LABORATOIRE DES SCIENCES DU NUMÉRIQUE DE NANTES

TALN

Représentation multimodale de conversations pour la détection de messages abusifs

Richard Dufour

12/05/2022

Séminaire équipe TALEP

cnrs

CENTRALE NANTES

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Nantes Université

Inria

Equipe TALN

- Equipe d'accueil CNRS rattachée au LS2N (Laboratoire des Sciences du Numérique de Nantes)
- Environ 30 personnes
  - 11 permanents + 9 doctorants + stagiaires, ingénieurs…
- Thématiques
  - Analyse sémantique et discursive
  - Apprentissage & fouille de textes
  - Alignement multilingue et multimodal
- Domaines
  - Education
  - Santé
  - Documents scientifiques
- Chaire Unesco RELIA (Ressources Educatives Libres et Intelligence Artificielle)
  - Organisation conférence internationale OEGlobal 2022 fin mai

Plan de la présentation

- Contexte du travail
- Utilisation du contenu textuel
- Représentation par graphes conversationnels
- Utilisation conjointe texte et graphe
- Conclusions et Perspectives

Contexte général

- Problématique : Comment représenter au mieux le contenu d'un document en vue de son traitement automatique ?
  - Problème historique en traitement du langage
  - Quasi-impossible d'utiliser directement le contenu sans traitement préalable
    - Vocabulaire très grand
    - Spécificités de chaque langue (ex : homographes, règles grammaticales complexes, évolution au cours du temps…)
    - Nature des documents (message court, texte écrit, transcription d'un énoncé oral)
    - …
- Généralement, représentation en entrée d'autres problèmes en traitement automatique du langage - TAL (classification automatique, indexation, génération d'informations…)

Cadre applicatif

- Détection de messages abusifs sur Internet
  - Regroupement en communautés en ligne
  - Nombre croissant de messages à traiter
- Besoin de modération par les plateformes (règlements, lois…)
  - Tâche généralement manuelle (coûts importants)
  - Travaux automatiques menés en TAL
    - Modélisation statistique ou à base de règles
- Difficultés
  - Les utilisateurs « camouflent » de plus en plus leurs commentaires
    - Création de termes compréhensibles par les humains mais inconnus par les systèmes d'apprentissage
  - Données considérées au niveau du message seul : manque de corpus
- Systèmes automatiques actuels : efficaces pour aide mais encore insuffisants pour une modération automatique
- Domaine : hate speech

Automatisation de la modération

- Cas d'usage
  - Simple assistance : attirer l'attention des modérateurs sur des cas d'abus potentiels
  - Modération complète : détection des abus et application des sanctions
  - Des questions sociétales se posent néanmoins sur le côté « automatique »
- Problème difficile - cf. Google Perspective API [Hosseini17]
  - Gère mal les négations
  - Gère mal le « masquage » (ex : stu.ipd)
- Travaux ici dans le cadre d'une discussion (et non de messages simplement isolés)

[Hosseini17] Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:1702.08138.

Détection d'abus

- Approches liées au contenu direct du message
  - Dictionnaires de mots abusifs
  - Règles prédéfinies
  - Approches à bases de n-grammes de mots
  - Modèles sac-de-mots (TF–IDF)
  - Approches récentes par apprentissage automatique (embeddings contextuels + caractéristiques linguistiques, réseaux de neurones récurrents, Bi-LSTM
- Approches liées au contexte
  - Contenu des messages voisins
  - Modèles d'utilisateurs (langage, comportement)
  - Interactions hors-discussion (ex. abonnements)
- Travail mêlant ici le message ainsi que son contexte

