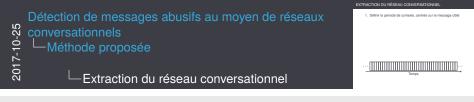
Détection de messages abusifs au moyen de réseaux Méthode proposée 201 -Méthode proposée

MÉTHODE PROPOSÉE

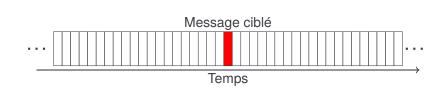
- 1. Extraction du réseau conversationne Graphe pondéré non-orienté
- Relatif à un message ciblé Couvre une certaine période de context
- Neuris utilisateurs intervenant nervient la nériorie Liens : échange de message entre deux utilisateurs
- Poids : intensité de l'interaction 2. Calcul des mesures topologiques
- 3. Entrainement d'un SVM

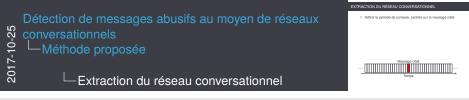
1. Définir la période de contexte, centrée sur le message ciblé



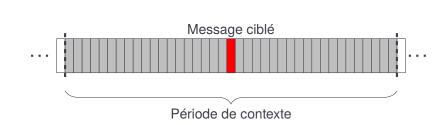


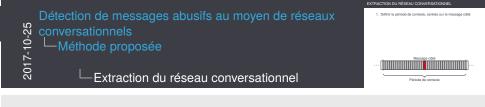
1. Définir la période de contexte, centrée sur le message ciblé



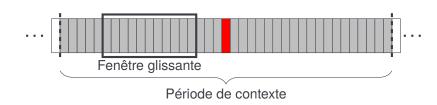


1. Définir la période de contexte, centrée sur le message ciblé





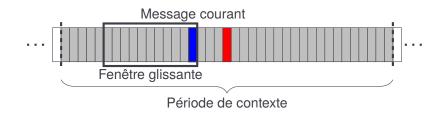
- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant



201 Extraction du réseau conversationnel Période de contexte



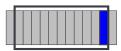
- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant



Détection de messages abusifs au moyen de réseaux 201 Extraction du réseau conversationnel



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un *message* courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

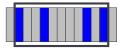
201

Extraction du réseau conversationnel

- 3. Calculer poids des liens concernés
- Hyp.2 : auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un *message* courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- 3. Calculer poids des liens concernés
- Hyp.2 : auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un *message* courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- 3. Calculer poids des liens concernés
  - Hyp.2 : auteurs plus récents sont plus concernés
- Hyp.3 : auteurs mentionnés sont encore plus visés



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- 3. Calculer poids des liens concernés
- Hyp.2 : auteurs plus récents sont plus concernés
- Hyp.3 : auteurs mentionnés sont encore plus visés



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- 3. Calculer poids des liens concernés
- Hyp.2 : auteurs plus récents sont plus concernés
- Hyp.3 : auteurs mentionnés sont encore plus visés



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- 3. Calculer poids des liens concernés
  - Hyp.2 : auteurs plus récents sont plus concernés



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- 3. Calculer poids des liens concernés
  - Hyp.2: auteurs plus récents sont plus concernés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- 3. Calculer poids des liens concernés
- Hyp.2: auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

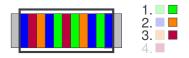
201

Extraction du réseau conversationnel

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés
- Hyp.2: auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés
- Hyp.2: auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés
  - Hyp.2: auteurs plus récents sont plus concernés
- Hyp.3 : auteurs mentionnés sont encore plus visés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés





Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés Hun 1 : message courant adressé aux autres intervenant
  - Hyp.2: auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés Hun 1 : message courant adressé aux autres intervenant
- Hyp.2: auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2 : auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés Hun 1 : message courant adressé aux autres intervenant
  - Hyp.2: auteurs plus récents sont plus concernés
  - Hyp.3 : auteurs mentionnés sont encore plus visés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un *message* courant
- 3. Calculer poids des liens concernés
  - Hyp.1: message courant adressé aux autres intervenants
  - Hyp.2: auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

EXTRACTION DU RÉSEAU CONVERSATIONNEI

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés Hun 1 : message courant adressé aux autres intervenant
  - Hyp.2 : auteurs plus récents sont plus concernés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1 : message courant adressé aux autres intervenants
  - Hyp.2: auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés
- 4. Mettre à jour le graphe



