



HAL
open science

Abusive Language Detection in Online Conversations by Combining Content- and Graph-based Features

Noé Cecillon, Vincent Labatut, Richard Dufour

► To cite this version:

Noé Cecillon, Vincent Labatut, Richard Dufour. Abusive Language Detection in Online Conversations by Combining Content- and Graph-based Features. Meetup LIAvignon, Nov 2022, Avignon, France. 2022. ⟨hal-03857928⟩

HAL Id: hal-03857928

<https://hal.science/hal-03857928v1>

Submitted on 17 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Abusive Language Detection in Online conversations by Combining Content- and Graph-based Features

Noé Cécillon¹, Vincent Labatut¹, Richard Dufour²

1: LIA – Laboratoire Informatique d'Avignon / 2: LS2N – Laboratoire des Sciences du Numérique de Nantes

Objective

- Detect abusive messages in online conversations through content- and graph-based methods.

Approach

1. Exhibit the importance of structural information in the context of online abuse detection.
2. Combine textual and structural information by proposing fusion methods of both sources of information.

Context

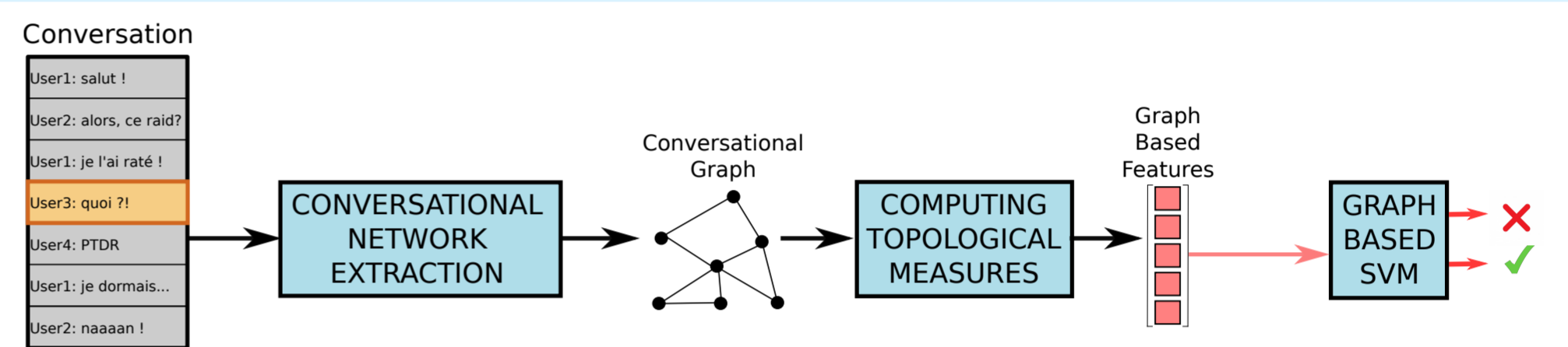
- 62% of french people experienced cyber-harassment at least once.
- Companies are responsible for content posted on their platforms.
- Billion daily users and messages sent on social networks.
- Moderation implies high human and financial costs.

Our proposition to automate the detection of abusive content online :

- A method based on a mix of standard NLP features.
- A graph-based approach that relies on interaction and completely ignores the textual content of messages.
- A combination of both approaches.

Graph-based method

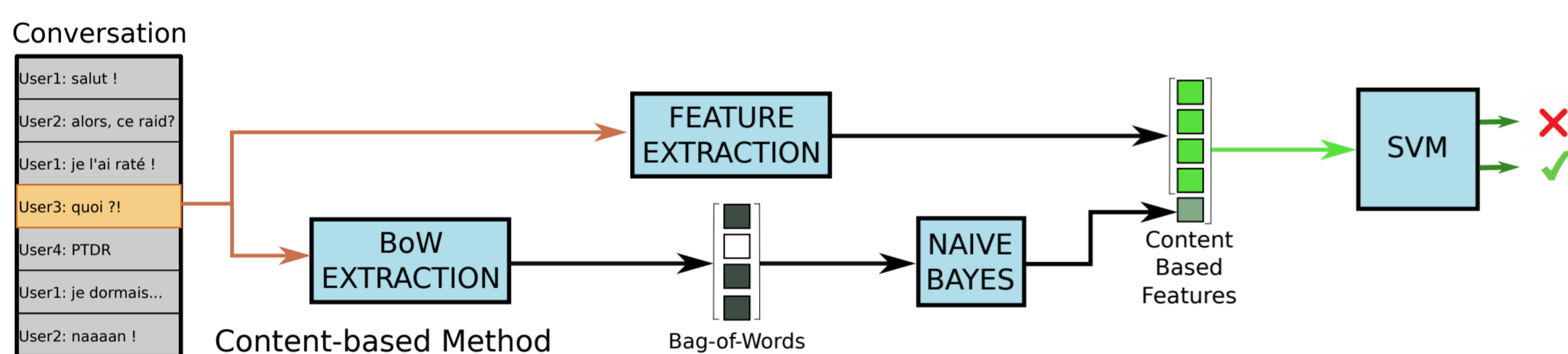
We extract a number of topological measures to describe the conversational graph. A SVM classifier is then trained using these values as features.



