Bibliography

[1] F. Abate, "The american heritage dictionary of the english language, ed. by joseph pickett and steve kleinedler," *Dictionaries: Journal of the Dictionary Society of North America*, vol. 34, no. 1, pp. 235–237, 2013.

[2] I. Abu El-Khair, "Effects of stop words elimination for arabic information retrieval: A comparative study," *International Journal of Computing Information Sciences*, vol. 4, pp. 119–133, 01 2006.

[3] M. Adda-Decker, "A corpus-based decompounding algorithm for german lexical modeling in lvcsr," in *Eighth European Conference on Speech Communication and Technology*. Citeseer, 2003.

[4] A. Ajees and G. Graham, "A hybrid approach for suffix separation in malayalam," in *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*. IEEE, 2018, pp. 1–4.

[5] M. Akasereh and J. Savoy, "Ad hoc retrieval with marathi language," in *Multilingual Information Access in South Asian Languages: Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*. Springer, 2013, pp. 23–37.

[6] B. Al-Shargabi, F. Olayah, and W. A. Romimah, "An experimental study for the effect of stop words elimination for arabic text classification algorithms," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 6, no. 2, pp. 68–75, 2011.

[7] A. Alajmi, E. Saad, and R. Darwish, "Toward an arabic stop-words list generation," *International Journal of Computer Applications*, vol. 46, no. 8, pp. 8–13, 2012.

[8] E. Alfonseca, S. Bilac, and S. Pharies, "Decompounding query keywords from compounding languages," in *Proceedings of ACL-08: HLT, Short Papers*, 2008, pp. 253–256.

[9] R. Alkula, "From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software," *Information Retrieval*, vol. 4, no. 3-4, pp. 195–208, 2001.

[10] G. Amati and C. J. Van Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 357–389, 2002.

[11] R. Aralikatte, N. Gantayat, N. Panwar, A. Sankaran, and S. Mani, "Sanskrit sandhi splitting using seq2(seq)2," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4909–4914. [Online]. Available: https://aclanthology.org/D18-1530

[12] T. V. Asubiaro, "Entropy-based generic stopwords list for yoruba texts," *International Journal of Computer and Information Technology*, vol. 2, no. 5, 2013.

[13] H. Ayral and S. Yavuz, "An automated domain specific stop word generation method for natural language text classification," in *2011 International Symposium on Innovations in Intelligent Systems and Applications*. IEEE, 2011, pp. 500–503.

[14] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[15] T. Bell, I. H. Witten, and J. G. Cleary, "Modeling for text compression," *ACM Computing Surveys (CSUR)*, vol. 21, no. 4, pp. 557–591, 1989.

[16] D. C. Blair, "Information retrieval, 2nd ed. c.j. van rijsbergen. london: Butterworths; 1979: 208 pp. price: $32.50," *Journal of the American Society for Information Science*, vol. 30, no. 6, pp. 374–375, 1979.

[17] M. Braschler and B. Ripplinger, "How effective is stemming and decompounding for german text retrieval?" *Information Retrieval*, vol. 7, no. 3-4, pp. 291–316, 2004.

[18] T. Brychcín and M. Konopík, "Hps: High precision stemmer," *Information Processing & Management*, vol. 51, no. 1, pp. 68–91, 2015.

[19] E. Cambria and B. White, "Jumping nlp curves: On natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[20] X. Chen, X. Qiu, C. Zhu, P. Liu, and X.-J. Huang, "Long short-term memory neural networks for chinese word segmentation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1197–1206.

[21] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] C. Courseault Trumbach and D. Payne, "Identifying synonymous concepts in preparation for technology mining," *Journal of Information Science*, vol. 33, no. 6, pp. 660–677, 2007.

[24] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[25] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[26] R. Dabre, A. Amberkar, and P. Bhattacharyya, "Morphological analyzer for affix stacking languages: A case study of marathi," in *Proceedings of COLING 2012: Posters*, 2012, pp. 225–234.

[27] J. Daiber, L. Quiroz, R. Wechsler, and S. Frank, "Splitting compounds by semantic analogy," in *Proceedings of the 1st Deep Machine Translation Workshop*. Praha, Czechia: ÚFAL MFF UK, 2015, pp. 20–28. [Online]. Available: https://aclanthology.org/W15-5703

[28] D. Das, K. Radhika, R. Rajeev, and R. Raj, "Hybrid sandhi-splitter for malayalam using unicode," in *In proceedings of National Seminar on Relevance of Malayalam in Information Technology*, 2012.

