Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Le rôle des valeurs aberrantes dans l'anticipation de nouvelles thématiques : approche par représentations vectorielles sur un corpus d'actualité

Evangelia Zve[1, 2] Benjamin Icard[1] Alice Breton[1] Lila Sainero[1]
Gauvain Bourgne[1] Jean-Gabriel Ganascia[1]

(1) LIP6, Sorbonne Université, CNRS, 4 Place Jussieu, 75005 Paris, France
(2) Infopro Digital, 20 Rue des Aqueducs, 94250 Gentilly, France

{prénom.nom}@lip6.fr

RÉSUMÉ

Cet article vise à analyser le rôle des données aberrantes (*outliers*), souvent assimilées à du bruit en modélisation thématique, en tant que signaux faibles de l'émergence de nouveaux thèmes dans des corpus d'actualités dynamiques. À partir de représentations vectorielles (*embeddings*) produites par différents modèles de langue à l'état de l'art, et d'une procédure de regroupement cumulatif (*clustering*), nous suivons leur évolution au fil du temps dans deux corpus médiatiques et institutionnelles en français et en anglais, centrés sur la responsabilité sociale des entreprises et le changement climatique. Les résultats mettent en évidence une régularité : au fil du temps, les données aberrantes tendent à se structurer en thèmes cohérents, et ce de manière robuste, indépendamment du modèle et de la langue considérés.

ABSTRACT

From Outliers to Topics in Language Models : Anticipating Trends in News Corpora

This paper examines how outliers, often dismissed as noise in topic modeling, can act as weak signals of emerging topics in dynamic news corpora. Using vector embeddings from state-of-the-art language models and a cumulative clustering approach, we track their evolution over time in French and English news datasets focused on corporate social responsibility and climate change. The results reveal a consistent pattern : outliers tend to evolve into coherent topics over time across both models and languages.

MOTS-CLÉS : Traitement automatique du langage, Modélisation thématique dynamique, Corpus d'actualité, Données aberrantes, Thématiques émergentes, Regroupement cumulatif, BERTopic, HDBSCAN, Conversion des outliers en thématiques.

KEYWORDS: Natural Language Processing, Dynamic Topic Modeling, News Corpora, Outliers, Emerging Topics, Cumulative Clustering, BERTopic, HDBSCAN, Outlier-to-Topic Conversion.

ARTICLE ACCEPTÉ À : ICNLSP 2025 : The 8th International Conference on Natural Language and Speech Processing, Southern Denmark University, Odense, Denmark, August 25-27, 2025. Obtained Best Paper Award.
URL : https://aclanthology.org/2025.icnlsp-1.38/

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 29–29.