Contexte du travail

- Travaux initiés avec Vincent Labatut (MCF), Georges Linarès (PR) et Etienne Papegnies au Laboratoire Informatique d'Avignon (LIA)
- Thèse de Noé Cecillon débutée en 2018
  - Travaux autour de la modélisation par embeddings et fusion d'informations
- Collaboration initiale avec l'entreprise Nectar de Code
- Données
  - Discussions sous forme de messages instantanés (tchat) du jeu en ligne SpaceOrigin
  - Deux classes : « abus » et « non abus »
  - 1320 messages annotés (50 % pour chaque classe)
  - Validation croisée en 10 parties
  - Evaluation en précision, rappel, f-mesure
  - Classifieur SVM pour toutes les expériences

Contexte du travail

UTILISATION DU CONTENU TEXTUEL

Représentation par graphes conversationnels
Utilisation conjointe texte et graphe
Conclusions et Perspectives

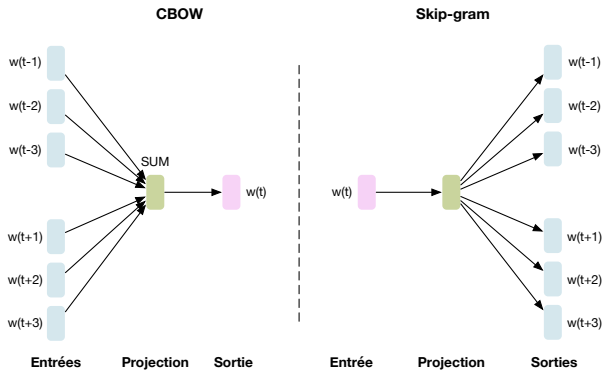
1ers travaux : Caractéristiques textuelles de base

- Extraction d'un ensemble de caractéristiques liées au message [Papegnies17]
  - Approche classique en TAL
- Caractéristiques morphologiques
  - Longueur du message (en mots et caractères)
  - Compression du message (i.e. allongement des mots)
  - …
- Caractéristiques liées à la langue
  - Pré-traitement des messages
  - Nombre d'occurrences de mots
  - Liste de mots abusifs
  - Moyenne des scores Term Frequency-Inverse Document Frequency (TF-IDF) du message
  - …
- Représentation du message : Vecteur comprenant les différentes caractéristiques

[Papegnies 17] Papegnies, E., Labatut, V., Dufour, R., & Linares, G. (2017). Impact of content features for automatic online abuse detection. In CLICLING (pp. 404-419). Springer, Cham.

Etude sur les représentations par plongement lexical (1/2)

- Intérêt : évaluer les approches de représentation de mots à l'état de l'art sur la tâche de détection d'abus
  - Permet aussi de s'abstraire des caractéristiques a priori
- Vecteur fixé
  - Quelque que soit le contexte, le mot est toujours représenté par le même vecteur
  - Word2vec [Mikolov13]
    - Approche par réseaux de neurones
    - Tente de conserver une proximité sémantique et/ou syntaxique entre les mots
    - Limitation sur les mots hors-vocabulaire



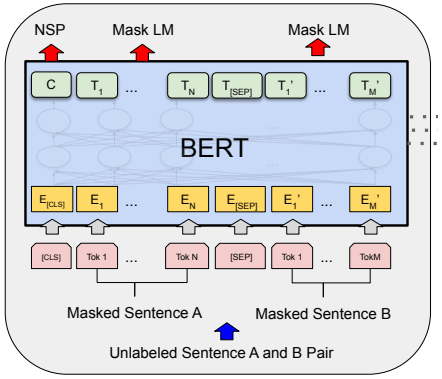
  - FastText [Bojanowski17]
    - Extension de word2vec : traite des n-grammes de caractères
    - Par exemple, avec n= 3, le mot « mode » serait représenté par (<mo,mod,ode,de>)
    - Vecteur associé à un mot est la somme de tous les n-grammes qui le composent et le mot