[29] P. Das and A. Das, "Bengali noun morphological analyzer," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2013, pp. 1538–1543.

[30] M. R. Davarpanah, M. Sanji, and M. Aramideh, "Farsi lexical analysis and stop word list," *Library Hi Tech*, vol. 27, no. 3, pp. 435–449, 2009.

[31] S. Dave, A. K. Singh, D. P. AP, and P. B. Lall, "Neural compound-word (sandhi) generation and splitting in sanskrit language," in *8th ACM IKDD CODS and 26th COMAD*, 2021, pp. 171–177.

[32] S. Deepa, K. Bali, A. G. Ramakrishnan, and P. Talukdar, "Automatic generation of compound word lexicon for hindi speech synthesis," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2004.

[33] R. D. Deshmukh and A. Kiwelekar, "Deep learning techniques for part of speech tagging by natural language processing," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 2020, pp. 76–81.

[34] V. Devadath, L. J. Kurisinkel, D. M. Sharma, and V. Varma, "A sandhi splitter for malayalam," in *Proceedings of the 11th International Conference on Natural Language Processing*, 2014, pp. 156–161.

[35] M. Diab, "Second generation amira tools for arabic processing: Fast and robust tokenization, pos tagging, and base phrase chunking," in *2nd International Conference on Arabic Language Resources and Tools*, vol. 110, 2009, p. 198.

[36] L. Dolamic and J. Savoy, "Unine at fire 2008: Hindi, bengali, and marathi ir," in *Working notes of the forum for information retrieval evaluation*. Citeseer, 2008, pp. 12–14.

[37] ——, "Comparative study of indexing and search strategies for the hindi, marathi, and bengali languages," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 9, no. 3, p. 11, 2010.

[38] V. J. Easton and J. H. McColl, "Statistics glossary v1. 1," 1997.

[39] N. Erbs, P. B. Santos, T. Zesch, and I. Gurevych, "Counting what counts: Decompounding for keyphrase extraction," in *Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction*, 2015, pp. 10–17.

[40] C. Fautsch and J. Savoy, "Algorithmic stemmers or morphological analysis? an evaluation," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 8, pp. 1616–1624, 2009.

[41] C. Fox, "A stop list for general text," in *ACM SIGIR Forum*, vol. 24, no. 1-2. ACM, 1989, pp. 19–21.

[42] W. N. Francis, H. Kucera, H. Kučera, and A. W. Mackie, *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin, 1982.

[43] D. Ganguly, J. Leveling, and G. J. Jones, "A case study in decompounding for bengali information retrieval," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2013, pp. 108–119.

[44] K. Ghosh and A. Bhattacharya, "Stopword removal: Why bother? a case study on verbose queries," in *Proceedings of the 10th Annual ACM India Compute Conference*, 2017, pp. 99–102.

[45] S. I. Hajeer, R. M. Ismail, N. L. Badr, and M. F. Tolba, "A new stemming algorithm for efficient information retrieval systems and web search engines," in *Multimedia Forensics and Security*. Springer, 2017, pp. 117–135.

[46] D. Harman, "How effective is suffixing?" *Journal of the american society for information science*, vol. 42, no. 1, pp. 7–15, 1991.

[47] C. Haruechaiyasak, S. Kongyoung, and M. Dailey, "A comparative study on thai word segmentation approaches," in *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, vol. 1. IEEE, 2008, pp. 125–128.

[48] B. He and I. Ounis, "A study of parameter tuning for term frequency normalization," in *Proceedings of the twelfth international conference on Information and knowledge management*, 2003, pp. 10–16.

[49] O. Hellwig, "Using recurrent neural networks for joint compound splitting and sandhi resolution in sanskrit," in *4th Biennial Workshop on Less-Resourced Languages*, 2015.

[50] O. Hellwig and S. Nehrdich, "Sanskrit word segmentation using character-level recurrent and convolutional neural networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2754–2763.

[51] D. Hiemstra, *Using language models for information retrieval*. Univ. Twente, 2001.

