

Des émotions au fil du récit : fabula, un package pour analyser les textes francophones par Transformers

Florian Cafiero^{1,2} , and Alexandre Lionnet³ 

¹ LRE, Ecole Pour l'Informatique et les Techniques Avancées (EPITA), Paris, France

² École nationale des chartes, Université Paris Sciences & Lettres (PSL), Paris, France

³ École Pratique des Hautes Études, Université Paris Sciences & Lettres (PSL), Paris, France

Abstract

Computational emotion analysis has become an established tool in computational literary studies for describing and comparing narratives. However, the solutions currently available for French remain limited and often rely on fixed lexicons that are not very robust to context, such as polysemy, negation, or irony, and that are difficult to relate back to the text in a fine-grained way. We present here `fabula-fr`, a Python package designed for emotion analysis in Francophone narratives. It is based on Transformer models while retaining a simple and reproducible pipeline. `fabula` offers several segmentation and smoothing strategies, an “in-context” mode to stabilize the analysis of long texts, preservation of probabilistic distributions, configurable smoothed arcs, and procedures providing a *minima* explainability for classification choices. Our article situates these design choices within the state of the art and proposes an agenda for validation and extension.

Mots-clés: python, analyse de sentiments, détection d'émotions, BERT, études littéraires computationnelles

Keywords: python, sentiment analysis, emotion detection, BERT, computational literary studies

1 Introduction

L'ambition de quantifier l'évolution émotionnelle d'un récit — ce que le formalisme russe nomme le *syuzhet* — est devenue l'un des chantiers les plus visibles de la « lecture distante » (*distant reading*). L'objectif est notamment de vérifier empiriquement l'hypothèse formulée par l'écrivain Kurt Vonnegut, selon laquelle les histoires suivent des courbes émotionnelles simples et universelles, telles que « l'homme dans un trou » (*Man in a Hole*) ou « Cendrillon » [23].

Cependant, la traduction de cette intuition littéraire en algorithmes soulève des questions méthodologiques fondamentales : que mesure-t-on exactement ? à quelle échelle du texte ? avec quelle comparabilité entre œuvres ? et avec quel niveau d'interprétabilité pour permettre le retour au texte ? Cet article procède en trois temps : (1) un état de l'art centré sur les usages de l'analyse des émotions en humanités numériques ; (2) une discussion des choix méthodologiques qu'implique l'opérationnalisation de ces usages, à partir du cas canonique de *syuzhet* [5] ; (3) la présentation de `fabula`, un package Python reposant sur des modèles Transformer (plutôt que sur des lexiques fixes), et des perspectives de validation et d'extension.

2 Usages et choix méthodologiques : de la fouille de corpus aux arcs narratifs

La littérature récente mobilise l'analyse des émotions et du sentiment comme un instrument de description, de comparaison et d'exploration des récits. De récentes synthèses proposent des typologies convergentes de ces usages [7]. L'intérêt de ce champ est qu'il articule des objectifs interprétatifs (décrire, comparer, naviguer, inférer) à des choix techniques très concrets (segmentation,

représentation du signal, modèles, lissage, évaluation). En pratique, ces choix deviennent d'autant plus visibles que l'on cherche à produire des sorties comparables (entre œuvres, genres, périodes) et suffisamment interprétables pour permettre un retour au texte.

Un premier ensemble de travaux s'inscrit dans une logique de fouille de corpus et de navigation dans des collections : il s'agit d'indexer des textes à partir d'affects dominants, de repérer des régularités comme les « fins heureuses » [25], ou de construire des profils émotionnels rendant des œuvres comparables à grande échelle [6]. Cette orientation prolonge aussi des discussions plus anciennes sur l'usage de méthodes de classification et de regroupement pour l'étude littéraire [24]. À ce niveau, une première décision méthodologique est centrale : l'unité d'analyse. Indexer un roman « globalement » ou indexer des segments (chapitres, paragraphes, phrases) ne produit pas la même information, ni les mêmes possibilités de navigation. Le choix de segmentation conditionne aussi la sensibilité aux effets de style, aux scènes dialoguées, ou aux ruptures narratives.

