IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, June 2023

Exploring Latent Themes-Analysis of Various Topic Modelling Algorithms

Reetesh Kumar Srivastava, Shalini Sharma, Dr. Piyush Pratap Singh
School of Computer and Systems Sciences
Jawaharlal Nehru University, New Delhi, India

Abstract: *This research explores the effectiveness of four common topic modelling methods for identifying latent themes and topics in unstructured text data: Latent Dirich- let Allocation(LDA), Non-Negative Matrix Factorization(NMF), Top2Vec, and BERTopic. Topic modelling is an essential method for gaining insights from massive amounts of textual data. Top2Vec and BERTopic are recent approaches that use unsupervised neural networks to develop distributed representations of texts and words, whereas NMF and LDA are traditional techniques frequently utilised for topic modelling. This document gives a timeline of important advances in topic modelling, including the development of NMF and LDA, as well as many refinements and additions to LDA. According to the study's findings, BERTopic surpasses the other approaches, particularly in recognising overlapping and fine- grained subjects. This work emphasises the significance of text processing quality, the variety of subjects in the text, and the right selection of topic modelling methods in efficiently breaking down topics.*

Keywords: LDA, NMF, Top2Vec, BERTopic

REFERENCES

[1] Egger, R. & Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology 7 (2022). [Online; accessed 2023-05-06].
[2] Rodriguez-Garcia, P., Li, Y., Lopez-Lopez, D. & Juan, A. A. Strategic decision making in smart home ecosystems: A review on the use of artificial intelligence and internet of things. Internet of Things 100772 (2023).
[3] Lyu, Y. Dockerized knowledge-oriented multi-modal social event detection system, 1–6 (IEEE, 2022).
[4] Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999). [Online; accessed 2023-05-09].
[5] Lee, D. & Seung, H. S. Leen, T., Dietterich, T. & Tresp, V. (eds) Algorithms for non-negative matrix factorization. (eds Leen, T., Dietterich, T. & Tresp, V.) Advances in Neural Information Processing Systems, Vol. 13 (MIT Press, 2000). URL https://t.ly/q 17.
[6] Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).
[7] Hoffman, M. D., Blei, D. M. & Bach, F. Online learning for latent dirichlet allocation, NIPS'10, 856–864 (Curran Associates Inc., Red Hook, NY, USA, 2010).
[8] Blei, D. M. & McAuliffe, J. D. Supervised topic models, NIPS'07, 121–128 (Curran Associates Inc., Red Hook, NY, USA, 2007).
[9] Blei, D. M. & Lafferty, J. D. Dynamic topic models, ICML '06, 113–120 (Asso- ciation for Computing Machinery, New York, NY, USA, 2006). URL https://doi.org/10.1145/1143844.1143859.
[10] Angelov, D. Top2vec: Distributed representations of topics (2020). 2008.09470.
[11] Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure (2022). 2203.05794.
[12] Vijayarani, S., Ilamathi, M. J., Nithya, M. et al. Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks 5, 7–16 (2015).
[13] Rahimi, Z. & Homayounpour, M. M. The impact of preprocessing on word embedding quality: A comparative study. Language Resources and Evaluation 57, 257–291 (2023).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, June 2023

[14] Qader, W., M. Ameen, M. & Ahmed, B. An overview of bag of words;importance, implementation, applications, and challenges, 200–204 (2019).

[15] Aizawa, A. An information-theoretic perspective of tf–idf measures. Information Processing Management 39, 45–65 (2003). URL https://www.sciencedirect.com/ science/article/pii/S0306457302000213.

[16] Gefen, D. et al. Identifying patterns in medical records through latent semantic analysis. Communications of the ACM 61, 72–77 (2018).

[17] Mardones-Segovia, C., Wheeler, J. M., Choi, H.-J., Wang, S. & Cohen, A. S. Model selection for latent dirichlet allocation in assessment data. Psychological Test and Assessment Modeling 65, 3–35 (2023).