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 Rochelle

Étienne Papégnies[1,2], Richard Dufour[1],
Vincent Labatut[1] & Georges Linarès[1]

prénom.nom@univ-avignon.fr
1 : LIA EA 4128 – Université d'Avignon et des Pays de Vaucluse
2 : Nectar de Code, Barbentane

2017-10-25

Détection de messages abusifs au moyen de réseaux conversationnels

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 Rochelle

Étienne Papégnies[1,2], Richard Dufour[1],
Vincent Labatut[1] & Georges Linarès[1]

prénom.nom@univ-avignon.fr
1 : LIA EA 4128 – Université d'Avignon et des Pays de Vaucluse
2 : Nectar de Code, Barbentane

OVERVIEW

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

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

COMMUNAUTÉS EN LIGNE & MODÉRATION

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

COMMUNAUTÉS EN LIGNE & MODÉRATION
Communautés en ligne
Médium important : niveau d'utilisation, impact sociétal, enjeu économique
Utilisateurs généralement anonymes

Bla bla blab la

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

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

Bla bla blab la

COMMUNAUTÉS EN LIGNE & MODÉRATION

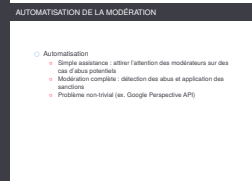
- Communautés en ligne
  - Médium important : niveau d'utilisation, impact sociétal, enjeu économique
  - Utilisateurs généralement anonymes
- Comportement abusif
  - Violation des règles d'utilisation
  - Conséquences : baisse de qualité, poursuites légales
- Modération
  - Détection de cas d'abus et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

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

Bla bla blab la

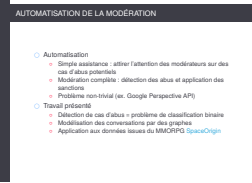
AUTOMATISATION DE LA MODÉRATION

- 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)



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 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*)

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 (F–idf)

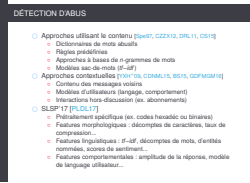
DÉTECTION D'ABUS

- 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 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 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...



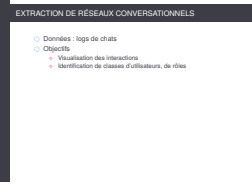
EXTRACTION DE RÉSEAUX CONVERSATIONNELS

- Données : logs de chats

EXTRACTION DE RÉSEAUX CONVERSATIONNELS
Données : logs de chats

EXTRACTION DE RÉSEAUX CONVERSATIONNELS

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



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
  - 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]

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
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]

MÉTHODE PROPOSÉE

1. Extraction du réseau conversationnel

MÉTHODE PROPOSÉE
1. Extraction du réseau conversationnel

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

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

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

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

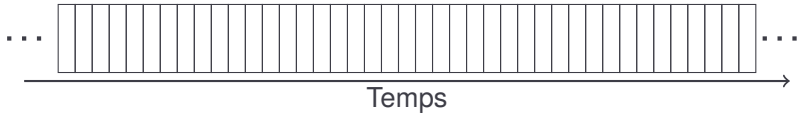
MÉTHODE PROPOSÉE

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
3. Entrainement d'un SVM

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
3. Entrainement d'un SVM

EXTRACTION DU RÉSEAU CONVERSATIONNEL

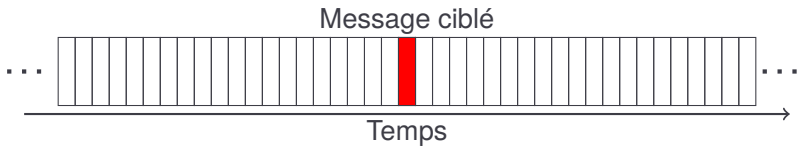
1. Définir la période de contexte, centrée sur le message ciblé



EXTRACTION DU RÉSEAU CONVERSATIONNEL
1. Définir la période de contexte, centrée sur le message ciblé
Temps