[Mikolov13] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

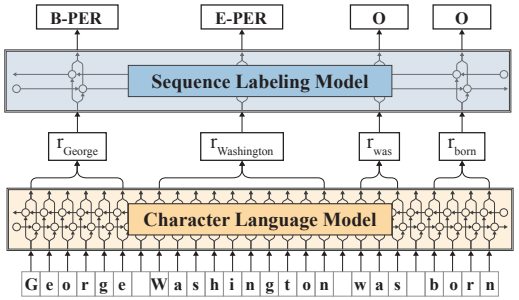
[Bojanowski17] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

Etude sur les représentations par plongement lexical (2/2)

- Vecteur variant selon le contexte
  - Un même mot utilisé dans des contextes différents obtiendra des représentations vectorielles différentes
  - Approche BERT [Devlin18]
    - S'appuie sur une architecture Transformer
    - Intègre le concept de self-attention et multi-head attention
    - Modèle de langage masqué
  - Approche Flair [Akbik18]
    - Modèle de langage + labélisation de séquences
    - Traite des caractères
    - LSTM bi-directionnel



Pre-training

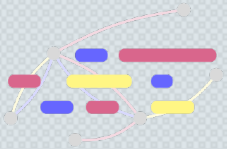


[Devlin18] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[Akbik18] Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In Proceedings of the 27th international conference on computational linguistics (pp. 1638-1649).

Résultats traitement contenu textuel

Échelle	Méthode	Dimension	F-mesure
-	Référence-Texte	29	79,89
Mot	Word2vec	300	75,21
	fastText	300	81,24
	Flair	2 048	82,56
	CamemBERT-W	1 024	89,79
	FlauBERT-W	1 024	85,12
Message	CamemBERT-M	1 024	89,62
	FlauBERT-M	1 024	79,79

Cecillon, N., Dufour, R., & Labatut, V. (2022). Approche multimodale par plongements de texte et de graphes pour la détection de messages abusifs. Revue TAL.



Contexte du travail
Utilisation du contenu textuel

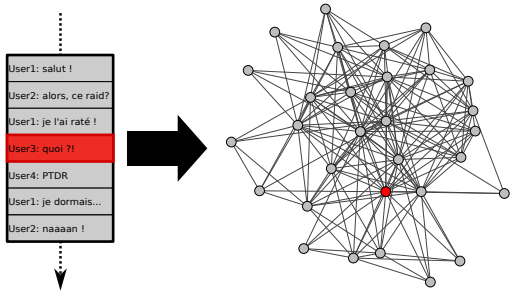
REPRÉSENTATION PAR GRAPHES CONVERSATIONNELS

Utilisation conjointe texte et graphe
Conclusions et Perspectives

LS2N

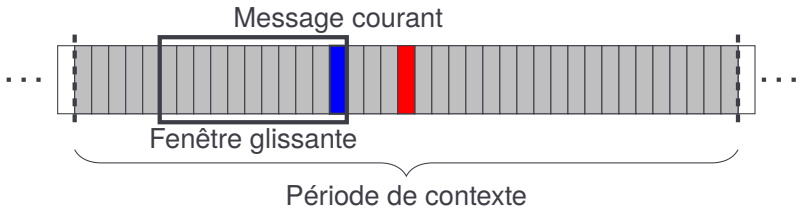
Contexte

- Limites précédemment évoquées
  - Assez simple de contrer les systèmes de TAL dans ce contexte
  - « Masquage » du langage, messages courts, données d'apprentissage limitées et liées à une langue…
- Proposition
  - Considération des interactions entre utilisateurs en dehors de tout contenu textuel
  - Modélisation de la structure des conversations sous forme d'un graphe conversationnel [Papegnies19]
  - Chaque noeud représente un utilisateur
  - Chaque lien représente une réponse
  - En rouge, l'utilisateur du message courant
- Avantages
  - Masquage quasi-impossible par les utilisateurs
  - Potentiellement indépendant de la langue