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

EXTRACTION DU RÉSEAU CONVERSATIONNEI

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés Hun 1 : message courant adressé aux autres intervenant Hyp.2 : auteurs plus récents sont plus concernés
- Hyp.3 : auteurs mentionnés sont encore plus visés







- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1 : message courant adressé aux autres intervenants
  - Hyp.2: auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés
- 4. Mettre à jour le graphe



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

EXTRACTION DU RÉSEAU CONVERSATIONNEI

- Définir la nériode de contexte, centrée sur le message nihi
- Hun 1 : message courant adressé aux autres intervenant Hyp.2 : auteurs plus récents sont plus concernés

3. Calculer poids des liens concernés

- Hyp.3 : auteurs mentionnés sont encore plus visés



- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1 : message courant adressé aux autres intervenants
  - Hyp.2: auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés
- 4. Mettre à jour le graphe



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

EXTRACTION DU RÉSEAU CONVERSATIONNEI

- Définir la nériode de contexte, centrée sur le message nihi
- Hun 1 : message courant adressé aux autres intervenant Hyp.2 : auteurs plus récents sont plus concernés

3. Calculer poids des liens concernés

Hyp.3 : auteurs mentionnés sont encore plus visés







- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1 : message courant adressé aux autres intervenants
  - Hyp.2: auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés
- 4. Mettre à jour le graphe



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

EXTRACTION DU RÉSEAU CONVERSATIONNEI

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés Hun 1 : message courant adressé aux autres intervenant
- Hyp.2 : auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés





- 1. Définir la période de contexte, centrée sur le message ciblé
- 2. Parcourir via une fenêtre glissante relative à un message courant
- 3. Calculer poids des liens concernés
  - Hyp.1 : message courant adressé aux autres intervenants
  - Hyp.2: auteurs plus récents sont plus concernés
  - Hyp.3: auteurs mentionnés sont encore plus visés
- 4. Mettre à jour le graphe



Détection de messages abusifs au moyen de réseaux

Méthode proposée

201

Extraction du réseau conversationnel

EXTRACTION DU RÉSEAU CONVERSATIONNEI

- Définir la nériode de contexte, centrée sur le message nihi
- 3. Calculer poids des liens concernés Hun 1 : message courant adressé aux autres intervenant
- Hyp.2 : auteurs plus récents sont plus concernés Hyp.3 : auteurs mentionnés sont encore plus visés





### MESURES TOPOLOGIQUES

- Mesures locales
  - Degré, Eigenvector, PagRank, Hub & Authority
  - Betweeness, Closeness, Eccentricity, Coreness

Détection de messages abusifs au moyen de réseaux conversationnels

Méthode proposée

Mesures topologiques

MESURES TOPOLOGIQUES

Degré, Eigenvector, PagRank, Hub & Authority

#### **MESURES TOPOLOGIQUES**

- Mesures locales
  - Degré, Eigenvector, PagRank, Hub & Authority
  - Betweeness, Closeness, Eccentricity, Coreness
- Mesures globales
  - o Nombres de nœuds et de liens, densité
  - Diamètre, distance moyenne
  - Nombre de cliques
  - Assortativité du degré
  - Moyenne de chaque mesure locale

Détection de messages abusifs au moyen de réseaux 10-25

Méthode proposée

201

-Mesures topologiques

MESURES TOPOLOGIQUES

Mesures locales Degré, Eigenvector, PagRank, Hub & Authority

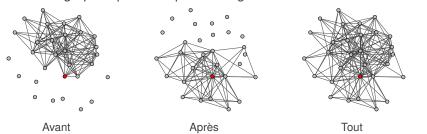
Mesures globales Nombres de nœuds et de liens, densité Diamètre, distance moyenne

Nombre de climas

Moyenne de chaque mesure local

# MESURES TOPOLOGIQUES

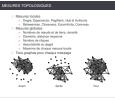
- Mesures locales
  - Degré, Eigenvector, PagRank, Hub & Authority
  - Betweeness, Closeness, Eccentricity, Coreness
- Mesures globales
  - Nombres de nœuds et de liens, densité
  - o Diamètre, distance moyenne
  - Nombre de cliques
  - Assortativité du degré
  - Moyenne de chaque mesure locale
- Trois graphes pour chaque message



Détection de messages abusifs au moyen de réseaux conversationnels

Méthode proposée

Mesures topologiques



# DONNÉES ET PROTOCOLE EXPÉRIMENTAL

#### Données

- Chat du MMORPG SpaceOrigin
- 4 029 343 messages instantanés
  - o 779 signalés et confirmés comme abusifs
  - Échantillon de 2 000 messages supposés non-abusifs
  - Tous les messages pris dans des conversations distinctes





Détection de messages abusifs au moyen de réseaux conversationnels
Résultats

Échantillon de 2 000 messages supposés non-abusil

779 signalés et confirmés comme abusifs.