EXTRACTION DU RÉSEAU CONVERSATIONNEL

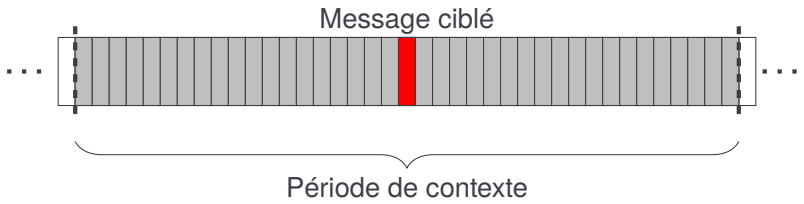
1. Définir la période de contexte, centrée sur le message ciblé



EXTRACTION DU RÉSEAU CONVERSATIONNEL
1. Définir la période de contexte, centrée sur le message ciblé
Message ciblé
Temps

EXTRACTION DU RÉSEAU CONVERSATIONNEL

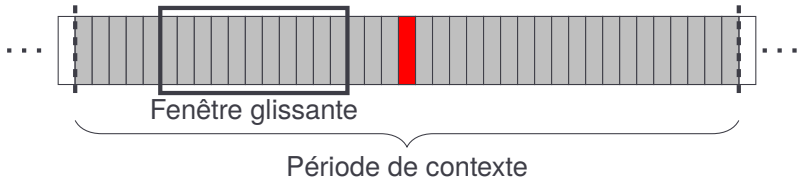
1. Définir la période de contexte, centrée sur le message ciblé



EXTRACTION DU RÉSEAU CONVERSATIONNEL
1. Définir la période de contexte, centrée sur le message ciblé
Message ciblé
Période de contexte

EXTRACTION DU RÉSEAU CONVERSATIONNEL

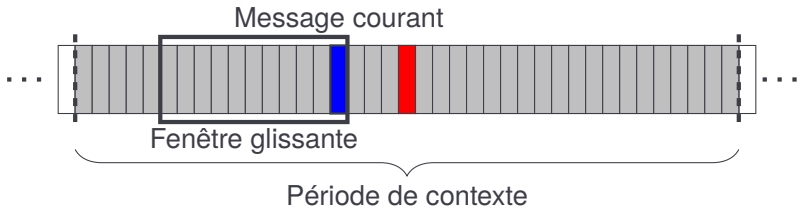
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*



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*
Fenêtre glissante
Période de contexte

EXTRACTION DU RÉSEAU CONVERSATIONNEL

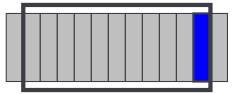
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*



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*
Message courant
Fenêtre glissante
Période de contexte

EXTRACTION DU RÉSEAU CONVERSATIONNEL

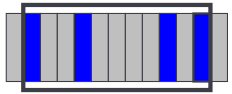
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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



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

EXTRACTION DU RÉSEAU CONVERSATIONNEL

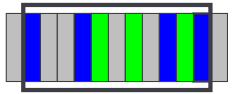
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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



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

EXTRACTION DU RÉSEAU CONVERSATIONNEL

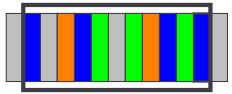
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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



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

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



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

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

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

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

1.

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

1.

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

1. ■
2. ■

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

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

1.
2.
3.

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

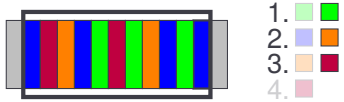
EXTRACTION DU RÉSEAU CONVERSATIONNEL

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

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

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

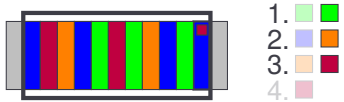


1.
2.
3.
4.

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

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

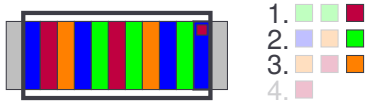


1.
2.
3.
4.

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

EXTRACTION DU RÉSEAU CONVERSATIONNEL

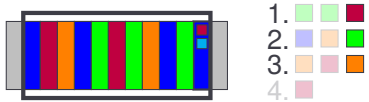
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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



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

EXTRACTION DU RÉSEAU CONVERSATIONNEL

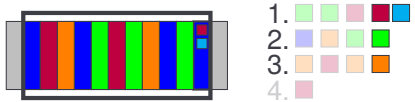
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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



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

EXTRACTION DU RÉSEAU CONVERSATIONNEL

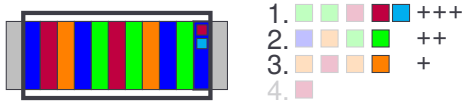
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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