Ces usages de description se prolongent vers une approche plus formelle des dynamiques narratives : plutôt que d'assigner une tonalité globale à un roman, l'enjeu devient de suivre l'évolution du signal affectif et d'identifier des trajectoires récurrentes, comme l'ont illustré des analyses à grande échelle des arcs émotionnels [20]. Dans la même veine, plusieurs travaux ont examiné le lien entre développement émotionnel et genres [10] ou discuté l'existence de trajectoires « prototypiques » selon les sous-genres [11]. L'arc narratif met alors au premier plan un second choix méthodologique : comment transformer une série de scores en une forme comparable. Deux opérations reviennent fréquemment : (1) le ré-échantillonnage (ramener des textes de longueurs différentes à une grille commune), et (2) le lissage (réduire le bruit local pour privilégier la tendance). Ces opérations répondent à un objectif comparatif et de lisibilité, mais elles introduisent des paramètres (taille de fenêtre, intensité du lissage, méthode de filtrage) qui doivent être rendus explicites pour éviter les glissements d'interprétation.

Dans ce contexte, certaines chaînes de traitement se sont imposées comme des solutions canoniques parce qu'elles standardisent ces décisions et les rendent facilement reproductibles. Le package R *syuzhet* [5] joue ici un rôle important : il propose un pipeline « texte → série → arc » et popularise l'idée d'une visualisation stable des dynamiques affectives. Ses choix techniques typiques — extraction lexicale à partir de dictionnaires (dont le *NRC Word-Emotion Association Lexicon* [16]) et lissage du signal (initialement via FFT, puis alternatives) — ont contribué à diffuser une pratique comparatiste clé-en-main. Mais cette standardisation rend également visibles les points sensibles du protocole : d'une part, une controverse mathématique a porté sur l'usage de la FFT et les artefacts de bord induits sur un signal fini (phénomène de Gibbs) [22], ce à quoi ont répondu des variantes (moyennes mobiles, LOESS, DCT) [4]; d'autre part, l'approche par dictionnaire reste structurellement limitée par sa faible prise en compte du contexte, notamment face à la polysémie, à la négation et à l'ironie, difficultés classiques des méthodes *bag-of-words* [19]. Autrement dit, les débats autour de *syuzhet* ne sont pas seulement « un cas » : ils cristallisent des questions générales sur ce que mesure un score affectif, à quelle échelle, et avec quelle robustesse.

Une partie de la recherche a transposé ces méthodes à des corpus diachroniques ou historiques, où les contraintes se renforcent : variation linguistique, décalage sémantique, bruit d'OCR et nécessité de protocoles d'évaluation adaptés [21]. Dans ces contextes, la comparabilité des signaux affectifs devient une hypothèse à tester plutôt qu'un acquis, comme le montrent des travaux sur la fiction à long terme [17]. Cela souligne un troisième choix méthodologique : l'alignement entre objectif et procédure d'évaluation. Un pipeline utile pour la navigation exploratoire peut être insuffisant pour des inférences diachroniques, où l'on attend une stabilité sémantique et des contrôles plus stricts.

Enfin, un autre ensemble de travaux se structure autour d'analyses centrées sur les personnages : l'objectif est d'attribuer des affects à des entités et de décrire des relations émotionnelles dirigées (qui ressent quoi envers qui). Des approches pionnières ont exploré l'extraction de réseaux

affectifs dans des corpus dramatiques [18]; d'autres ont proposé de cartographier des relations émotionnelles au fil d'un récit [3]. Plus récemment, l'annotation de rôles sémantiques de l'émotion a contribué à stabiliser ce type d'analyse [9], et des modèles ont été entraînés pour classifier explicitement des relations émotionnelles entre personnages [8]. Ici, la question méthodologique dominante est celle de l'attribution (à quel personnage associer l'émotion) et du retour au texte : pour être utile en narratologie, le signal doit rester traçable, segmentable et justifiable.

À ces familles s'ajoutent des usages transversaux où le signal affectif sert de variable secondaire dans des analyses critiques et sociologiques : il peut éclairer des phénomènes macroscopiques, comme des hiérarchies de genre dans la fiction contemporaine [13] ou des effets de catégorisation associés aux « livres de femmes » [12]. Dans des travaux plus appliqués, l'émotion peut aussi servir au filtrage, à la recherche ciblée, ou à l'étude de configurations discursives spécifiques [1; 14; 15; 26].