Graph-based pipeline. topological measures include density, edge count, degree centrality, eccentricity..

Content-based method

We extract NLP features from the content of each considered message to train a SVM classifier to distinguish between *Abuse* and *Non-abuse* class.



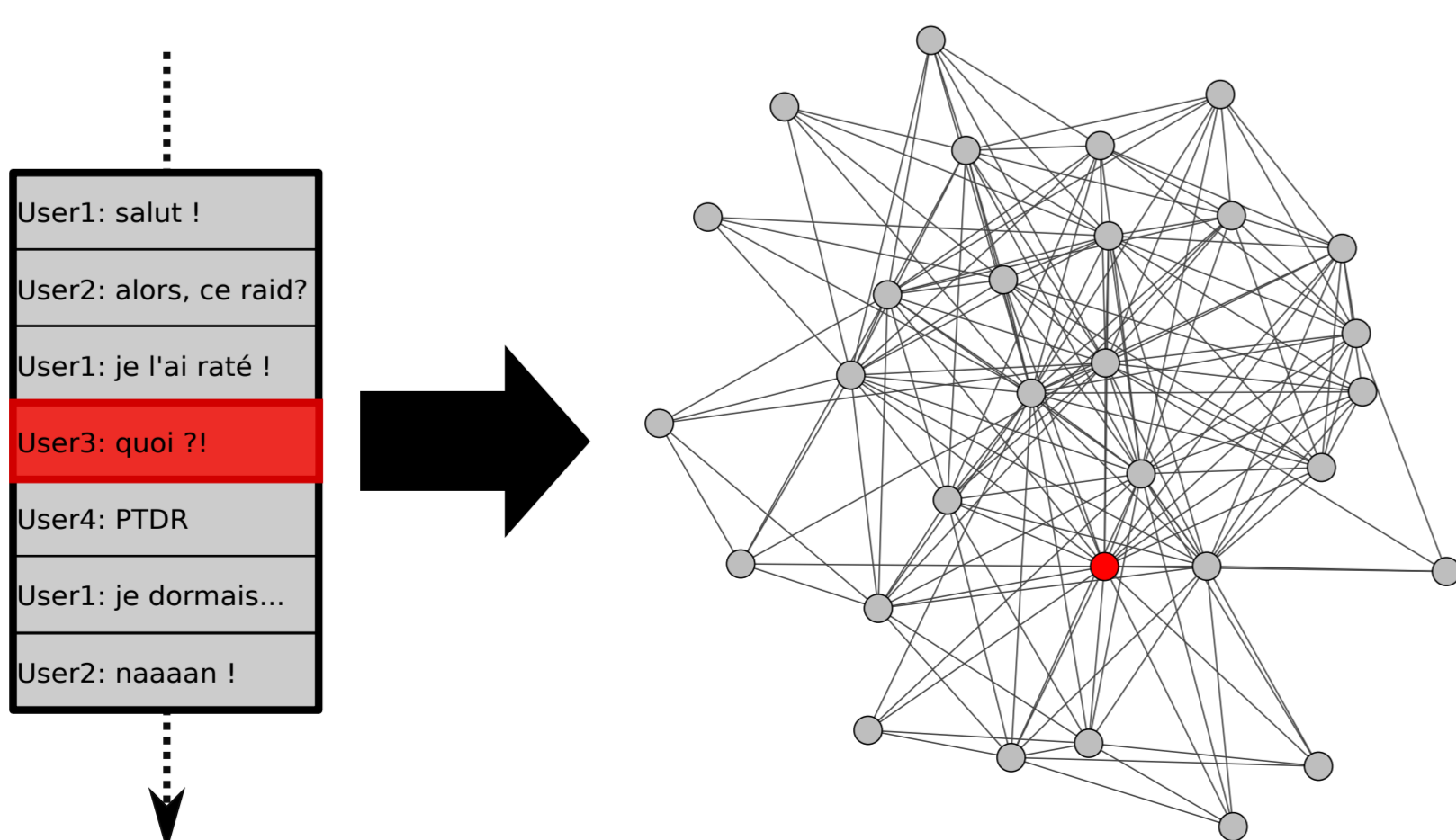
Content-based pipeline. NLP features include the count of unique characters in the message, classes of characters, proportion of capital letters and numbers.