[52] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[54] V. Hollink, J. Kamps, C. Monz, and M. De Rijke, "Monolingual document retrieval for european languages," *Information retrieval*, vol. 7, no. 1-2, pp. 33–52, 2004.

[55] D. A. Hull, "Stemming algorithms: A case study for detailed evaluation," *Journal of the American Society for Information Science*, vol. 47, no. 1, pp. 70–84, 1996.

[56] M. IŞIK and H. DAĞ, "The impact of text preprocessing on the prediction of review ratings," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 28, no. 3, pp. 1405–1421, 2020.

[57] K. Jasleen and R. S. Jatinderkumar, "Pos word class based categorization of gurmukhi language stemmed stop words," in *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2*. Springer, 2016, pp. 3–10.

[58] R. Jayashree, K. S. Murthy, and B. S. Anami, "Effect of stop word removal on the performance of naïve bayesian methods for text classification in the kannada language," *International Journal of Artificial Intelligence and Soft Computing*, vol. 4, no. 2-3, pp. 264–282, 2014.

[59] G. N. Jha, M. Agrawal, S. K. Mishra, D. Mani, D. Mishra, M. Bhadra, S. K. Singh *et al.*, "Inflectional morphology analyzer for sanskrit," in *Sanskrit computational linguistics*. Springer, 2007, pp. 219–238.

[60] V. Jha, N. Manjunath, P. D. Shenoy, and K. Venugopal, "Hsra: Hindi stopword removal algorithm," in *2016 international conference on microelectronics, computing and communications (MicroCom)*. IEEE, 2016, pp. 1–5.

[61] M. F. Kabir, K. Abdullah-Al-Mamun, and M. N. Huda, "Deep learning based parts of speech tagger for bengali," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. IEEE, 2016, pp. 26–29.

[62] K. Kettunen and E. Airio, "Is a morphologically complex language really that complex in full-text retrieval?" in *International Conference on Natural Language Processing (in Finland)*. Springer, 2006, pp. 411–422.

[63] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[64] Y. Kitagawa and M. Komachi, "Long short-term memory for Japanese word segmentation," in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Hong Kong: Association for Computational Linguistics, 1–3 Dec. 2018. [Online]. Available: https://aclanthology.org/Y18-1033

[65] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary: Association for Computational Linguistics, Apr. 2003.

[66] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004, pp. 625–633.

[67] L. Kovriguina, I. Shilin, A. Shipilo, and A. Putintseva, "Russian tagging and dependency parsing models for stanford corenlp natural language toolkit," in *International Conference on Knowledge Engineering and the Semantic Web*. Springer, 2017, pp. 101–111.

[68] A. Krishna, P. Satuluri, and P. Goyal, "A dataset for sanskrit word segmentation," in *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2017, pp. 105–114.

[69] R. Krovetz, "Viewing morphology as an inference process," *Artificial intelligence*, vol. 118, no. 1-2, pp. 277–294, 2000.

[70] H. Kucera, H. Kučera, and W. N. Francis, *Computational analysis of present-day American English*. Brown university press, 1967.

[71] D. Kumar and P. Rana, "Design and development of a stemmer for punjabi," *International Journal of Computer Applications*, vol. 11, no. 12, pp. 18–23, 2010.

[72] A. T. Kwee, F. S. Tsai, and W. Tang, "Sentence-level novelty detection in english and malay," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2009, pp. 40–51.

[73] D. J. Ladani and N. P. Desai, "Stopword identification and removal techniques on tc and ir applications: A survey," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2020, pp. 466–472.

[74] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[75] T. Laureys, V. Vandeghinste, and J. Duchateau, "A hybrid approach to compounds in lvcsr," in *Seventh International Conference on Spoken Language Processing*, 2002.

[76] J. Leveling, W. Magdy, and G. J. Jones, "An investigation of decompounding for cross-language patent search," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 1169–1170.

[77] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.

[78] R. T.-W. Lo, B. He, and I. Ounis, "Automatically building a stopword list for an information retrieval system," in *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, vol. 5, 2005, pp. 17–24.

[79] J. B. Lovins, "Development of a stemming algorithm," *Mech. Transl. Comput. Linguistics*, vol. 11, no. 1-2, pp. 22–31, 1968.

[80] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of research and development*, vol. 1, no. 4, pp. 309–317, 1957.

