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

Noé Cécillon[1] Vincent Labatut[1]
Richard Dufour[1] Georges Linarès[1]

[1]Laboratoire Informatique d'Avignon, Avignon Université – LIA EA 4128
{firstname.lastname}@univ-avignon.fr

Soc2Net: International workshop on Modeling and mining Social-Media-driven Complex Networks
Munich, Germany, June 11 2019

LABORATOIRE INFORMATIQUE D'AVIGNON

AVIGNON UNIVERSITÉ

Outline

Context
Online Communities & Moderation

- Online communities
  - Important medium: widely used, high socio-economical impact
  - Users are usually anonymous
- Abusive behavior
  - Violation of community rules
  - Can lead to: community degradation, legal consequences

→ Moderation
  - Detecting abusive users and applying sanctions
  - Usually done by hand: costly task (time, money)

Context
Online Communities & Moderation

- Online communities
  - Important medium: widely used, high socio-economical impact
  - Users are usually anonymous
- Abusive behavior
  - Violation of community rules
  - Can lead to: community degradation, legal consequences

→ Moderation
  - Detecting abusive users and applying sanctions
  - Usually done by hand: costly task (time, money)

Context
Online Communities & Moderation

- Online communities
  - Important medium: widely used, high socio-economical impact
  - Users are usually anonymous
- Abusive behavior
  - Violation of community rules
  - Can lead to: community degradation, legal consequences

→ Moderation
  - Detecting abusive users and applying sanctions
  - Usually done by hand: costly task (time, money)

Context
Automated moderation

- Automation
  - Assistance: raise messages to moderator's attention
  - Full moderation: detect abuse and apply sanctions
  - Not a trivial problem
    - Noise (can be intentional)
    - Natural language
    - Context
- In this work:
  - Detection of abusive messages as a binary classification task
  - Application to data from the SpaceOrigin MMORPG

Context
Automated moderation

- Automation
  - Assistance: raise messages to moderator's attention
  - Full moderation: detect abuse and apply sanctions
  - Not a trivial problem
    - Noise (can be intentional)
    - Natural language
    - Context
- In this work:
  - Detection of abusive messages as a binary classification task
  - Application to data from the SpaceOrigin MMORPG

Literature
Quick Review of Abuse Detection Works

- **Content**-Based Approaches [Spe97; Che+12; DRL11; CS15]
  - Badwords dictionaries
  - Static rules
  - Word n-gram approaches
  - Bag-of-Words models (tf–idf)
- **Context**-Based approaches [Yin+09; CDL15; BS15; Gar+16]
  - Content of neighboring messages
  - User models (language, behavior)
  - Interactions outside of discussions (ex. subscriptions)
- CORIA'17 [Pap+17a] & TransCSS'18 [Pap+19]
  - Morphological features: char. count, compression rate
  - Graph-based modeling of the conversations
  - Conversational network

Literature
Quick Review of Abuse Detection Works

- **Content**-Based Approaches [Spe97; Che+12; DRL11; CS15]
  - Badwords dictionaries
  - Static rules
  - Word n-gram approaches
  - Bag-of-Words models (tf–idf)
- **Context**-Based approaches [Yin+09; CDL15; BS15; Gar+16]
  - Content of neighboring messages
  - User models (language, behavior)
  - Interactions outside of discussions (ex. subscriptions)
- CORIA'17 [Pap+17a] & TransCSS'18 [Pap+19]
  - Morphological features: char. count, compression rate
  - Graph-based modeling of the conversations
  - Conversational network

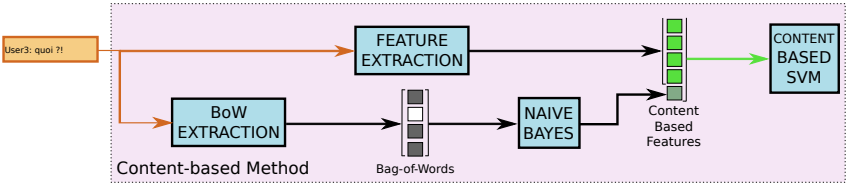
Literature
Quick Review of Abuse Detection Works