DONNÉES ET PROTOCOLE EXPÉRIMENTAL

-Données et protocole expérimental

### DONNÉES ET PROTOCOLE EXPÉRIMENTAL

#### Données

- Chat du MMORPG SpaceOrigin
- 4 029 343 messages instantanés
  - 779 signalés et confirmés comme abusifs
  - Échantillon de 2 000 messages supposés non-abusifs
  - o Tous les messages pris dans des conversations distinctes





#### Classification

- SVM (Sklearn C-Support Vector Classification)
- Validation croisée 70–30%
- Estimation du pouvoir discriminant des features par ExtraTreesClassifier (Sklearn)

Détection de messages abusifs au moyen de réseaux conversationnels

 $^{igspace}$ Données et protocole expérimental

Données
 Obarde MACROS (Superiority):
 400 345 missappe instantanés
 400 345 missappe instantanés
 50 signales et confirmires comme abustis
 50 des missappe produces acceptes en orabustis
 50 bas les messappes privines des conversation diefocie

DONNÉES ET PROTOCOLE EXPÉRIMENTAL

afon

SVM (Sklearn C-Support Vector Classification)
 Validation croisée 70–30%
 Estimation du pouvoir discriminant des features par

20.

# RÉSULTATS DE LA CLASSIFICATION

#### Performances pour la classe Abus

Classificateur	Précision	Rappel	F-mesure
Référence aléatoire	0,28	0,50	0,36
Features contenu/contexte	0,70	0,74	0,72
Features graphe	0,77	0,77	0,77

Détection de messages abusifs au moyen de réseaux conversationnels

Résultats

RÉSULTATS DE LA CLASSIFICATION

Résultats de la classification

### RÉSULTATS DE LA CLASSIFICATION

#### Performances pour la classe Abus

Classificateur	Précision	Rappel	F-mesure
Référence aléatoire	0,28	0,50	0,36
Features contenu/contexte	0,70	0,74	0,72
Features graphe	0,77	0,77	0,77

- Performances améliorées bien que le contenu soit ignoré
  - Explication possible : meilleure utilisation de l'information post-abus

Détection de messages abusifs au moyen de réseaux
conversationnels
Résultats
Résultats de la classification

Performance pour la risage Alou
Testalidade pour Missian Rept. Personal P

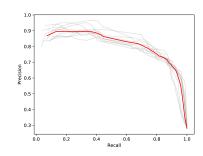
RÉSULTATS DE LA CLASSIFICATION

### RÉSULTATS DE LA CLASSIFICATION

#### Performances pour la classe Abus

Classificateur	Précision	Rappel	F-mesure
Référence aléatoire	0,28	0,50	0,36
Features contenu/contexte	0,70	0,74	0,72
Features graphe	0,77	0,77	0,77

- Performances améliorées bien que le contenu soit ignoré
  - Explication possible : meilleure utilisation de l'information post-abus
- Classificateur apte à assister un modérateur



Détection de messages abusifs au moyen de réseaux conversationnels

Résultats

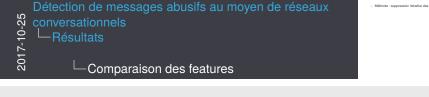
Résultats

Résultats de la classification

Performances position facilities (1997) | Performances | Performan

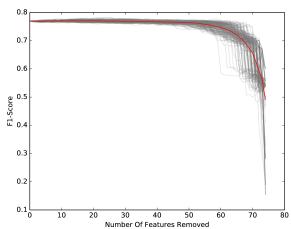
RÉSULTATS DE LA CLASSIFICATION

Méthode : suppression itérative des features



COMPARAISON DES FEATURES

Méthode : suppression itérative des features



Détection de messages abusifs au moyen de réseaux conversationnels

Résultats



COMPARAISON DES FEATURES

-Comparaison des features

12

Méthode : suppression itérative des features
 Features les plus discriminantes pour l'approche à base de graphe