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

EXTRACTION DU RÉSEAU CONVERSATIONNEL

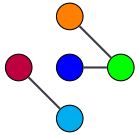
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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

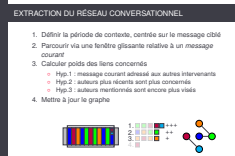


EXTRACTION DU RÉSEAU CONVERSATIONNEL
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents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

EXTRACTION DU RÉSEAU CONVERSATIONNEL

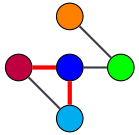
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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





EXTRACTION DU RÉSEAU CONVERSATIONNEL

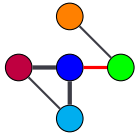
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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



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
4. Mettre à jour le graphe

EXTRACTION DU RÉSEAU CONVERSATIONNEL

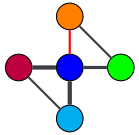
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents sont plus concernés
   - Hyp.3 :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



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
4. Mettre à jour le graphe

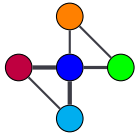
EXTRACTION DU RÉSEAU CONVERSATIONNEL

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
4. Mettre à jour le graphe



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
4. Mettre à jour le graphe



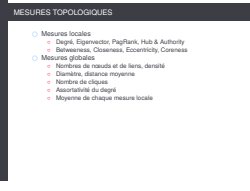
EXTRACTION DU RÉSEAU CONVERSATIONNEL
1. Définir la période de contexte, centrée sur le message ciblé
2. Parcourir via une fenêtre glissante relative à un *message courant*
3. Calculer poids des liens concernés
   - Hyp.1 : message courant adressé aux autres intervenants
   - Hyp.2 : auteurs plus récents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

MESURES TOPOLOGIQUES

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

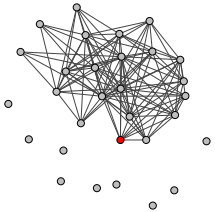
MESURES TOPOLOGIQUES

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

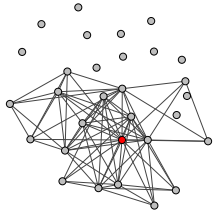


MESURES TOPOLOGIQUES

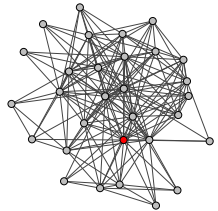
- Mesures locales
  - Degré, Eigenvector, PagRank, Hub & Authority
  - Betweeness, Closeness, Eccentricity, Coreness
- Mesures globales
  - Nombres de nœuds et de liens, densité
  - Diamètre, distance moyenne
  - Nombre de cliques
  - Assortativité du degré
  - Moyenne de chaque mesure locale
- Trois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



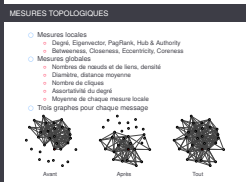
Avant



Après

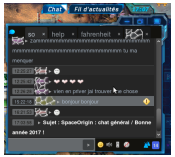


Tout



DONNÉES ET PROTOCOLE EXPÉRIMENTAL

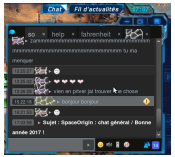
- Données
  - Chat du MMORPG SpaceOrigin
  - 4 029 343 messages instantanés
    - 779 signalés et confirmés comme abusifs
    - Échantillon de 2 000 messages supposés non-abusifs
    - Tous les messages pris dans des conversations distinctes



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
Tous les messages pris dans des conversations distinctes

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
    - 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)

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

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

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

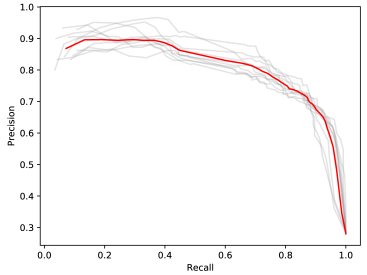
RÉSULTATS DE LA CLASSIFICATION
Performances pour la classe Abus
Classificateur | Précision | Rappel | F-mesure
Référence aléatoire | 0,28 | 0,50 | 0,36
Features contenu/contexte | 0,70 | 0,74 | 0,72
Features graphe | 0,77 | 0,77 | 0,77
Performances améliorées bien que le contenu soit ignoré
Explication possible : meilleure utilisation de l'information post-abus