- **Content**-Based Approaches [Spe97; Che+12; DRL11; CS15]
  - Badwords dictionaries
  - Static rules
  - Word n-gram approaches
  - Bag-of-Words models (tf–idf)
- **Context**-Based approaches [Yin+09; CDL15; BS15; Gar+16]
  - Content of neighboring messages
  - User models (language, behavior)
  - Interactions outside of discussions (ex. subscriptions)
- CORIA'17 [Pap+17a] & TransCSS'18 [Pap+19]
  - Morphological features: char. count, compression rate
  - Graph-based modeling of the conversations
  - Conversational network

Proposed Method
Overview

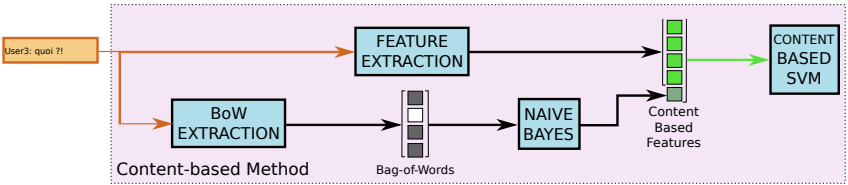
1. Combine content- and graph-based methods
2. Based on the two previously developed methods
   - 3 fusion strategies
   - Constitution of a global feature set containing all content- and graph-based features
   - Computation of two scores corresponding to the output probability of each message to be abusive given by the content- and graph-based methods
3. Train new classifiers using these features/scores

Proposed Method
Content-based approach



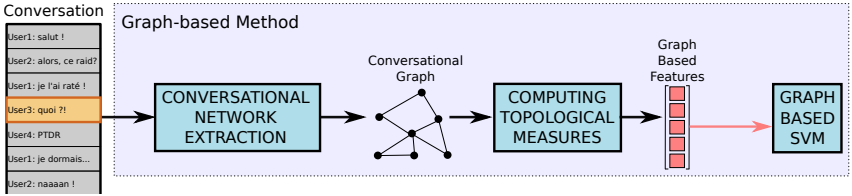
- Bag of Words representing the messages are used to train a Naive Bayes classifier
- We extract classic features from the raw message (no preprocessing)
- Naive Bayes output is used as input

Proposed Method
Content-based approach



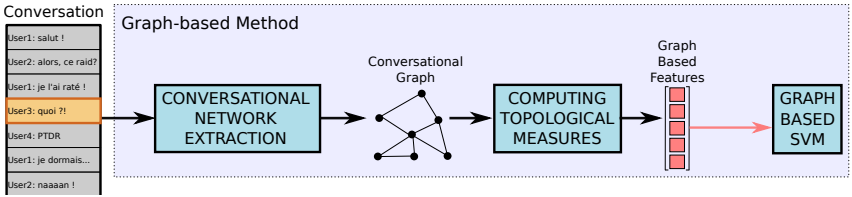
- Bag of Words representing the messages are used to train a Naive Bayes classifier
- We extract classic features from the raw message (no preprocessing)
- Naive Bayes output is used as input

Proposed Method
Graph-based approach



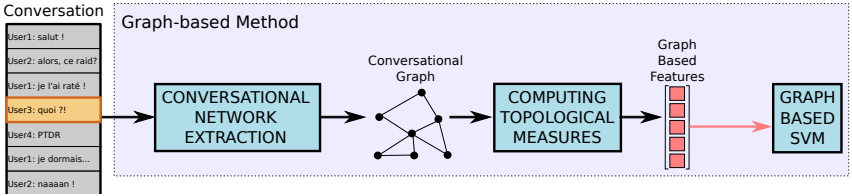
1. Extract conversational networks
   - Weighted directed graph
   - Build around a *targeted message*
   - Spawns a predefined *context period*
   - Nodes: active users within the context period
   - Links: message-based interactions between users
   - Weights: intensity of the interaction
2. Compute topological measures
3. Use them as features to train a classifier

