6. Conclusion and Future Work

Text classification is one of the major applications of machine learning. The proposed method uses text mining algorithms for webpage classification. The Porter and WordNet dictionary are used to calculate the semantic distances between keywords, i.e., words with the greatest similarity. Then, documents are classified based on the extracted keywords using machine learning algorithms, i.e., decision tree, K-NN, SVM, and Naïve Bayes. The performance analysis of supervised machine learning algorithms for text classification shows that the decision tree algorithm gave better results based on prediction accuracy when compared to the other three algorithms. In future work, we planning to use the best algorithm from this study to classify positive webpages for information extraction and create a datatahub portal.

7. References

[1] S. A. Ozel and E. Sarac, "Focused crawler for finding professional events based on user interests," in *2008 23rd International Symposium on Computer and Information Sciences*, 2008, pp. 1–4.

[2] A. Abbasi, F. ``Mariam'' Zahedi, and S. Kaza, "Detecting Fake Medical Web Sites Using Recursive Trust Labeling," *ACM Trans. Inf. Syst.*, vol. 30, no. 4, p. 22:1--22:36, Nov. 2012.

[3] H. B. Kazemian and S. Ahmed, "Comparisons of machine learning techniques for detecting malicious webpages," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1166–1177, 2015.

[4] E. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," *WSEAS Trans. Comput.*, vol. 4, pp. 966–974, 2005.

[5] Menaka S. and Radha N., "Text Classification using Keyword Extraction Technique," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 12, pp. 734–740, 2013.

[6] G. A. Miller, "WordNet: A Lexical Database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[7] F. R. Lucini *et al.*, "Text mining approach to predict hospital admissions using early medical records from the emergency department," *Int. J. Med. Inform.*, vol. 100, no. Supplement C, pp. 1–8, 2017.

[8] D. Johnson, F. J. Oles, T. Zhang, and T. Goetz, "A Decision-Tree-Based Symbolic Rule Induction System for Text Categorization," *IBM Syst. J.*, vol. 41, pp. 428–437, 2002.

[9] H.-S. Lim, "Improving kNN Based Text Classification with Well Estimated Parameters," vol. 3316. pp. 516–523, 2004.

[10] J. Shanahan and N. Roma, "Improving SVM Text Classification Performance through Threshold Adjustment," vol. 2837. pp. 361–372, 2003.

[11] S.-B. Kim, H.-C. Rim, D. Yook, and H. Lim, "Effective Methods for Improving Naive Bayes Text Classifiers," in *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, 2002, pp. 414–423.

[12] K.-M. Schneider, "Techniques for Improving the Performance of Naive Bayes for Text Classification," in *In Proceedings of CICLing 2005*, 2005, pp. 682–693.

[13] M. A. Kłopotek, "Very large Bayesian multinets for text classification," *Futur. Gener. Comput. Syst.*, vol. 21, no. 7, pp. 1068–1082, 2005.

[14] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier Science, 2011.