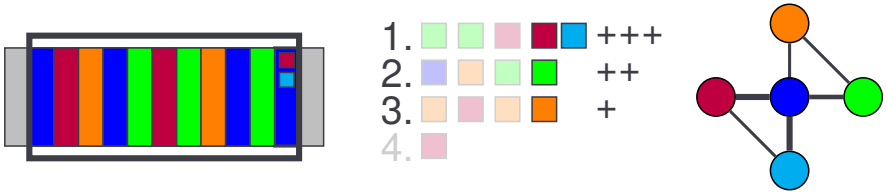


Extraction du graphe conversationnel

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



- 3. Calculer les poids des liens concernés
  - Hyp1 : message courant adressé aux autres intervenants
  - Hyp2 : auteurs plus récents sont plus concernés
  - Hyp3 : auteurs mentionnés sont encore plus visés



- 4. Mise à jour du graphe complet (i.e. construit à partir de toutes les fenêtres glissantes)

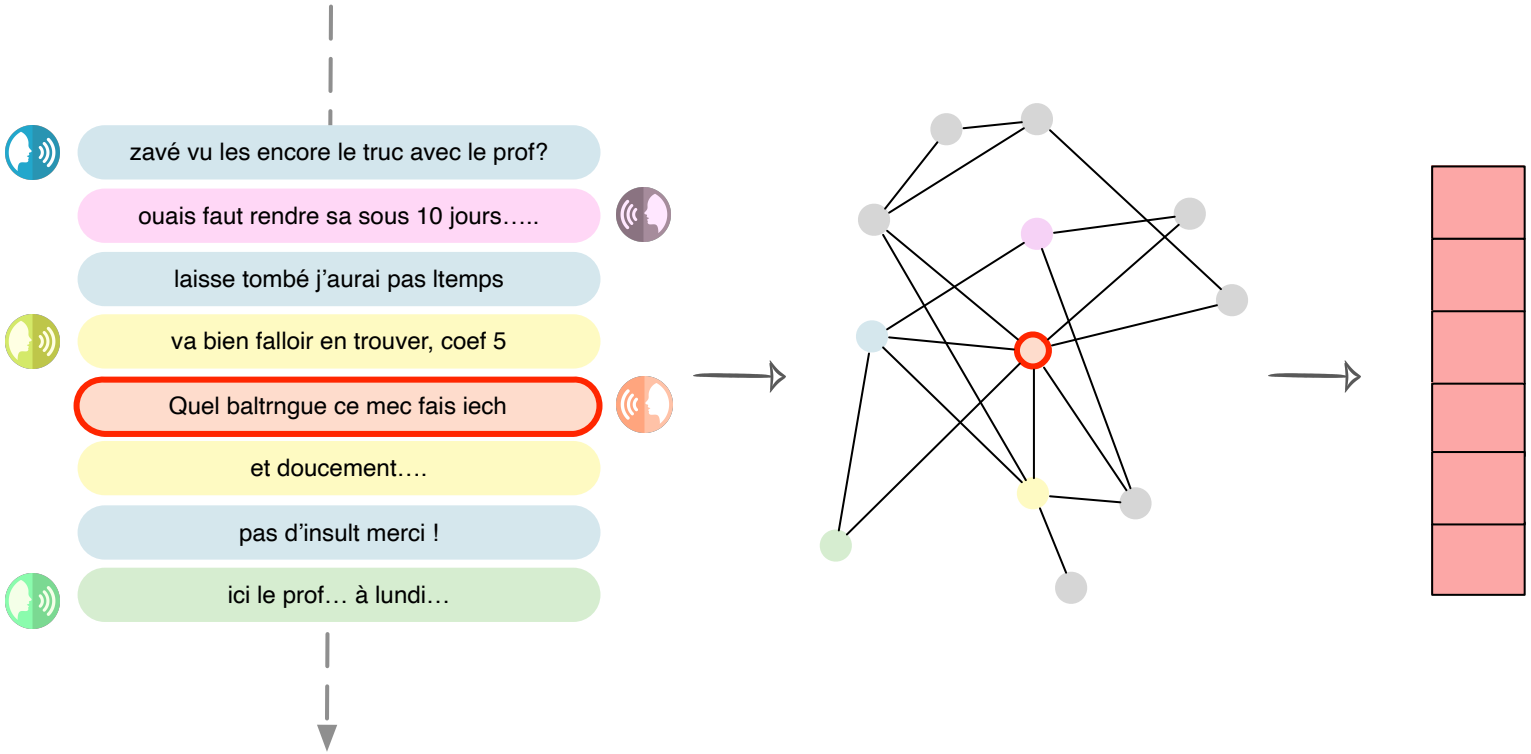
1er travaux : Mesures topologiques

- À l'instar des approches classiques en TAL, l'idée est ici d'extraire un ensemble de mesures du graphe
  - Le vecteur représentant alors un graphe est la concaténation de l'ensemble de ces mesures
  - Fourni, comme pour le texte, en entrée d'un classifieur
- Choix fait de prendre un très grand nombre de mesures classiques, puis d'étudier leur importance [Papegnies19]
- Extraction à différents niveaux
  - Graphe entier : diamètre, densité…
  - Noeud individuel : degré, centralité de proximité…
