	Decision Tree	**89,29%**	**89,28%**	87,80%	88,10%

Berdasarkan hasil perbandingan pada tabel 8, akurasi tertinggi terdapat pada penelitian sebelumnya yang memiliki akurasi sebesar 99,40% dengan model bahasa Naïve Bayes dan *feature extraction* N-Gram, adapun juga model bahasa SVM dengan akurasi 98,80% dengan *feature extraction* TF-IDF yang dicapai oleh penelitian sebelumnya. Untuk *Random Forest* dan *Decision Tree* pengujian yang dilakukan memiliki akurasi tertinggi dibanding penelitian sebelumnya sebesar 97,21% untuk *Random Forest* dan 90,36% untuk *Decision Tree* dengan *feature extraction* N-Gram. Didapatkan hasil pebandingan sebagai berikut dimana penelitian yang dilakukan Nugraha dan Romadhony menggunakan jumlah *data training* dan *data test* yang lebih besar dibandingkan pengujian yang dilakukan pada tugas akhir ini sehingga memiliki akurasi yang lebih tinggi dari seluruh hasil pengujian yang dilakukan.

5. Kesimpulan

Identifikasi bahasa daerah dengan menggunakan metode Multinomial Naïve Bayes dapat menghasilkan akurasi yang baik untuk menggunakan jumlah n-gram yang singkat yakni 1-gram (unigram) hingga 2-gram (bigram) dan akurasi model bahasa meningkat dengan bertambahnya jumlah data latih yang diberikan. Rata-rata akurasi yang dihasilkan pada Naïve Bayes dan F-measure sebesar 98,86%. Adapun untuk *range* n-gram 3-gram dan 4-gram nilai akurasi yang dihasilkan hanya sebesar 49,5% dikarenakan kalimat yang digunakan umunya dicampur dengan bahasa lain sehingga saat melakukan klasifikasi kata yang asalnya bahasa asing menjadi fitur dalam suatu tersebut. Adapun jumlah data latih mempengaruhi akurasi pada model bahasa, sehingga semakin banyak data latih maka akurasi pada model bahasa semakin tinggi. Dengan hasil perbandingan yang dilakukan model yang dibangun saat ini belum dapat mencapai hasil yang lebih tinggi dari penelitian sebelumnya,

Daftar Pustaka

[1] Lui, M., Lau, J. H., & Baldwin, T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, *2*, 27-40.

[2] Ibrohim, M. O., & Budi, I. (2019, August). Multi-label hate speech and abusive language detection in Indonesian twitter. In Proceedings of the Third Workshop on Abusive Language Online (pp. 46-57)

[3] Moryossef, A., Tsochantaridis, I., Aharoni, R., Ebling, S., & Narayanan, S. (2020, August). Real-time sign language detection using human pose estimation. In European Conference on Computer Vision (pp. 237-248). Springer, Cham.

[4] Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2019). Automatic language identification in texts: A survey. Journal of Artificial Intelligence Research, 65, 675-782.

[5] Zhang, Y., Riesa, J., Gillick, D., Bakalov, A., Baldridge, J., & Weiss, D. (2018). A fast, compact, accurate model for language identification of codemixed text. *arXiv preprint arXiv:1810.04142*.

[6] Nishijima, M., & Liu, Y. (2021). Native Language Identification and Reconstruction of Native Language Relationship Using Japanese Learner Corpus. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (pp. 368-376).

[7] Wijonarko, P., & Zahra, A. (2022). Spoken language identification on 4 Indonesian local languages using deep learning. Bulletin of Electrical Engineering and Informatics, 11(6), 3288-3293

[8] Nugraha, A. B., & Romadhony, A. (2023). Identification of 10 Regional Indonesian Languages Using Machine Learning. *Sinkron: jurnal dan penelitian teknik informatika*, *8*(4), 2203-2214.

[9] Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

[10] Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 403.

[11] Baldwin, T., & Lui, M. (2010, June). Language identification: The long and the short of the matter. In Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics (pp. 229-237).

[12] Sasaki, Y. (2007). The truth of the F-measure. *Teach tutor mater*, *1*(5), 1-5.

[13] ARDIYANTI SURYANI, ARIE; Widyantoro, Dwi Hendratmo; Purwarianti, Ayu; Sudaryat, Yayat, 2022, "Sundanese-Indonesian Parallel Corpus", https://doi.org/10.34820/FK2/HDYWXW, Telkom University Dataverse, V1

[14] Sujaini, H. (2020). Improving the role of language model in statistical machine translation (Indonesian-Javanese). *International Journal of Electrical and Computer Engineering*, *10*(2), 2102.

[15] Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., ... & Ruder, S. (2022). Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. *arXiv preprint arXiv:2205.15960*.
[16] Jurafsky, Daniel & Martin, James. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Lampiran