Proposed Method
Graph-based approach



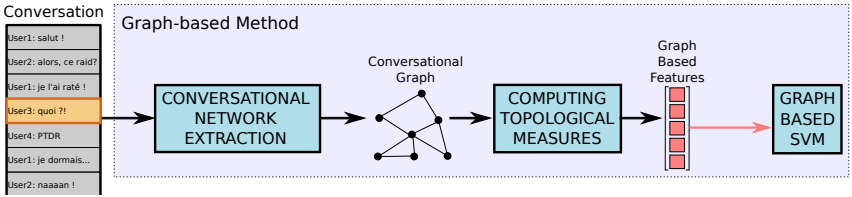
1. Extract conversational networks
   - Weighted directed graph
   - Build around a *targeted message*
   - Spawns a predefined *context period*
   - Nodes: active users within the context period
   - Links: message-based interactions between users
   - Weights: intensity of the interaction
2. Compute topological measures
3. Use them as features to train a classifier

Proposed Method
Graph-based approach



1. Extract conversational networks
   - Weighted directed graph
   - Build around a *targeted message*
   - Spawns a predefined *context period*
   - Nodes: active users within the context period
   - Links: message-based interactions between users
   - Weights: intensity of the interaction
2. Compute topological measures
3. Use them as features to train a classifier

Proposed Method
Graph-based approach



1. Extract conversational networks
   - Weighted directed graph
   - Build around a *targeted message*
   - Spawns a predefined *context period*
   - Nodes: active users within the context period
   - Links: message-based interactions between users
   - Weights: intensity of the interaction
2. Compute topological measures
3. Use them as features to train a classifier

Graph-based approach
Topological Measures

- Standard measures covering all scopes and scales

	Graph-scale	Node-scale
Macroscopic	Component counts, Adhesion, Cohesion, Articulation points, Radius, Diameter, Average distance	Spectral centralities, Subgraph centrality, Betweenness, Closeness, Eccentricity, Articulation point
Mesoscopic	Clique count, Communities, Modularity	Coreness Score Participation, Diversity Intensities, Heterogeneity
Microscopic	Node & Link counts, Density, Reciprocity, Global Transitivity, Degree Assortativity	Degree, Strength, Local Transitivity, Burt's Constraint

- Most measures have directed and/or weighted variants
- Additional graph-scale measures: average nodal measures

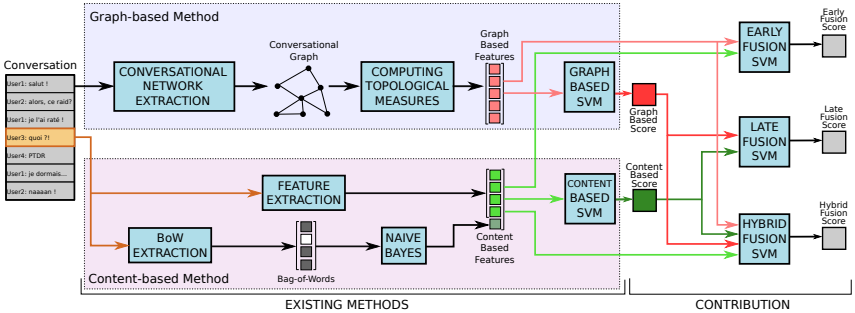
Graph-based approach
Topological Measures

- Standard measures covering all scopes and scales

	Graph-scale	Node-scale
Macroscopic	Component counts, Adhesion, Cohesion, Articulation points, Radius, Diameter, Average distance	Spectral centralities, Subgraph centrality, Betweenness, Closeness, Eccentricity, Articulation point
Mesoscopic	Clique count, Communities, Modularity	Coreness Score Participation, Diversity Intensities, Heterogeneity
Microscopic	Node & Link counts, Density, Reciprocity, Global Transitivity, Degree Assortativity	Degree, Strength, Local Transitivity, Burt's Constraint

