7.0 REFERENCE

[1] N. Y. Saiyad, H. B. Prajapati and V. K. Dabhi, "A survey of document clustering using semantic approach," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 2555-2562, doi: 10.1109/ICEEOT.2016.7755154.

[2] Facebookresearch, "facebookresearch/XLM: PyTorch original implementation of Cross-lingual Language Model Pretraining.," GitHub. [Online]. Available: https://github.com/facebookresearch/XLM. [Accessed: 13-Sep-2021].

[3] M. F. Fayaza and S. Ranathunga, "Tamil News Clustering Using Word Embeddings," 2020 Moratuwa Engineering Research Conference (MERCon), 2020.

[4] P. Nanayakkara and S. Ranathunga, "Clustering Sinhala News Articles Using Corpus-Based Similarity Measures," 2018 Moratuwa Engineering Research Conference (MERCon), 2018.

[5] "A Fast and Powerful Scraping and Web Crawling Framework," Scrapy. [Online]. Available: https://scrapy.org/. [Accessed: 13-Sep-2021].

[6] Sparck Jones Karen, " A statistical interpretation of term specificity and its application in retrieval," Document Retrieval Systems, pp. 132–142, 1988.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, en J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality", in Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Lake Tahoe, Nevada, 2013, bll 3111–3119.

[8] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. Mcclosky, "The Stanford CoreNLP Natural Language Processing Toolkit," Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014.

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.

[10] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, en T. Mikolov, "Learning Word Vectors for 157 Languages", arXiv [cs.CL]. 2018.

[11] O. Melamud, J. Goldberger, en I. Dagan, "context2vec: Learning Generic Context Embedding with Bidirectional LSTM", in Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016, bll 51–61.

[12] B. McCann, J. Bradbury, C. Xiong, en R. Socher, "Learned in Translation: Contextualized Word Vectors", in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017, bll 6297–6308.

[13] M. E. Peters et al., "Deep Contextualized Word Representations", in Proceedings of the 2018 Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, bll 2227–2237.
[14] J. Devlin, M.-W. Chang, K. Lee, en K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.
[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, en I. Sutskever, “Language Models are Unsupervised Multitask Learners”, 2018.
[16] Google-Research, “google-research/bert: TensorFlow code and pre-trained models for BERT,” GitHub. [Online]. Available: https://github.com/google-research/bert. [Accessed: 17-Sep-2021].
[17] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, arXiv [cs.LG]. 2020.
[18] M. Müller, M. Salathé, en P. E. Kummervold, “COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter”, arXiv [cs.CL]. 2020.
[19] U. Naseem, I. Razzak, S. K. Khan, en M. Prasad, “A Comprehensive Survey on Word Representation Models: From Classical to State-Of-The-Art Word Representation Language Models”, arXiv [cs.CL]. 2020.
[20] S. Doddapaneni, G. Ramesh, A. Kunchukuttan, P. Kumar, en M. M. Khapra, “A Primer on Pretrained Multilingual Language Models”, arXiv [cs.CL]. 2021.
[21] H. W. Chung, T. Févry, H. Tsai, M. Johnson, en S. Ruder, “Rethinking embedding coupling in pre-trained language models”, arXiv [cs.CL]. 2020.
[22] X. Ouyang et al., “ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora”, arXiv [cs.CL]. 2021.
[23] XTREME. [Online]. Available: https://sites.research.google/xtreme. [Accessed: 25-Sep-2021].
[24] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, en M. Johnson, “XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization”, arXiv [cs.CL]. 2020.
[25] T. Pires, E. Schlinger, en D. Garrette, “How multilingual is Multilingual BERT?”, arXiv [cs.CL]. 2019.
[26] G. Lample en A. Conneau, “Cross-lingual Language Model Pretraining”, arXiv [cs.CL]. 2019.
[27] A. Conneau et al., “Unsupervised Cross-lingual Representation Learning at Scale”, arXiv [cs.CL]. 2020.
[28] Z. Chi et al., “XLM-E: Cross-lingual Language Model Pre-training via ELECTRA”, arXiv [cs.CL]. 2021.
[29] Google-Research, “bert/multilingual.md at master · google-research/bert,” GitHub, 18-Oct-2019. [Online]. Available: https://github.com/google-research/bert/blob/master/multilingual.md. [Accessed: 13-Sep-2021].