  - Au total, 459 caractéristiques

[Papegnies19] Papegnies, E., Labatut, V., Dufour, R., & Linarès, G. (2019). Conversational networks for automatic online moderation. IEEE Transactions on Computational Social Systems, 6(1), 38-55.

Approches par plongement (1/3)

- Permettent de transformer des noeuds, des liens ou même des graphes complets en vecteurs numériques de taille fixée, tout en préservant une partie de leurs propriétés structurelles



Approches par plongement (2/3)

- Approches par plongement de noeuds
  - Production d'un vecteur pour chaque noeud [Goyal18]
  - Exemple : Node2vec avec marches aléatoires [Grover 16]
    - Parallèle avec l'approche SkipGram de word2vec

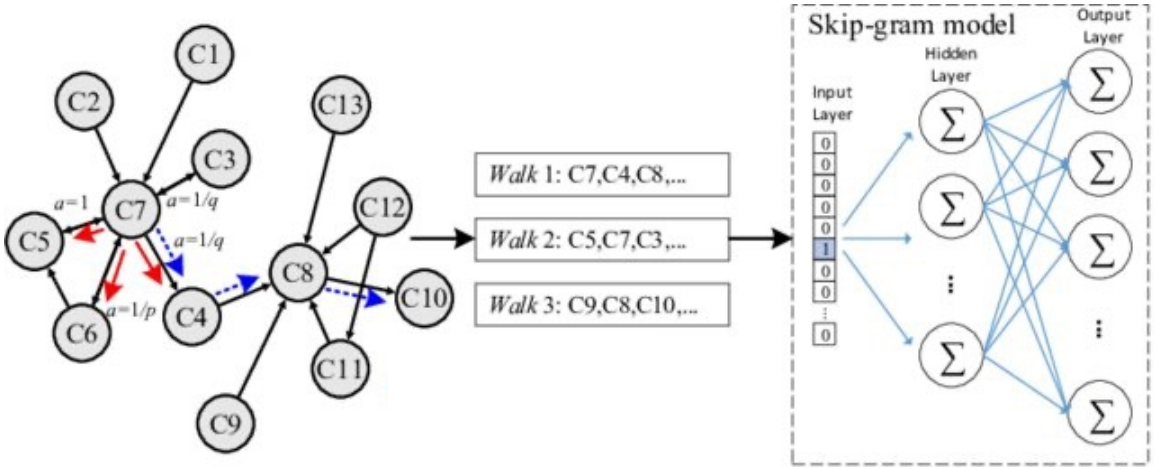


Figure : [Qu18]

[Qu18] Qu, Y., Liu, T., Chi, J., Jin, Y., Cui, D., He, A., & Zheng, Q. (2018). node2defect: using network embedding to improve software defect prediction. In 2018 33rd IEEE/ACM International Conference on ASE (pp. 844-849). IEEE.