Ces usages convergent vers un constat : les approches lexicales ont eu le mérite de stabiliser des pipelines reproductibles, mais elles atteignent rapidement une limite sémantique — la difficulté à représenter le contexte — dès que l'on veut interpréter finement des segments ou comparer des œuvres hétérogènes. L'émergence des modèles de langue pré-entraînés basés sur l'architecture Transformer, tels que BERT [2], modifie précisément ce point : au lieu d'associer un mot à une valence fixe, le modèle produit des représentations contextualisées où le sens dépend du voisinage syntaxique et sémantique. Pour les humanités numériques, l'enjeu n'est pas de remplacer un outil par un autre, mais de conserver des sorties utiles aux usages établis (navigation, arcs, comparaisons) tout en rendant explicites les décisions qui conditionnent la robustesse et l'interprétabilité du signal [7].

3 Le package `fabula` : fonctionnalités déduites des usages et fonctionnement actuel

Le package `fabula` vise à décrire l'évolution des émotions ou du sentiment tout au long d'un texte narratif. Son choix central est de conserver le geste opératoire qui a fait le succès de `syuzhet` (texte → série → arc), tout en remplaçant l'approche lexicale par des modèles de langue contextuels. Concrètement, le pipeline se veut minimal et transparent : découper le texte en segments, attribuer une distribution de scores à chaque segment à l'aide d'un modèle de langue, puis transformer cette série de scores en un « arc » lissé comparable d'un texte à l'autre. `fabula` propose deux modes d'analyse — sentiment (positif vs négatif) et émotion (joie, tristesse, etc.) — afin de s'adapter aux traditions de recherche déjà établies (valence globale ou répartition d'affects).

3.1 Segmentation : choix du niveau d'observation

Le découpage du texte est décisif : il détermine la granularité de lecture, et conditionne la plupart des usages recensés (exploration de collections, arcs et genres, textes longs, personnages). `fabula` fournit quatre stratégies. Par défaut, les documents sont segmentés par phrases. Ils peuvent également être traités paragraphe par paragraphe, ou en suivant des fenêtres de tokens glissantes (algorithme roulant).

3.2 Le mode « in-context » : réinjecter du contexte dans les textes longs

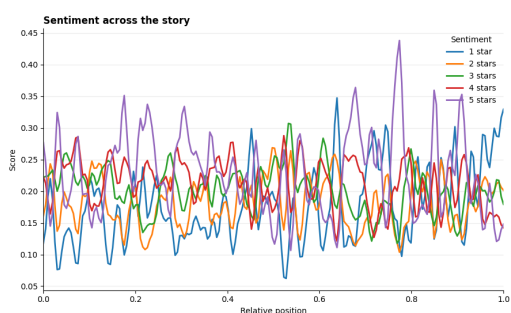
Les Transformers offrent un avantage net par rapport aux approches lexicales, mais ils restent contraints par une limite de longueur d'entrée : ils ne « lisent » qu'un nombre fini de tokens. Cette contrainte peut nuire à la compréhension contextuelle d'un passage, surtout dans les textes longs. En l'absence de modèles francophones de type Longformer entraînés spécifiquement pour l'émotion ou le sentiment, `fabula` propose un mode alternatif à évaluer.

L'intuition est la suivante : si une phrase isolée diverge fortement du ton général d'un passage, il est possible qu'elle doive être interprétée à la lumière du contexte (ironie, contraste, rupture narrative). Autrement dit, une phrase très négative au milieu d'un passage globalement positif ne devrait pas nécessairement inverser l'évaluation globale.

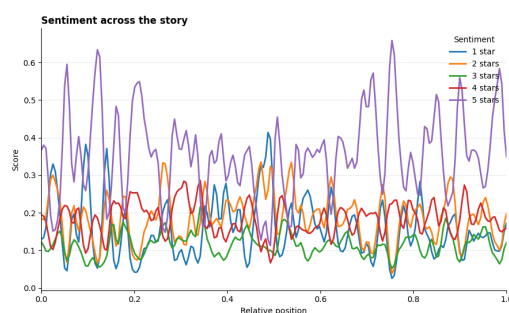
Le mode « in-context » combine donc deux niveaux de lecture. Le texte est d'abord découpé en phrases (segments fins), puis en grands blocs de tokens (segments larges). Chaque niveau est scoré séparément par le modèle. Ensuite, pour chaque phrase, *fabuLa* calcule une moyenne pondérée des scores des blocs voisins : les blocs proches comptent davantage que les blocs lointains, afin d'injecter du contexte global sans effacer la dynamique locale. Enfin, les scores « phrase » et « bloc » sont fusionnés via un poids réglable (par défaut 0,3 pour les blocs). Ce mécanisme vise à conserver la finesse de l'analyse locale tout en stabilisant l'interprétation globale des textes longs.