A major limitation of the content-based methods is their sensitivity to intentional text obfuscation. Example :

Pi&ce Of sh1t, D4mn, f*ck y, bul..hit**

Conversational graph extraction

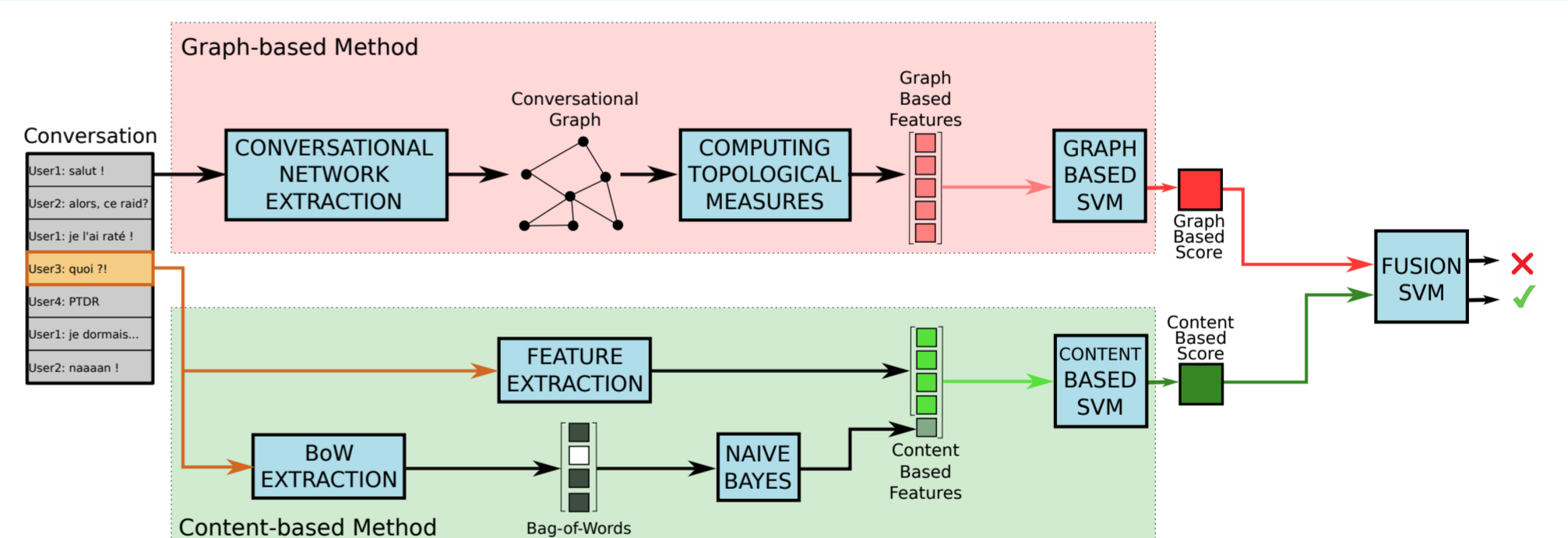
- Extraction of a conversational graph from a sequence of messages.
- Vertices represent users and edges model interactions between them.



A sequence of messages and the corresponding graph.

Fusion

We propose a method seeking to take advantage of both previous ones by combining them.



Graph- and Content-based methods combined through a third SVM.

Results

- Graph-based method performs better than Content-based method.
 - Fusion method obtains better results than them
- Graph- and Content-based methods are complementary

Method	Number of features	Precision	Recall	F-measure
Content-Based	29	78.59	83.61	81.02
Graph-Based	459	90.21	87.63	88.90
Fusion	488 (2)	94.10	92.43	93.26

Comparison of the performances obtained with the methods Content-based, Graph-based and their Fusion. The dataset is extracted from the in-game chat of an online video-game.

Perspectives

- Replace NLP features extraction with text embedding.
- Replace graph topological measures with whole graph embedding.



CV HAL



Noé Cécillon
noe.cecillon@univ-avignon.fr

PhD student in Computer Science at Avignon University.
Main research interests :

- Representation learning
- Natural Language Processing