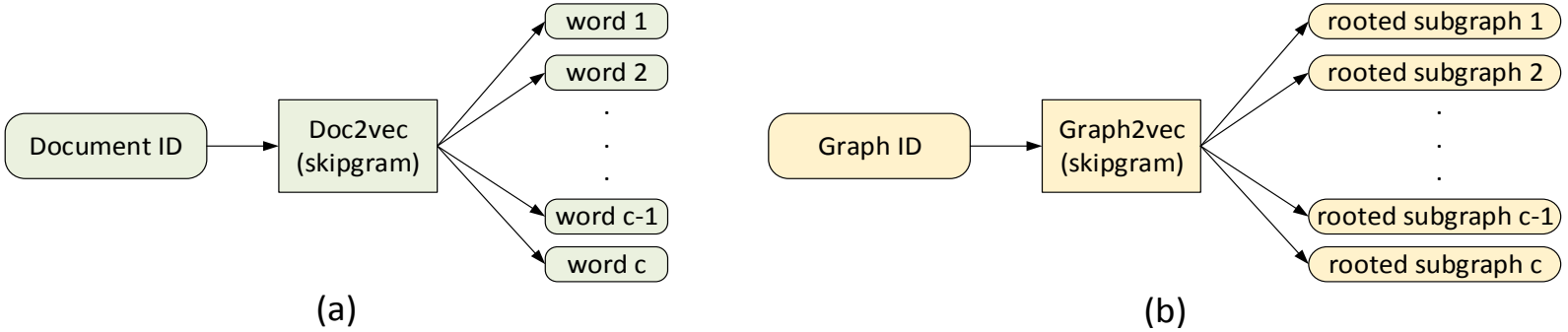
[Grover16] Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

[Goyal18] Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems, 151, 78-94.

[Cecillon21] N. Cecillon, V. Labatut, R. Dufour, G. Linarès (2021). Graph embeddings for Abusive Language Detection. Springer Nature (SN) Computer Science. 24 pages.

Approches par plongement (3/3)

- Approches par plongement de graphes entiers
  - Production d'un vecteur pour le graphe
  - Exemple : Graph2vec [Narayanan17]
    - Extension de doc2vec pour le TAL



[Narayanan17] Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., & Jaiswal, S. (2017). graph2vec: Learning distributed representations of graphs. arXiv preprint arXiv:1707.05005.

Résultats approche par graphes conversationnels

Échelle	Méthode	Dimension	F-mesure
-	Référence-Graphe	459	88,08
Nœuds	Node2vec	128	83,15
	GraphWave	200	82,89
Graphe entier	Spectral Features	128	79,59
	Graph2vec	128	81,91
	Graph2vec-auteur	128	82,21
	Graph2vec-cible	128	81,86
	Graph2vec-distance	128	84,30

Cecillon, N., Dufour, R., & Labatut, V. (2022). Approche multimodale par plongements de texte et de graphes pour la détection de messages abusifs. Revue TAL.

Contexte du travail
Utilisation du contenu textuel
Représentation par graphes conversationnels