3.3 Scoring probabiliste : conserver l'incertitude

Au lieu d'un score binaire, chaque segment reçoit une distribution de probabilités sur les étiquettes du modèle (sentiments ou émotions). Cette distribution complète est conservée, ce qui permet des analyses plus fines que la simple classe gagnante (par exemple, suivre la compétition entre « tristesse » et « peur »). Pour les arcs « scalarisés », *fabuLa* calcule ensuite un score synthétique : une valence (positif – négatif) en mode sentiment, ou la probabilité maximale en mode émotion. Ce choix maintient un lien direct avec la sortie brute du modèle, tout en restant interprétable pour un public non spécialiste.



(a) Exemple de répartitions des résultats de valence à travers *Le rouge et le Noir* de Stendhal. Ré-échantillonnage à 200 points.



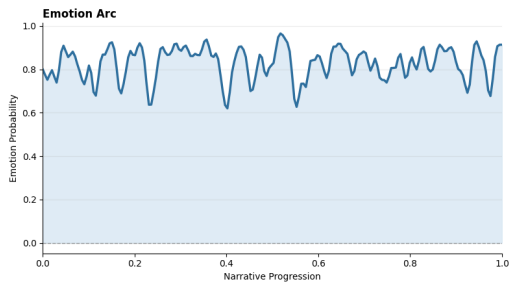
(b) Exemple de répartitions des résultats de valence à travers *Dans les Forêts de Sibérie* de Sylvain Tesson. Ré-échantillonnage à 200 points.

3.4 Arc narratif lissé : comparabilité et lisibilité

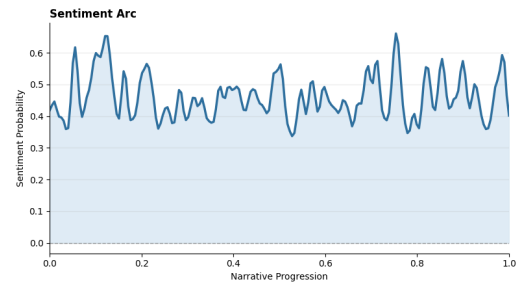
Une narration n'est pas un signal régulier ; *fabuLa* transforme donc la série irrégulière de scores en un nombre fixe de points (ré-échantillonnage) et applique un lissage configurable (moyenne mobile, gaussien ou aucun lissage). Ce choix est justifié par un objectif comparatif : rendre des textes de longueurs différentes lisibles sur une même échelle, tout en limitant le bruit local. L'utilisateur peut garder la série brute ou ajuster la quantité de lissage selon son protocole analytique.

3.5 Interprétabilité : expliquer quels mots influencent la décision

Pour répondre à la critique fréquente des « boîtes noires », *fabuLa* inclut une procédure d'explicabilité simple : pour un segment donné, il mesure l'impact de chaque token en le supprimant puis en recalculant la probabilité de la classe cible. Le résultat est une liste de tokens accompagnés d'une variation de probabilité (delta), ce qui identifie les mots qui contribuent le plus à la décision du



(a) Exemple d'un arc narratif produit à partir des scores d'émotion sur *Le Rouge et le Noir* de Stendhal. Ré-échantillonnage à 200 points, lissage gaussien.



(b) Exemple d'un arc narratif produit à partir des scores de sentiment sur *Dans les Forêts de Sibérie* de Sylvain Tesson. Ré-échantillonnage à 200 points, lissage gaussien.

modèle. L'option est accessible via l'API (et exposée côté CLI), ce qui permet de l'utiliser comme outil d'analyse qualitative, ou comme justification méthodologique dans un protocole DH. Cette méthode a l'avantage d'être directement compréhensible pour un lecteur non spécialiste : « retirer tel mot fait baisser la confiance du modèle ». Nous soulignons toutefois qu'il s'agit d'une heuristique exploratoire : la suppression d'un token (souvent un sous-mot) peut modifier l'encodage global et ne constitue pas, en soi, une attribution causale stable. Ces scores doivent être interprétés qualitativement et triangulés avec d'autres indices.