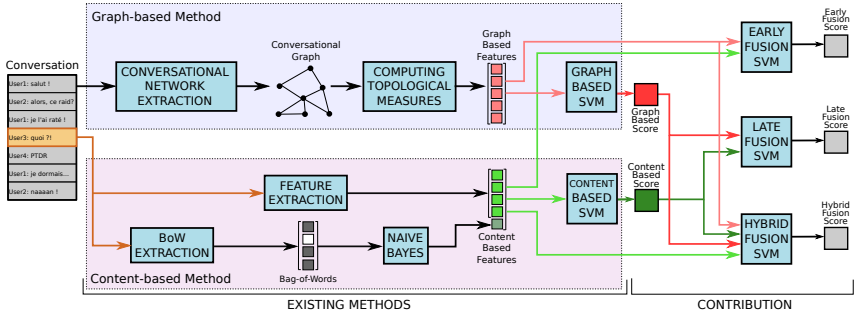
- Most measures have directed and/or weighted variants
- Additional graph-scale measures: average nodal measures

Proposed Method
Fusion approach



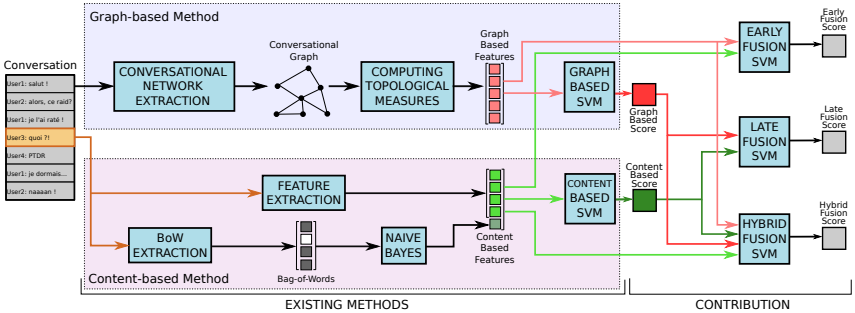
- *Early Fusion:* Global feature set containing all content- and graph-based features
- *Late Fusion:* Two scores corresponding to the output probability of each message to be abusive
- *Hybrid Fusion:* Create a feature set containing the content- and graph-based features and both scores

Proposed Method
Fusion approach



- *Early Fusion:* Global feature set containing all content- and graph-based features
- *Late Fusion:* Two scores corresponding to the output probability of each message to be abusive
- *Hybrid Fusion:* Create a feature set containing the content- and graph-based features and both scores

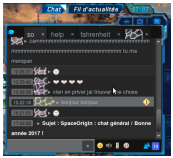
Proposed Method
Fusion approach



- *Early Fusion:* Global feature set containing all content- and graph-based features
- *Late Fusion:* Two scores corresponding to the output probability of each message to be abusive
- *Hybrid Fusion:* Create a feature set containing the content- and graph-based features and both scores

Results
Dataset & Experimental Protocol

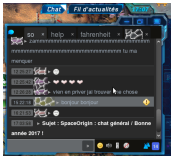
- Dataset
  - Chat logs from the SpaceOrigin MMORPG
  - $4{,}029{,}343$ instant messages
    - 779 messages flagged and later confirmed as abusive
    - Sample of 779 messages assumed non-abusive
    - All messages taken from distinct conversations



- Classification
  - SVM (Sklearn C-Support Vector Classification)
  - Cross validation with $70{-}30\%$ split
  - Feature importance estimated using ExtraTreesClassifier (Sklearn)

Results
Dataset & Experimental Protocol

- Dataset
  - Chat logs from the SpaceOrigin MMORPG
  - $4{,}029{,}343$ instant messages
    - 779 messages flagged and later confirmed as abusive
    - Sample of 779 messages assumed non-abusive
    - All messages taken from distinct conversations



- Classification
  - SVM (Sklearn C-Support Vector Classification)
  - Cross validation with $70{-}30\%$ split
  - Feature importance estimated using ExtraTreesClassifier (Sklearn)