UTILISATION CONJOINTE TEXTE ET GRAPHE

Conclusions et Perspectives

Contexte

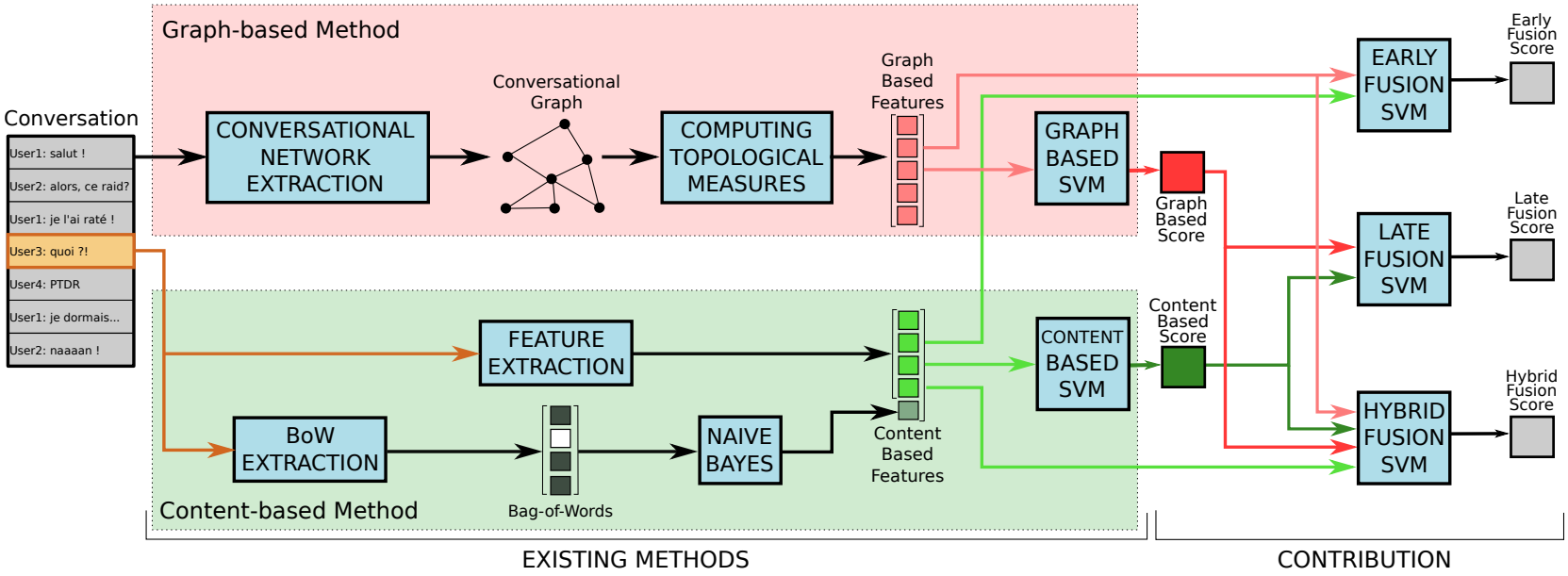
- Pour les approches classiques étudiées en texte et graphe, leur combinaison est intéressante et montre la complémentarité des sources d'information [Cecillon19]
  - Intéressant à étudier au niveau des plongements

Table 1. Comparison of the performances obtained with the methods (*Content-based*, *Graph-based*, *Fusion*) and their subsets of *Top Features* (TF). The total runtime is expressed as *h*:*min*:*s*. See text for details.

Method	Number of features	Total Runtime	Average Runtime	Precision	Recall	F-measure
Content-Based	29	0:52	0.02s	78.59	83.61	81.02
Content-Based TF	3	0:21	0.01s	75.82	82.57	79.05
Graph-Based	459	8:19:10	7.56s	90.21	87.63	88.90
Graph-Based TF	10	14:22	0.03s	88.72	84.87	86.75
Early Fusion	488	8:26:41	7.68s	91.25	89.45	90.34
Early Fusion TF	4	11:29	0.17s	89.09	87.12	88.09
Late Fusion	488 (2)	8:23:57	7.64s	94.10	92.43	93.26
Late Fusion TF	13	15:42	0.24s	91.64	89.97	90.80
Hybrid Fusion	490	8:27:01	7.68s	91.96	90.48	91.22
Hybrid Fusion TF	4	16:57	0.26s	90.74	89.00	89.86

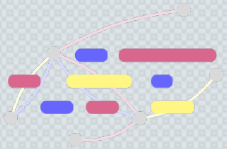
[Cecillon19] Cecillon, N., Labatut, V., Dufour, R., & Linarès, G. (2019). Abusive language detection in online conversations by combining content-and graph-based features. Frontiers in big Data, 2, 8.