3.6 Accès et formats : reproductibilité pratique

Enfin, *fabu1a* est utilisable en ligne de commande ou via une API Python. Les sorties sont en CSV ou JSON : soit un tableau de scores par segment (*score*), soit un arc lissé (*arc*). Les modèles par défaut sont francophones, mais tout modèle Hugging Face de classification séquentielle peut être fourni, ce qui rend l'outil adaptable à d'autres langues ou jeux d'étiquettes, sans modifier le code de base.

4 Perspectives

Cette première version de *fabu1a* propose une première étape pour remplir les usages que nous avons pu identifier. Mais plusieurs chantiers restent ouverts.

1. *Entraîner de nouveaux modèles* mieux adaptés à la détection d'émotions et de sentiments pour le français, pour différentes époques. L'offre est pour l'instant trop restreinte et pas toujours adaptée.
2. *Évaluer le mode « in-context »* : comparaison systématique avec le mode standard, analyses de sensibilité (poids des blocs, taille des fenêtres) et critères d'évaluation alignés avec les objectifs (robustesse des arcs, cohérence interprétative).
3. *Traiter structurellement la contrainte de longueur* : exploration de modèles alternatifs, en particulier des architectures à long contexte, et formalisation de recommandations pratiques (tailles de segments, fenêtres, lissage) pour les textes longs après expérimentations sur des corpus littéraires.
4. *Étendre à d'autres langues* : inclure la gestion par défaut de différents modèles d'autres langues adaptées au contexte littéraire.

Disponibilité des données

Le code source du package `fabula` et sa documentation sont disponibles en ligne : <https://github.com/floriancafiero/fabula>. Le package est disponible sur pypi : <https://pypi.org/project/fabula-fr>.

Références

- [1] CAFIERO, Florian et PUREN, Marie. « A Riddle in a Haystack : LLM Detection of Intricate Wordplays in Colette and Willy's Novels for Authorship Attribution ». In : *Digital Humanities 2025*. 2025.
- [2] DEVLIN, Jacob, CHANG, Ming-Wei, LEE, Kenton et TOUTANOVA, Kristina. « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 2019.
- [3] JHAVAR, Harshita et MIRZA, Paramita. « EMOFIEL : Mapping Emotions of Relationships in a Story ». In : *Companion Proceedings of The Web Conference 2018*. 2018, p. 243-246. DOI : 10.1145/3184558.3186989.
- [4] JOCKERS, Matthew L. « Requiem for a Low Pass Filter ». Matthew Jockers Blog. 2015. URL : <http://www.matthewjockers.net/2015/04/06/requiem/>.
- [5] JOCKERS, Matthew L. « Revealing Sentiment and Plot Arcs with the Syuzhet Package ». Matthew Jockers Blog. 2015. URL : <http://www.matthewjockers.net/2015/02/02/syuzhet/>.
- [6] KAKKONEN, TUOMO et GALIC KAKKONEN, Gordana. « Sentiprofiler : Creating Comparable Visual Profiles of Sentimental Content in Texts ». In : *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*. 2011, p. 62-69.
- [7] KIM, Evgeny et KLINGER, Roman. « A Survey on Sentiment and Emotion Analysis for Computational Literary Studies ». In : *Zeitschrift für digitale Geisteswissenschaften* (2021). DOI : 10.17175/2019_008_v2.
- [8] KIM, Evgeny et KLINGER, Roman. « Frowning Frodo, Wincing Leia, and a Seriously Great Friendship : Learning to Classify Emotional Relationships of Fictional Characters ». In : *Proceedings of NAACL-HLT 2019, Volume 1 : Long and Short Papers*. 2019, p. 647-653. DOI : 10.18653/v1/N19-1067.
- [9] KIM, Evgeny et KLINGER, Roman. « Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions ». In : *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*. 2018, p. 1345-1359.
- [10] KIM, Evgeny, PADÓ, Sebastian et KLINGER, Roman. « Investigating the Relationship between Literary Genres and Emotional Plot Development ». In : *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 2017, p. 17-26. DOI : 10.18653/v1/W17-2203.
- [11] KIM, Evgeny, PADÓ, Sebastian et KLINGER, Roman. « Prototypical Emotion Developments in Adventures, Romances, and Mystery Stories ». In : *Digital Humanities 2017 : Conference Abstracts*. 2017, p. 288-291.
- [12] KOOLEN, Corina. « Women's Books versus Books by Women ». In : *Digital Humanities 2018 : Book of Abstracts*. 2018.