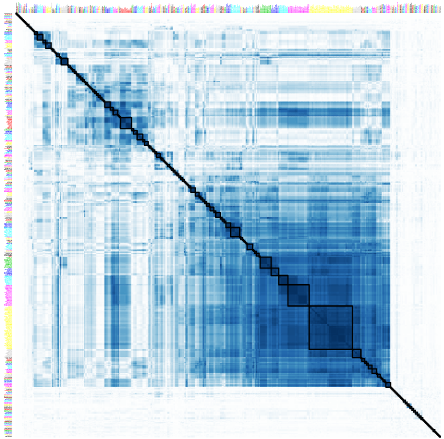
Scores relative to the *Abuse* class:

	Approach	Precision	Recall	F-measure
Baseline	Content-based [Pap+17b]	0.79	0.84	0.81
	Graph-based [Pap+19]	0.90	0.88	0.89
Contribution	Early Fusion	0.91	0.89	0.90
	Late Fusion	**0.94**	**0.92**	**0.93**
	Hybrid Fusion	0.92	0.90	0.91

- Better performances using both content and graph
- Classifier can be used to assist moderators

Results
Feature Correlation Study

- Some features are highly correlated
  - Directed/weighted variants of the same measure
  - Measures based on similar principles
  - Application-specific reasons
- Cluster analysis
  - Measures in the same cluster are considered equivalent



Results
Feature Selection

- Application of a feature ablation method (Sklearn)
- Top-Features (TF): minimal subset of features reaching 97% of the original performance

Method	Number of features	Total Runtime	Average Runtime	F-measure
Content-Based	29	0:52	0.02 s	0.81
Content-Based TF	3	0:21	0.01 s	0.79
Graph-Based	459	8:19:10	7.56 s	0.89
Graph-Based TF	10	14:22	0.03 s	0.87
Early Fusion	488	8:26:41	7.68 s	0.90
Early Fusion TF	4	11:29	0.17 s	0.88
Late Fusion	488(2)	8:23:57	7.64 s	0.93
Late Fusion TF	13(2)	15:42	0.24 s	0.91
Hybrid Fusion	490	8:27:01	7.68 s	0.91
Hybrid Fusion TF	4	16:57	0.26 s	0.90

Results
Feature Selection

- Application of a feature ablation method (Sklearn)
- Top-Features (TF): minimal subset of features reaching 97% of the original performance

Method	Number of features	Total Runtime	Average Runtime	F-measure
Content-Based	29	0:52	0.02 s	0.81
Content-Based TF	3	0:21	0.01 s	0.79
Graph-Based	459	8:19:10	7.56 s	0.89
Graph-Based TF	10	14:22	0.03 s	0.87
Early Fusion	488	8:26:41	7.68 s	0.90
Early Fusion TF	4	11:29	0.17 s	0.88
Late Fusion	488(2)	8:23:57	7.64 s	0.93
Late Fusion TF	13(2)	15:42	0.24 s	0.91
Hybrid Fusion	490	8:27:01	7.68 s	0.91
Hybrid Fusion TF	4	16:57	0.26 s	0.90

Conclusions & Perspectives

- Main results
  - Better results combining graph- and content-based approaches than using them separately
  - Performance good enough for moderation support, not full automation
  - Limits: computational cost, no real-time
  - Possible to keep 97% of the original performance using a small subset of relevant features
- Perspectives
  - Find other corpora to test our methods at a much higher scale
  - Test our methods on an other language than French
  - Explore representation learning to derive more efficient features

Conclusions & Perspectives

- Main results
  - Better results combining graph- and content-based approaches than using them separately
  - Performance good enough for moderation support, not full automation
  - Limits: computational cost, no real-time
  - Possible to keep 97% of the original performance using a small subset of relevant features
- Perspectives
  - Find other corpora to test our methods at a much higher scale
  - Test our methods on an other language than French
  - Explore representation learning to derive more efficient features

Questions?