Fusion


Graph-based Method
Conversation
User1: salut !
User2: alors, ce raid?
User1: je l'ai raté !
User3: quoi ?!
User4: PTDR
User1: je dormais...
User2: naaaan !
CONVERSATIONAL NETWORK EXTRACTION
Conversational Graph
COMPUTING TOPOLOGICAL MEASURES
Graph Based Features
GRAPH BASED SVM
Graph Based Score
FEATURE EXTRACTION
BoW EXTRACTION
Bag-of-Words
NAIVE BAYES
Content Based Features
CONTENT BASED SVM
Content Based Score
Content-based Method
EARLY FUSION SVM
Early Fusion Score
LATE FUSION SVM
Late Fusion Score
HYBRID FUSION SVM
Hybrid Fusion Score
EXISTING METHODS
CONTRIBUTION

Résultats fusion texte et graphe

	Précoce		Tardive		Hybride	
Méthodes	**Dim.**	***F*-m.**	**Dim.**	***F*-m.**	**Dim.**	***F*-m.**
Réf. Graphe + Réf. Texte	488	90,44	2	89,07	490	90,92
Cam-W + Cam-M	2 048	90,12	2	89,72	2 050	90,15
Cam-W + FlauBERT-W	2 048	89,69	2	89,86	2 050	90,09
Graph2vec-dist. + Node2vec	256	*87,07	2	*85,12	258	*84,14
Graph2vec-dist. + Cam-W	1 152	93,52	2	92,83	1 154	93,62
Node2vec + Cam-W	1 152	92,86	2	92,65	1 154	93,02
Réf. Graphe + Cam-W	1 483	94,23	2	93,54	1 485	94,43



Contexte du travail
Utilisation du contenu textuel
Représentation par graphes conversationnels
Conclusions et Perspectives

CONCLUSIONS ET PERSPECTIVES

Résumé des performances

Type	**Méthode(s)**	**Dim.**	***F*-m.**
Référence	Référence-Graphe Référence-Texte	459 29	88,08 79,89
Texte	CamemBERT-W	1 024	89,79
Graphe	Graph2vec-distance	128	84,30
Fusion texte-texte	CamemBERT-W + CamemBERT-M	2 050	90,15
Fusion graphe-graphe	Graph2vec-dist + Node2vec	256	87,07
Fusion graphe-texte	Référence-Graphe + CamemBERT-W	1 485	94,43

Conclusions

- Proposition d'une représentation par graphes conversationnels
  - Permet de rivaliser avec l'analyse du contenu textuel alors que la quantité de données d'apprentissage est négligeable comparativement
- Etude d'approches par plongement liant le contenu textuel et les interactions utilisateurs
- Les plongements surpassent les approches classiques
  - Evitent de définir une liste de caractéristiques
  - Plus rapides à extraire
- Proposition de stratégies d'étiquetage des noeuds pour les plongements par graphe [Cecillon20]

[Cecillon20] Cecillon, N., Dufour, R., Labatut, V., & Linares, G. (2020). Tuning Graph2vec with Node Labels for Abuse Detection in Online Conversations. In MARAMI.

Perspectives

- Étudier l'impact des méthodes de classification
- Vérifier sur d'autres jeux de données et dans d'autres langues
- Prendre en compte l'aspect dynamique des conversations
- Intégrer les informations multimodales dans une même architecture
  - Ne plus travailler au niveau message mais conversation (ex : Flair)
  - Intégrer des informations issues d'autres modalités dans les architectures (ex : mécanisme d'attention dans BERT, en entrée des systèmes de représentation…)
- Vérifier la prise en compte du contexte (et en particulier le positionnement des mots) par rapport aux représentations contextuelles
- Est-ce que les travaux de la communauté vont dans la bonne direction ?

MERCI DE VOTRE ATTENTION

Questions ?