[30] M. Marcińczuk, M. Gniewkowski, T. Walkowiak, en M. Będkowski, “Text Document Clustering: Wordnet vs. TF-IDF vs. Word Embeddings”, in Proceedings of the 11th Global Wordnet Conference, 2021, bll 207–214.
[31] J. Park, C. Park, J. Kim, M. Cho, en S. Park, “ADC: Advanced document clustering using contextualized representations”, Expert Systems with Applications, vol 137, bll 157–166, 2019.
[32] L. Stankevičius en M. Lukoševičius, “Testing pre-trained Transformer models for Lithuanian news clustering”, arXiv [cs.IR]. 2020.
[33] N. Reimers en I. Gurevych, “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 11 2020.
[34] “Multilingual-Models¶,” Multilingual-Models - Sentence-Transformers documentation. [Online]. Available: https://www.sbert.net/examples/training/multilingual/README.html#available-pre-trained-models. [Accessed: 29-Sep-2021].
[35] C. Vu, “A complete guide to transfer learning from English to other Languages using Sentence Embeddings...,” Medium, 22-May-2020. [Online]. Available: https://towardsdatascience.com/a-complete-guide-to-transfer-learning-from-english-to-other-languages-using-sentence-embeddings-8c427f8804a9. [Accessed: 29-Sep-2021].
[36] “The open parallel corpus,” OPUS. [Online]. Available: https://opus.nlpl.eu/. [Accessed: 29-Sep-2021].
[37] S. Dutta, “‘Alignment is All You Need’: Analyzing Cross-Lingual Document Similarity for Domain-Specific Applications”, in CLEOPATRA@WWW, 2021.
[38] M. Artetxe en H. Schwenk, “Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond”, arXiv [cs.CL]. 2019.
[39] I. Gusev en I. Smurov, “Russian News Clustering and Headline Selection Shared Task”, arXiv [cs.CL]. 2021.
[40] V. A. S and S. E. Y, “Russian News Similarity Detection with SBERT: pre-training and fine-tuning,” Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021,” Jun. 2021.
[41] M. Linger en M. Hajaiej, “Batch Clustering for Multilingual News Streaming”, arXiv [cs.CL]. 2020.
[42] S. Miranda, A. Znotiņš, S. B. Cohen, en G. Barzdins, “Multilingual Clustering of Streaming News”, arXiv [cs.CL]. 2018.
[43] G. Leban, B. Fortuna and M. Grobelnik, "Using News Articles for Real-time Cross-Lingual Event Detection and Filtering," in Proc. of the NewsIR’16 Workshop at ECIR, Padua, Italy, 2016. pp. 33-38

[44] S. Wu, A. Conneau, H. Li, L. Zettlemoyer, en V. Stoyanov, “Emerging Cross-lingual Structure in Pretrained Language Models”, arXiv [cs.CL]. 2020.
[45] S. Wu en M. Dredze, “Do Explicit Alignments Robustly Improve Multilingual Encoders?”, arXiv [cs.CL]. 2020.
[46] V. Vithulan, "multilingual-doc-clustering/template.json at main · VIthulan/multilingual-doc-clustering", GitHub, 2022. [Online]. Available: https://github.com/VIthulan/multilingual-doc-clustering/blob/main/src/template.json. [Accessed: 29- Mar- 2022].
[47] V. Vithulan, "multilingual-doc-clustering/src/data/valid_records at main · VIthulan/multilingual-doc-clustering", GitHub, 2022. [Online]. Available: https://github.com/VIthulan/multilingual-doc-clustering/tree/main/src/data/valid_records. [Accessed: 29- Mar- 2022].
[48] "Interest in Different Types of News - Digital News Report 2013", Reuters Institute Digital News Report, 2022. [Online]. Available: https://www.digitalnewsreport.org/survey/2013/interest-in-different-types-of-news-2013/. [Accessed: 29- Mar- 2022].
[49] "What are the different types of news?", Quora, 2022. [Online]. Available: https://www.quora.com/What-are-the-different-types-of-news. [Accessed: 29- Mar- 2022].
[50] V. Vithulan, "multilingual-doc-clustering/documents_annotated_v2_1.csv at main · VIthulan/multilingual-doc-clustering", GitHub, 2022. [Online]. Available: https://github.com/VIthulan/multilingual-doc-clustering/blob/main/src/processed_data/reviewed/documents_annotated_v2_1.csv. [Accessed: 29- Mar- 2022].
[51] V. Vithulan, "multilingual-doc-clustering/data_validator.py at main · VIthulan/multilingual-doc-clustering", GitHub, 2022. [Online]. Available: https://github.com/VIthulan/multilingual-doc-clustering/blob/main/src/data_validator.py. [Accessed: 29- Mar- 2022].
[52] V. Vithulan, "multilingual-doc-clustering/review_iaa_calc_export_team_1.xlsx at main · VIthulan/multilingual-doc-clustering", GitHub, 2022. [Online]. Available: https://github.com/VIthulan/multilingual-doc-clustering/blob/main/src/IAA/review_iaa_calc_export_team_1.xlsx. [Accessed: 29- Mar- 2022].
[53] V. Vithulan, "multilingual-doc-clustering/review_iaa_calc_export_team_2.xlsx at main · VIthulan/multilingual-doc-clustering", GitHub, 2022. [Online]. Available: https://github.com/VIthulan/multilingual-doc-clustering/blob/main/src/IAA/review_iaa_calc_export_team_2.xlsx. [Accessed: 29- Mar- 2022].
[54] V. Vithulan, "multilingual-doc-clustering/Inter-Annotater-Agreement.ipynb at main · VIthulan/multilingual-doc-clustering", GitHub, 2022. [Online]. Available: https://github.com/VIthulan/multilingual-doc-clustering/blob/main/src/IAA/Inter-Annotater-Agreement.ipynb. [Accessed: 29- Mar- 2022].

[55] V. Vithulan, "Google Colaboratory", Colab.research.google.com, 2022. [Online]. Available: https://colab.research.google.com/drive/1qoGMjyiFmuC-ehWluOaxePKBOWV0FDex?usp=sharing. [Accessed: 29- Mar- 2022].
[56] V. Vithulan, "multilingual-doc-clustering/tf-idf-baseline.ipynb at main · VIthulan/multilingual-doc-clustering", GitHub, 2022. [Online]. Available: https://github.com/VIthulan/multilingual-doc-clustering/blob/main/src/Baseline/tf-idf-baseline.ipynb. [Accessed: 29- Mar- 2022].
[57] M. Lewis κ.ά., 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension'. arXiv, 2019.
[58] "Evaluation of clustering", Nlp.stanford.edu. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html. [Accessed: 05- Jul- 2022].