Graphe	Feature	F-mesure avant suppression
Tout	Betweenness moyenne	0,76
Avant	Coreness moyenne	0,75
Après	Nombre de liens	0,75
Après	Densité	0,73
Tout	Score de Hub	0,73
Après	Centralité de degré	0,68
Avant	Nombre de liens	0,67
Tout	Eccentricity moyenne	0,58
Avant	Eigenvector moyenne	0,57
Tout	Eccentricity	0,35

Détection de messages abusifs au moyen de réseaux conversationnels
Résultats

-Comparaison des features

Melhodo - copproposition beature des industres
 Feaburies lais palar de confinencia pour Tapponde à base de gra
 Teatre de la companya del la companya de la companya del la companya de la companya de la companya del la companya de la companya del la companya de

COMPARAISON DES FEATURES

Méthode : suppression itérative des features

Features les plus discriminantes pour l'approche à base de graphe

	Betweenness moyenne Coreness moyenne	0,76
	Coreness movenne	
Avant (	Corchess moyenne	0,75
Après I	Nombre de liens	0,75
Après I	Densité	0,73
Tout 5	Score de Hub	0,73
Après (	Centralité de degré	0,68
Avant I	Nombre de liens	0,67
Tout I	Eccentricity moyenne	0,58
Avant I	Eigenvector moyenne	0,57
Tout I	Eccentricity	0,35

#### Observations

- o Mesures hétérogènes par leur façon de caractériser le graphe
- o Groupes de features très corrélées et quasi-interchangeables
- Considérer avant/après améliore la discrimination
- o Présence de l'excentricité individuelle, mais aussi moyenne

Détection de messages abusifs au moyen de réseaux conversationnels

—Comparaison des features

Midmode - suppression l'étative des teatures
 Schartes de Justice d'active des teatures
 Schartes de Justice d'active de l'extremiser pour de l'extremiser pour suppression
 Total Balescerena moyere
 Aprile Des l'Active de l'extremiser pour suppression
 Aprile Des l'extremiser de l'

COMPARAISON DES FEATURES

20.

Observations
 Mesures hétérogènes par leur façon de caractériser le ;
 Groupes de features très cométées et quisil-interchange

Groupes de features très comélées et quasi-interchi
 Considérer avant'après améliore la discrimination
 Présence de l'expentriché intérchalle, mais aussi m

#### **CONCLUSIONS & PERSPECTIVES**

- Résultats principaux
  - Approche simple
  - Robuste aux problèmes de prétraitement textuel
  - Meilleurs résultats qu'en utilisant le contenu/contexte
  - Performance suffisante pour assistance, pas automatisation complète
  - Limites : coût computationnel, pas de temps réel

Détection de messages abusifs au moyen de réseaux conversationnels

Conclusions & perspective

10-25

201

Conclusions & perspectives

CONCLUSIONS & PERSPECTIVES

Résultats principaux

Approche simple

Robuste aux problèmes de prétraitement textuel
 Meilleurs résultats qu'en utilisant le contenuicontexte

Performance suffisante pour assistance, pas automatisation committe

Limites : coût computationnel, pas de temps

#### **CONCLUSIONS & PERSPECTIVES**

- Résultats principaux
  - Approche simple
  - Robuste aux problèmes de prétraitement textuel
  - Meilleurs résultats qu'en utilisant le contenu/contexte
  - Performance suffisante pour assistance, pas automatisation complète
  - Limites : coût computationnel, pas de temps réel
- Perspectives
  - Explorer l'effet des différents paramètres d'extraction
  - Considérer plus de mesures topologiques (notamment mesoscopiques)
  - Combiner avec les features contenu/contexte
  - Modéliser sous forme de réseaux dynamiques

Détection de messages abusifs au moyen de réseaux

-25

ė

201

-Conclusions & perspectives

CONCLUSIONS & PERSPECTIVES

Réquitats principau

- Robuste aux problèmes de prétraitement textuel Meilleurs résultats qu'en utilisent le contenuirontest
- Performance sufficente nour assistance, nas automotisation

- Considérer plus de mesures topologiques (notamme
- Combiner avec les features contenu/context Modéliser sous forme de réseaux dynamiques

Détection de messages abusifs au moyen de réseaux conversationnels

Conclusions & perspectives

# Questions

[BS15] K. Balci and A. A. Salah. Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. Computers in Human Behavior, 53:517-526, 2015.