AVIGNON UNIVERSITÉ

Additional Material
Top-features

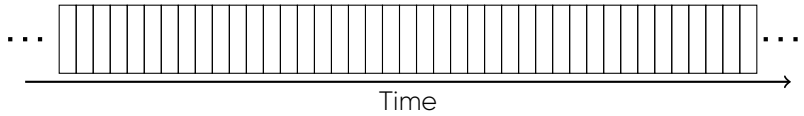
Method	Top Features
Content-Based	Naive Bayes $tf-idf$ Abuse Score Character Capital Ratio
Graph-Based	Coreness Score PageRank Centrality Strength Centrality Vertex Count Closeness Centrality Closeness Centrality Authority Score Hub Score Reciprocity Closeness Centrality

Additional Material
Top-features

Method	Top Features
Early Fusion	Coreness Score Coreness Score Eccentricity Naive Bayes
Late Fusion	*Content-Based* TF ∪ *Graph-Based TF*
Hybrid Fusion	Graph-based output Content-based output Strength Centrality Coreness Score

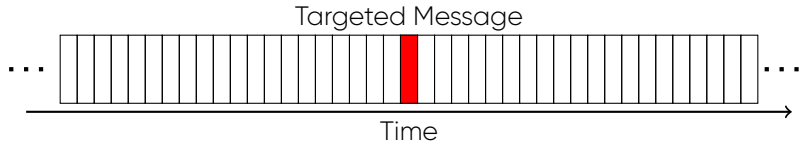
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



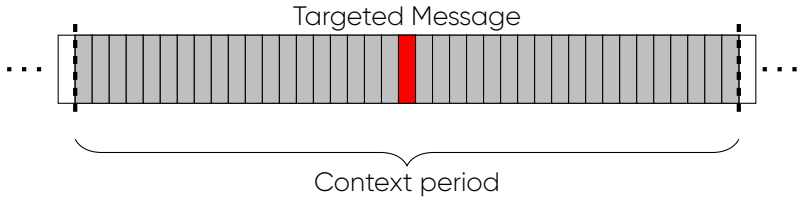
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



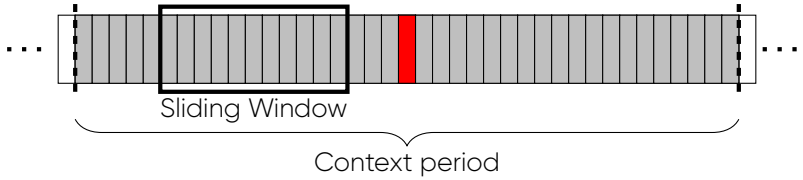
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



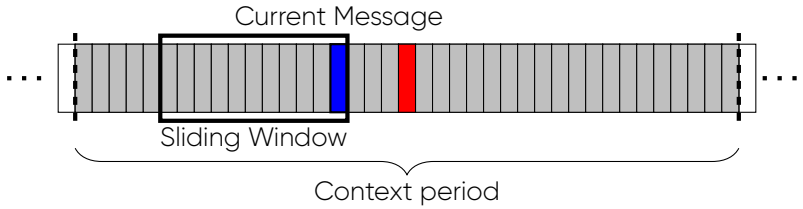
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



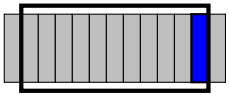
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



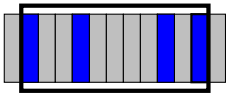
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



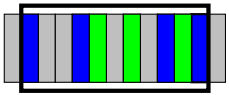
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



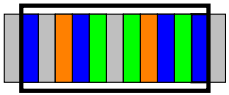
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



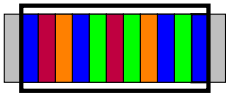
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



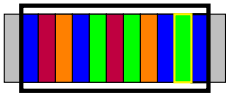
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



1. ■

Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph

Additional Material
Conversational Network Extraction

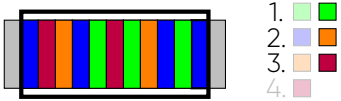
1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph

Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph

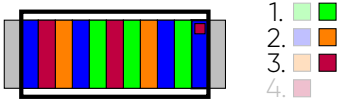
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



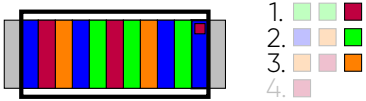
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



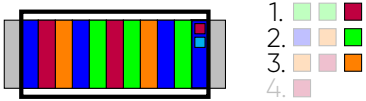
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



