BAB 6

KESIMPULAN DAN SARAN

6.1 Kesimpulan

Setelah melakukan proses analisis, perancangan, implementasi, dan pengujian, maka dapat diambil beberapa kesimpulan, yaitu:

1. N-gram berhasil diimplementasikan pada level karakter untuk mencari jumlah kemunculan karakter pada suatu dokumen. Digunakan n-gram hanya sampai pada trigram (tiga karakter) karena pada tahap trigram sudah memiliki tingkat akurasi yang cukup untuk klasifikasi dokumen.

2. Dalam proses klasifikasi dokumen digunakan algoritma euclidean untuk mencari jarak dokumen yang diuji dengan *cluster* bahasa yang ada. Algoritma euclidean tersebut berhasil diimplementasikan dan berhasil untuk mengklasifikasi dokumen.

3. Dari pengujian klasifikasi yang dimulai dari 50 hingga 500 dokumen training dari tiap cluster, dokumen sudah berhasil diklasifikasi namun untuk dokumen bahasa Indonesia dan bahasa Melayu masih kurang akurat karena lima n-gram teratas dari bahasa Indonesia dan bahasa Melayu mirip.

4. Dari hasil pengujian pelatihan dokumen, didapatkan bahwa semakin banyak dokumen yang diuji waktu yang dibutuhkan juga semakin lama.

6.2 Saran

Saran yang dapat diberikan pada skripsi ini sebagai berikut :

1. Agar penelitian ini dapat terus berkembang, bahasa yang diuji bisa diperluas lebih dari lima bahasa.

2. Metode n-gram dapat dikembangkan hingga pada level per kata.

3. Karakter yang digunakan yang akan diklasifikasi dapat dikembangkan hingga karakter non-ASCII

4. Penelitian dapat dikembangkan dengan tidak memakai *case folding* karena untuk beberapa bahasa memiliki perbedaan karakter jika tidak dilakukan *case folding*.

5. Penelitian ini bertujuan untuk mengklasifikasi dokumen yang belum diketahui bahasa dokumennya. Agar penelitian ini dapat terus berkembang, hasil dari penelitian ini dapat dikembangkan agar dapat dimanfaatkan dalam bidang dan tujuan lain.

DAFTAR REFERENSI

[1] Jain, A. K. (2010) Data clustering: 50 years beyond k-means. *Algorithmica*, **31**, 651–666.

[2] Jurafsky, D. dan Martin, J. H. (2000) *Speech and Language Processing*, 2rd edition. Alan Apt, New Jersey.

[3] Cavnar, W. B. dan Trenkle, J. M. (1994) N-gram-based text categorization. *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, US, 9 September, pp. 161–175. Environmental Research Institute of Michigan.

[4] Hamzah, A. (2010) Deteksi bahasa untuk dokumen teks berbahasa indonesia. Skripsi. IST AKPRIND, Yogyakarta.

[5] Gurusamy, V. dan Kannan, S. (2014) Preprocessing techniques for text mining. Technical report. Madurai Kamaraj University, Palkalai Nagar, Madurai, Tamil Nadu 625021, India.

[6] Gantz, J. F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., dan Toncheva, A. (2007) The diverse and exploding digital universe. *An Updated Forecast of Worldwide Information Growth Through 2011*, **1**, 2–10.

[7] Dictionary, M.-W. O. (2008) *"cluster analysis"*. https://www.merriam-webster.com/dictionary/cluster. 15 january 2018.

[8] Tan, P.-N., Steinbach, M., dan Kumar, V. (2005) *Introduction to Data Mining*, second edition edition. Addison Wesley, Boston.