In the future, we aim at studying more rules to capture the cases which the SimString algorithm fails and deploying a sensitive keyword detection system to use in practice with a large number of keywords. We also aim to use this research results in other problems, such as automatic discovery of variations of brand keywords, spell checking for keywords in search engine.

References

[1] A. Bhonde, S. Agarwal, "Multi-layer stacking with diverse supervised models to determine quality of product titles," in *International Conference on Information and Knowledge Management*, *AnalytiCup*, 2017.

[2] M. Nicosia, A. Moschitti, "Lazada Product Title Quality Challenge: Constructing Features for a Diversified Ensemble of Classifiers," in *International Conference on Information and Knowledge Management*, *AnalytiCup*, 2017.

[3] K.Singh, V.Sunder, "Lazada Product Title Quality Challenge: An Ensemble of Deep and Shallow Learning to predict the Quality of Product Titles," in *International Conference on Information and Knowledge Management*, *AnalytiCup*, 2017.

[4] Y. Zhang, M. Zhu, D. Wang, and S.Feng, "Logo Detection and Recognition Based on Classification, International Conference on Web-Age Information Management," in *2014 WAIM Conference on Web-Age Information Management*, 2014, pp. 805-816

[5] T. Mudumbi, N. Bian, Y. Zhang and F. Hazoume, "An Approach Combined the Faster RCNN and Mobilenet for Logo Detection," *Journal of Physics Conference Series*, 2019

[6] V. Murugan, V.R. vijaykumar, and A. Nidhila, "Vehicle Logo Recognition using RCNN for Intelligent Transportation Systems," in *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET),* 2019, pp. 107-111

[7] M. Bozbura, H. Tunc, M. Kusak, C. Sakar, "Detection of e-Commerce Anomalies using LSTM-recurrent Neural Networks," in *Proceedings of the 8th International Conference on Data Science, Technology and Applications*, 2019, pp. 217-224

[8] Z. Yang, S. Cao, and B. Yan, "Using linear discrim-inant analysis and data mining approaches to identifyE-commerce anomaly," in *Proceedings of 2011 7th Inter-national Conference on Natural Computation*, 2011, pp. 2406–2410.

[9] A. R. Yelundur, S. H. Sengamedu, and B. Mishra, "E-commerce Anomaly Detection: A Bayesian Semi-Supervised Tensor Decomposition Approach usingNatural Gradients," *arXiv e-prints. arXiv:1804.03836*, 2018

[10] L. Zheng, G. Liu, C. Yan; C. Jiang, "Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity," *IEEE Transactions on Computational Social Systems*, vol. 5, 2018, pp. 796 - 806

[11] S. Cao, X. Yang, C. Chen, J. Zhou, X. Li, and Y. Qi, "TitAnt: Online Real-time Transaction Fraud Detection in Ant Financial," in *Proceedings of the VLDB Endowment*, 12(12), 2019, pp. 2082 - 2093

[12] Z. Zhang, X. Zhou, X. Zhang, L. Wang, and P. Wang, "A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection," *Security and Communication Networks*, 2018

[13] J.Y. Jiang, Y.Y. Ke, P.Y. Chien, "Learning user reformulation behavior for query auto-completion," in: *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval*, 2014, pp. 445–454

[14] L. Ramachandran, U. Murthy, "Ghosting: contextualized query auto-completion on Amazon search," in: *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 1377–1378

[15] Jose G. Camargo de Souza, M. Kozielski, P. Mathur, E. Chang, M. Guerini, M. Negri, M. Turchi, and E. Matusov, "Generating E-Commerce Product Titles and Predicting their Quality," in *Proceedings of the 11th International Conference on Natural Language Generation*, Association for Computational Linguistics, 2018, pp. 233–243

[16] M. R. Mane, S. Kedia, A. Mantha, S. Guo, and K. Achan, "Product Title Generation for Conversational Systems using BERT," *arXiv preprint arXiv:2007.11768*, 2020

[17] A. Hartveld, M. V. Keulen, D. Mathol, T.V. Noort, T. Plaatsman, F. FrasincarEmail, and K. Schouten, "An LSH-Based Model-Words-Driven Product Duplicate Detection Method," in *International Conference on Advanced Information Systems Engineering*, 2018, pp. 409-423

[18] B. West, K.A. Jadda, U. Ahsan, H. Qu, and X. Cui, "Interpretable Methods for Identifying Product Variants," in *WWW '20: Companion Proceedings of the Web Conference*, 2020 pp. 448–453

[19] T. Zhu, Y. Wang, H. Li, Y. Wu, X. He, and B. Zhou, "Multimodal Joint Attribute Prediction and Value Extraction for E-commerce Product," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2020, pp. 2129–2139

[20] Henzinger, Monika, "Finding near-duplicate web pages: a large-scale evaluation of algorithms," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 284– 291

[21] G.S. Manku, G. Singh, A. Jain, and A.D. Sarma, "Detecting near-duplicates for web crawling," in *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, 2007, pp. 141–150

[22] S. Kurtz, J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich, "REPuter: the manifold applications of repeat analysis on a genomic scale," *Nucleic Acids Research*, 29(22): 4633–4642, 2001

[23] G. M. Landau, J. P. Schmidt, and D. Sokol. "An algorithm for approximate tandem repeats," *Journal of Computational Biology*, 8(1):1–18, 2001.

[24] G. S. Manku, A. Jain, and A. Das Sarma, "Detecting near-duplicates for web crawling," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 141–150

[25] N. Okazaki, and J. Tsujii, "Simple and efficient algorithm for approximate dictionary matching," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 851–859

[26] A. Cislak, Sz. Grabowski, "A practical index for approximate dictionary matching with few mismatches", in *COMPUTING AND INFORMATICS*, VOL 36, NO 5, 2017, pp. 1088-1106

[27] Yoshua Bengio, R. Ducharme, Pascal Vincent, Christian Janvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, 3(6):932-938, 2003

[28] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013

[29] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov ,"Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguist*, 2017, 5, pp. 135–146.

[30] M.E. Peters, M. Neumann, M. Iyyer, and M. Gardner, "Deep contextualized word representations," in *Proceedings of the NAACL-HLT*, 2018, pp. 2227–2237

[31] José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, Rafael Valencia-García, "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings," *Future Generation Computer Systems*, vol. 114, 2021, pp. 506-518

[32] D.T. Do, T.Q.T. Le, N.Q.K. Le, "Using deep neural networks and biological subwords to detect protein S-sulfenylation sites," *Briefings in Bioinformatics*, 2020