[81] M. M. Majgaonker, "Discovering suffixes: A case study for marathi language," 2010.

[82] P. Majumder, M. Mitra, S. K. Parui, G. Kole, P. Mitra, and K. Datta, "Yass: Yet another suffix stripper," *ACM transactions on information systems (TOIS)*, vol. 25, no. 4, pp. 18–es, 2007.

[83] M. Makrehchi and M. S. Kamel, "Automatic extraction of domain-specific stopwords from labeled documents," in *European Conference on Information Retrieval*. Springer, 2008, pp. 222–233.

[84] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press, 2008.

[85] J. Mayfield and P. McNamee, "Single n-gram stemming," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 415–416.

[86] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.

[87] U. Mishra and C. Prakash, "Maulik: an effective stemmer for hindi language," *International Journal on Computer Science and Engineering*, vol. 4, no. 5, p. 711, 2012.

[88] K. Mohnot, N. Bansal, S. P. Singh, and A. Kumar, "Hybrid approach for part of speech tagger for hindi language," *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, vol. 4, no. 1, pp. 25–30, 2014.

[89] C. Monz and M. d. Rijke, "Shallow morphological analysis in monolingual information retrieval for dutch, german, and italian," in *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 2001, pp. 262–277.

[90] J. Nair, S. S. Nair, and U. Abhishek, "Sanskrit stemmer design: A literature perspective," in *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 3*. Springer, 2022, pp. 117–128.

[91] R. L. Ott and M. T. Longnecker, *An introduction to statistical methods and data analysis*. Cengage Learning, 2015.

[92] J. H. Paik, M. Mitra, S. K. Parui, and K. Järvelin, "Gras: An effective and efficient stemming algorithm for information retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 29, no. 4, pp. 1–24, 2011.

[93] J. H. Paik, D. Pal, and S. K. Parui, "A novel corpus-based stemming algorithm using co-occurrence statistics," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 863–872.

[94] J. H. Paik and S. K. Parui, "A fast corpus-based stemmer," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 10, no. 2, pp. 1–16, 2011.

[95] B. P. Pande, P. Tamta, and H. S. Dhami, "Generation, implementation, and appraisal of an n-gram-based stemming algorithm," *Digital Scholarship in the Humanities*, vol. 34, no. 3, pp. 558–568, 2019.

[96] P. Patel, K. Popat, and P. Bhattacharyya, "Hybrid stemmer for gujarati," in *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, 2010, pp. 51–55.

[97] T. Pellegrini and L. Lamel, "Automatic word decompounding for asr in a morphologically rich language: Application to amharic," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 5, pp. 863–873, 2009.

[98] M. F. Porter *et al.*, "An algorithm for suffix stripping." *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[99] U. Prajitha, C. Sreejith, and P. R. Raj, "Lalitha: A light weight malayalam stemmer using suffix stripping method," in *2013 International Conference on Control Communication and Computing (ICCC)*. IEEE, 2013, pp. 244–248.

[100] B. Premjith, K. Soman, and P. Poornachandran, "A deep learning based part-of-speech (pos) tagger for sanskrit language by embedding character level features." in *FIRE*, 2018, pp. 56–60.

[101] A. Priyadarshi and S. K. Saha, "Towards the first maithili part of speech tagger: Resource creation and system development," *Computer Speech & Language*, vol. 62, p. 101054, 2020.

[102] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[103] M. M. Rahman, M. Kutlu, T. Elsayed, and M. Lease, "Efficient test collection construction via active learning," in *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, 2020, pp. 177–184.

[104] R. M. Rakholia and J. R. Saini, "Lexical classes based stop words categorization for gujarati language," in *2016 2nd International Conference on Advances in Computing, Communication, Automation (ICACCA) (Fall)*, 2016, pp. 1–5.

[105] A. Ramanathan and D. D. Rao, "A lightweight stemmer for hindi," in *the Proceedings of EACL*, 2003.

[106] J. Raulji, J. Saini *et al.*, "Morphological analyzer for sanskrit language," 2019.

[107] J. K. Raulji and J. R. Saini, "Generating stopword list for sanskrit language," in *2017 IEEE 7th international advance computing conference (IACC)*. IEEE, 2017, pp. 799–802.