- [13] KRAICER, Eve et PIPER, Andrew. « Social Characters : The Hierarchy of Gender in Contemporary English-Language Fiction ». In : *Journal of Cultural Analytics* (2019). DOI : 10 . 22148/16 . 032.
- [14] LIONNET-ROLLIN, Alexandre et CAFIERO, Florian. « Never Care for What They Say? Platform vs. Genre Rules in Online Horror Narratives (2007–2024) ». In : *Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities (NLP4DH 2026)*. Association for Computational Linguistics, 2026.
- [15] LIONNET-ROLLIN, Alexandre et CAFIERO, Florian. « What Makes a Creepypasta Stick? On the Recipe for Canonicity of Online Literary Productions ». hal-05406444. 2025.
- [16] MOHAMMAD, Saif M. et TURNEY, Peter D. « Crowdsourcing a Word–Emotion Association Lexicon ». In : *Computational Intelligence* 29, no. 3 (2013), p. 436-465.
- [17] MORIN, Olivier et ACERBI, Alberto. « Birth of the Cool : A Two-Century Decline in Emotional Expression in Anglophone Fiction ». In : *Cognition and Emotion* (2017).
- [18] NALISNICK, Eric T. et BAIRD, Henry S. « Extracting Sentiment Networks from Shakespeare’s Plays ». In : *2013 12th International Conference on Document Analysis and Recognition*. 2013, p. 758-762.
- [19] PANG, Bo et LEE, Lillian. « Opinion Mining and Sentiment Analysis ». In : *Foundations and Trends in Information Retrieval* 2, no. 1–2 (2008), p. 1-135. DOI : 10 . 1561/1500000011.
- [20] REAGAN, Andrew J., MITCHELL, Lewis, KILEY, Dilan, DANFORTH, Christopher M. et DODDS, Peter Sheridan. « The Emotional Arcs of Stories are Dominated by Six Basic Shapes ». In : *EPJ Data Science* 5, no. 1 (2016), p. 31. DOI : 10 . 1140/epjds/s13688-016-0093-1.
- [21] SPRUGNOLI, Rachele, TONELLI, Sara, MARCHETTI, Alessandro et MORETTI, Giovanni. « Towards Sentiment Analysis for Historical Texts ». In : *Digital Scholarship in the Humanities* 31, no. 4 (2016), p. 762-772. DOI : 10 . 1093/11c/fqv027.
- [22] SWAFFORD, Annie. « Why Syuzhet Breaks : A Critical Assessment of Jockers’ Sentiment Analysis Package ». 2015. URL : <https://annieswafford.wordpress.com/2015/03/02/why-syuzhet-breaks/>.
- [23] VONNEGUT, Kurt. *A Man Without a Country*. New York : Seven Stories Press, 2005.
- [24] YU, Bei. « An Evaluation of Text Classification Methods for Literary Study ». In : *Literary and Linguistic Computing* 23, no. 3 (2008), p. 327-343. DOI : 10 . 1093/11c/fqn015.
- [25] ZEHE, Albin, BECKER, Martin, HETTINGER, Lena, HOTHO, Andreas, REGER, Isabella et JANNIDIS, Fotis. « Prediction of Happy Endings in German Novels Based on Sentiment Information ». In : *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing 2016*. 2016, p. 9-16.
- [26] ZRIBI, Yaelle, CAFIERO, Florian, LÉPINAY, Vincent et VIDAL-GORÈNE, Chahan. « Timing in Stand-up Comedy : Text, Audio, Laughter, Kinesics (TIC-TALK) : Pipeline and Database for the Multimodal Study of Comedic Timing ». In : *Proceedings of the 2nd Workshop on Computational Humor (CHum 2026)*. Association for Computational Linguistics, 2026.