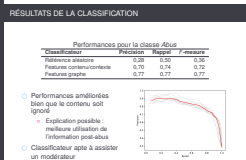
RÉSULTATS DE LA CLASSIFICATION

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





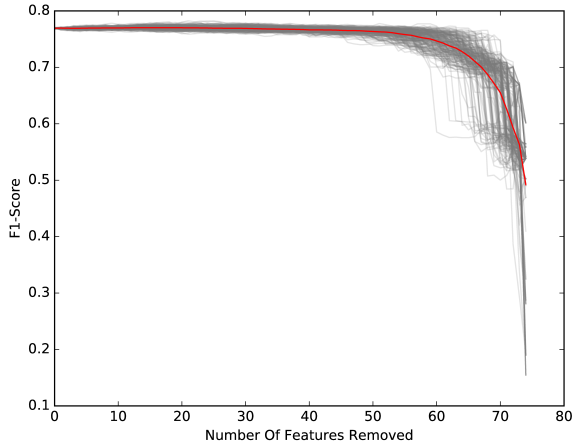
COMPARAISON DES FEATURES

- Méthode : suppression itérative des features

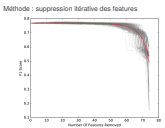
COMPARAISON DES FEATURES
Méthode : suppression itérative des features

COMPARAISON DES FEATURES

- Méthode : suppression itérative des features



COMPARAISON DES FEATURES
Méthode : suppression itérative des features



COMPARAISON DES FEATURES

- 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

COMPARAISON DES FEATURES
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

COMPARAISON DES FEATURES

- 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

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

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

CONCLUSIONS & PERSPECTIVES

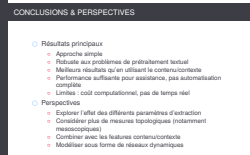
- Résultats principaux
  - Approche simple
  - Robuste aux problèmes de prétraitement textuel
  - Meilleurs résultats qu'en utilisant le contenu/contexte
  - Performance suffisante pour assistance, pas automatisation complète
  - Limites : coût computationnel, pas de temps réel

└Conclusions & perspectives

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

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



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.

[CDNML15] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec.
Antisocial behavior in online discussion communities.
arXiv :1504.00680 [cs.SI], 2015.

[CS15] V. S. Chavan and S. S. Shylaja.
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] P. Mutton.
Inferring and visualizing social networks on internet relay chat.
In *8th International Conference on Information Visualisation*, pages 35–43, 2004.

[PLDL17] E. Papegnies, V. Labatut, R. Dufour, and G. Linarès.
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.

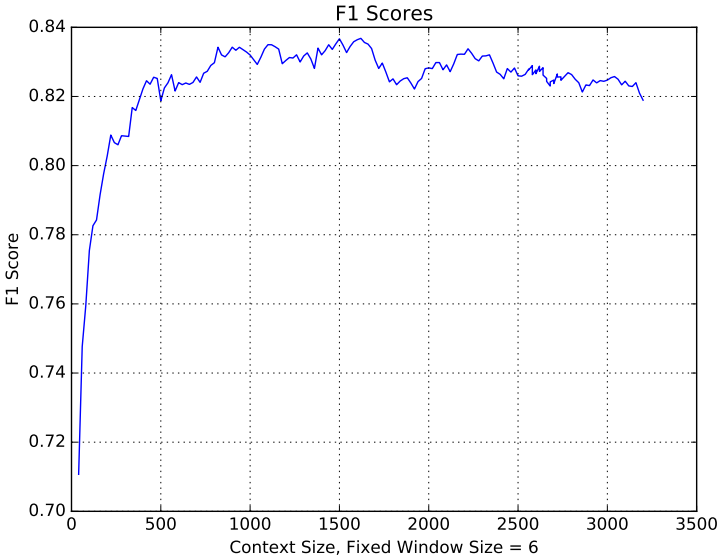
[SR14] T. Sinha and I. Rajasingh.
Investigating substructures in goal oriented online communities : Case study of Ubuntu IRC.
In *IEEE International Advance Computing Conference*, pages 916–922, 2014.

[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+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.

EFFET DE LA PÉRIODE DE CONTEXTE



EFFET DE LA PÉRIODE DE CONTEXTE