J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. [CDNML15] Antisocial behavior in online discussion communities.

arXiv:1504.00680 [cs.SI], 2015.

V. S. Chavan and S. S. Shvlaia. [CS15]

P. Mutton.

Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In IEEE ICACCI, pages 2354-2358, 2015.

[CZZX12] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, Detecting offensive language in social media to protect adolescent online safety. In PASSAT/SocialCom, pages 71-80, 2012.

[DRL11] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. 5th International AAAI Conference on Weblogs and Social Media, pages 11-17, 2011.

[GDFMGM16] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media.

In 9th ACM International Conference on Web Search and Data Mining, pages 33-42, 2016.

[Mut04] Inferring and visualizing social networks on internet relay chat. In 8th International Conference on Information Visualisation, pages 35-43, 2004,

E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. [PLDL17] Impact of content features for automatic online abuse detection. In International Conference on Computational Linguistics and Intelligent Text Processing, 2017.

[Spe97] E. Spertus.

Smokey: Automatic recognition of hostile messages.

In 14th National Conference on Artificial Intelligence and 9th Conference on Innovative Applications of Artificial Intelligence, pages 1058-1065, 1997.

Détection de messages abusifs au moyen de réseaux 0 201

- COMATE J. Own C. Devery Notices Med. and J. Lesboom
  - V. E. Olesson and E. E. Diyloja.

    Machine learning approach for deletionism of opter aggressive assessments by peers, on social media network.

    In IEEE/CACC... seems 2016—2016. 2016.
- [CZD12] Y. Own, Y. Zhou, S. Zho, and H. Xu.
- [DRLT1] K. Dinatar, K. Rebhari, and H. Liderman. Materiae for detector of including bulleting.
- [PLDL17] E. Papageira, V. Labetal, R. Dalton, and G. Linerin. Impact of certain final area for automatic order above detector.



[SR14] T. Sinha and I. Rajasingh. Investigating substructures in goal oriented online communities: Case study of Ubuntu IRC. In IEEE International Advance Computing Conference, pages 916–922, 2014.

[TMR10] S. Trausan-Matu and T. Rebedea.

A polyphonic model and system for inter-animation analysis in chat conversations with multiple participants.

In Computational Linguistics and Intelligent Text Processing, volume 6008 of Lecture Notes in Computer Science, pages 354–363. Springer, 2010.

[TMZ14] S. Tavassoli, M. Moessner, and K. A. Zweig. Constructing social networks from semi-structured chat-log data. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 146–149, 2014.

[YXH<sup>+</sup>09] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on Web 2.0.

In WWW Workshop: Content Analysis in the Web 2.0, 2009.

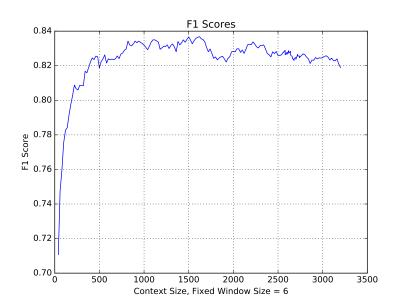
Détection de messages abusifs au moyen de réseaux conversationnels

— Références

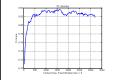
[SEA] Z. Schwarz M. Spinjah.
Sweptige Spinishnesses year desired order securities. Case shalp of Ubers PC.
SWED Commission of International Conference Spinishness spin Still Conference Spinishnesses
A SPI Commission of International Conference Spinishnesses
A SPINISHNESSES SPINISHNESSES SPINISHNESSES
A SPINISHNESSES SPINISHNESSES
A Commission of English Spinishnesses english and consecurities with multiple parties
A Commission of English Spinishnesses and analysis of English Spinishnesses
A Commission of English Spinishnesses
A Commission of English Spinishnesses
A SPINISHNESSES
A Commission of English Spinishnesses
A SPINISHNESSE

[YOU<sup>+</sup>OB] D. Yin, Z. Xun, L. Heng, B. D. Sanham, A. Karrinsiahin, and L. Eshamib. Distribution of humanoseni on Web 2 G. In 1990 O' Histology - Constent Analysis in the Sinh 2 G, 2000.

# EFFET DE LA PÉRIODE DE CONTEXTE







EFFET DE LA PÉRIODE DE CONTEXTE

Effet de la période de contexte