[108] V. Reddy, A. Krishna, V. Sharma, P. Gupta, V. M R, and P. Goyal, "Building a word segmenter for Sanskrit overnight," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1264

[109] S. Robertson, "On the history of evaluation in ir," *Journal of Information Science*, vol. 34, no. 4, pp. 439–456, 2008.

[110] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information science*, vol. 27, no. 3, pp. 129–146, 1976.

[111] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter, "Probabilistic models of indexing and searching." in *SIGIR*, vol. 80, 1980, pp. 35–56.

[112] M. Sadeghi and J. Vegas, "Automatic identification of light stop words for persian information retrieval systems," *Journal of information science*, vol. 40, no. 4, pp. 476–487, 2014.

[113] N. Saharia, U. Sharma, and J. Kalita, "Analysis and evaluation of stemming algorithms: a case study with assamese," in *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, 2012, pp. 842–846.

[114] S. S. Sahu and P. Mamgain, "A corpus-based decompounding in sanskrit." in *FDIA@ ESSIR*, 2019, pp. 110–114.

[115] S. S. Sahu and S. Pal, "Effect of stopwords in indian language ir," *Sādhanā*, vol. 47, no. 1, pp. 1–17, 2022.

[116] J. R. Saini and R. M. Rakholia, "On continent and script-wise divisions-based statistical measures for stop-words lists of international languages," *Procedia Computer Science*, vol. 89, pp. 313–319, 2016.

[117] R. P. Sandell, "Productivity in historical linguistics: Computational perspectives on word-formation in ancient greek and sanskrit," Ph.D. dissertation, UCLA, 2015.

[118] S. Sarica and J. Luo, "Stopwords in technical language processing," *PloS one*, vol. 16, no. 8, p. e0254937, 2021.

[119] J. Savoy, "A stemming procedure and stopword list for general french corpora," *Journal of the American Society for Information Science*, vol. 50, no. 10, pp. 944–952, 1999.

[120] J. Savoy, L. Dolamic, and M. Akasereh, "Information retrieval with hindi, bengali, and marathi languages: evaluation and analysis," in *Multilingual Information Access in South Asian Languages: Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*. Springer, 2013, pp. 334–352.

[121] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

[122] J. P. Shaffer, "Multiple hypothesis testing," *Annual review of psychology*, vol. 46, no. 1, pp. 561–584, 1995.

[123] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[124] Y. Shao, C. Hardmeier, J. Tiedemann, and J. Nivre, "Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 173–183. [Online]. Available: https://aclanthology.org/I17-1018

[125] M. R. Shree, S. Lakshmi, and B. Shambhavi, "A novel approach to sandhi splitting at character level for kannada language," in *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*. IEEE, 2016, pp. 17–20.

[126] S. Siddiqi and A. Sharan, "Construction of a generic stopwords list for hindi language without corpus statistics," *International Journal of Advanced Computer Research*, vol. 8, no. 34, pp. 35–40, 2018.

[127] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, vol. 3. IEEE, 2003, pp. 1661–1666.

[128] R. J. Simes, "An improved bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 73, no. 3, pp. 751–754, 1986.

[129] D. Singh, S. Bhingardive, and P. Bhattacharyya, "Multiword expressions dataset for indian languages," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 2331–2335.

[130] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '96. New York, NY, USA: Association for Computing Machinery, 1996, p. 21–29. [Online]. Available: https://doi.org/10.1145/243199.243206

[131] M. P. Sinka and D. W. Corne, "Towards modernised and web-specific stoplists for web document analysis," in *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*. IEEE, 2003, pp. 396–402.

[132] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[133] R. W. Sproat and M. V. Wilkes, *Morphology and computation*. MIT press, 1992.

[134] M. Straka, J. Straková, and J. Hajič, "Czech text processing with contextual embeddings: Pos tagging, lemmatization, parsing and ner," in *International Conference on Text, Speech, and Dialogue*. Springer, 2019, pp. 137–150.

[135] P. Strazny, *Encyclopedia of linguistics*. Routledge, 2013.

[136] K. Suba, D. Jiandani, and P. Bhattacharyya, "Hybrid inflectional stemmer and rule-based derivational stemmer for gujarati," in *Proceedings of the 2nd workshop on South Southeast Asian natural language processing (WSSANLP)*, 2011, pp. 1–8.