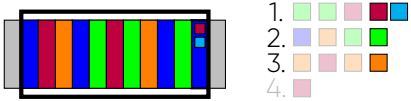
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



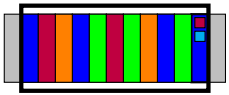
Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



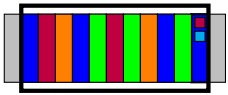
Additional Material
Conversational Network Extraction

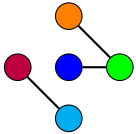
1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph



Additional Material
Conversational Network Extraction

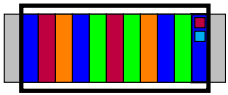
1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph

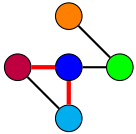




Additional Material
Conversational Network Extraction

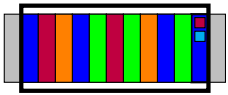
1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph

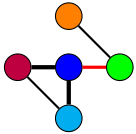




Additional Material
Conversational Network Extraction

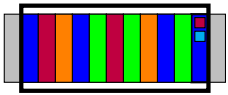
1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph

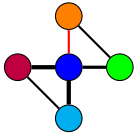




Additional Material
Conversational Network Extraction

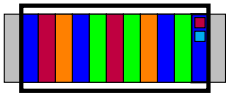
1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph

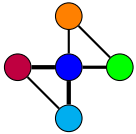




Additional Material
Conversational Network Extraction

1. Define context period, centered on *targeted message*
2. Slide *current message*-related window over conversation
3. Compute link weights
   - Hyp.#1: current message targeted towards other participants
   - Hyp.#2: message firstly addressed to last active users
   - Hyp.#3: directly referenced users even more targeted
4. Update graph





References I

[BS15] K. Balci and A. A. Salah. "Automatic analysis and identification of verbal aggression and abusive behaviors for online social games". In: *Computers in Human Behavior* 53 (2015), pp. 517–526. DOI: 10.1016/j.chb.2014.10.025.

[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 International Conference on Advances in Computing, Communications and Informatics*. 2015, pp. 2354–2358. DOI: 10.1109/ICACCI.2015.7275970.

[Che+12] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. "Detecting offensive language in social media to protect adolescent online safety". In: *International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*. 2012, pp. 71–80. DOI: 10.1109/SocialCom-PASSAT.2012.55.

[CDL15] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. "Antisocial Behavior in Online Discussion Communities". In: *9th International AAAI Conference on Web and Social Media*. 2015, pp. 61–70. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10469.

[DRL11] K. Dinakar, R. Reichart, and H. Lieberman. "Modeling the detection of Textual Cyberbullying". In: *5th International AAAI Conference on Weblogs and Social Media / Workshop on the Social Mobile Web*. 2011, pp. 11–17. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841.

References II

[Gar+16] 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*. 2016, pp. 33–42. DOI: 10.1145/2835776.2835792.

[Pap+19] É. Papégnies, V. Labatut, R. Dufour, and G. Linarès. "Conversational Networks for Automatic Online Moderation". In: *IEEE Transactions on Computational Social Systems* in press (2019). DOI: 10.1109/TCSS.2018.2887240.

[Pap+17a] É. Papégnies, V. Labatut, R. Dufour, and G. Linarès. "Detection of abusive messages in an on-line community". In: *14ème Conférence en Recherche d'Information et Applications*. 2017, pp. 153–168. DOI: 10.24348/coria.2017.16.

[Pap+17b] É. Papégnies, V. Labatut, R. Dufour, and G. Linarès. "Impact Of Content Features For Automatic Online Abuse Detection". In: *18th International Conference on Computational Linguistics and Intelligent Text Processing*. Vol. 10762. Lecture Notes in Artificial Intelligence. 2017, pp. 404–419. DOI: 10.1007/978-3-319-77116-8_30.

[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*. 1997, pp. 1058–1065. URL: http://dl.acm.org/citation.cfm?id=1867616.

References III

[Yin+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, pp. 1–7. URL: http://www.cse.lehigh.edu/~brian/pubs/2009/CAW2/.