[137] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[138] P. Tamta and B. P. Pande, "Simultaneous removal of prefix and suffix," *Vietnam Journal of Computer Science*, vol. 7, no. 02, pp. 129–144, 2020.

[139] A. Tolmachev, D. Kawahara, and S. Kurohashi, "Design and structure of the juman++ morphological analyzer toolkit," *Journal of Natural Language Processing*, vol. 27, no. 1, pp. 89–132, 2020.

[140] S. Tomlinson, "Lexical and algorithmic stemming compared for 9 european languages with hummingbird searchserver tm at clef 2003," in *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 2003, pp. 286–300.

[141] R. ul Haque, P. Mehera, M. Mridha, and M. A. Hamid, "A complete bengali stop word detection mechanism," in *2019 Joint 8th international conference on informatics, electronics & vision (ICIEV) and 2019 3rd international conference on imaging, vision & pattern recognition (icIVPR)*. IEEE, 2019, pp. 103–107.

[142] J. Xu and W. B. Croft, "Corpus-based stemming using cooccurrence of word variants," *ACM Transactions on Information Systems (TOIS)*, vol. 16, no. 1, pp. 61–81, 1998.

[143] N. Xue, "Chinese word segmentation as character tagging," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, 2003, pp. 29–48.

[144] M.-A. Yaghoub-Zadeh-Fard, B. Minaei-Bidgoli, S. Rahmani, and S. Shahrivari, "Pswg: An automatic stop-word list generator for persian information retrieval systems based on similarity function & pos information," in *2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)*. IEEE, 2015, pp. 111–117.

[145] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley, 1949.

[146] F. Zou, F. L. Wang, X. Deng, and S. Han, "Automatic identification of chinese stop words," *Research on Computing Science*, vol. 18, pp. 151–162, 2006.

Appendix

Table 1: List of suffixes stemmed by light stemmer

◌ः	◌ौ	◌ाः	म्
◌ै	न्	◌ान्	◌ानि
◌ं	◌ेन	◌ौः	◌ाय
यः	◌ोण	◌ा	◌ो
◌ेन	◌ेणा	णी	वे
◌ोः	ने	स्य	◌िन
◌ैः	◌ाणि	ना	णा
◌ेब	◌ात्	◌ेणु	◌ेही
◌ेषु	याः	◌ेण	योः
◌िभः	◌ानि	णां	यै
यां	भीः	◌िये	◌ियौ
◌ें भ्यः			

Table 2: Additionally a list of suffixes stemmed by aggressive stemmer

याम्	नाम्	णाम्	◌ानां
वान्	◌ेभ्यम्	◌ाभ्याम्	◌ियाम्
◌ानाम्	रम्	षाम्	व्यम्
यितुं	मेव	◌ियों	

Table 3: List of prefixes used in different Indian languages IR

अप	अभि	अति
ऩा	अक	मुख्य
दुर	परा	उच्च
निर	महा	पूर्व
उप	परि	प्रधान
प्रति	त्रि	प्रमुख

List of Publications

Journal Papers

- **Siba Sankar Sahu** and Sukomal Pal, Building a text retrieval system for the Sanskrit language: Exploring indexing, stemming, and searching issues, *Computer Speech and Language*, Elsevier (**SCI, IF-4.3**)
- **Siba Sankar Sahu** and Sukomal Pal, Effect of stopwords in Indian language IR, *Sādhanā*, 47 (1), 2022, Springer (**SCI, IF-1.6**)
- **Siba Sankar Sahu** and Sukomal Pal, A study on corpus-based stopword list in Indian language IR, *ACM Transactions on Asian and Low-Resource Language Information Processing* (**SCI, IF-2**)
- **Siba Sankar Sahu** and Sukomal Pal, A case study on decompounding in Inadian language IR, *Natural language processing* (**SCI, IF-2.5**)

- **Siba Sankar Sahu**, Debrup Dutta, Sukomal Pal, and Imran Rasheed, Effect of stopwords and stemming techniques in Urdu IR, *SN Computer Science* (Scopus)

Conference Papers

- **Siba Sankar Sahu** and Puneet Mamgain, A Corpus-based Decompounding in Sanskrit, *Symposium on Future Directions in Information Access*, Milan (Italy), July 2019