DAFTAR ISI

KATA PENGANTAR iv
UCAPAN TERIMA KASIH iv
ABSTRAK iv
ABSTRACT v
DAFTAR ISI vi
DAFTAR TABEL viii
DAFTAR GAMBAR x
DAFTAR LAMPIRAN x
BAB I PENDAHULUAN 1
1.1 Latar Belakang Penelitian 1
1.2 Rumusan Masalah 4
1.3 Tujuan Penelitian 4
1.4 Manfaat Penelitian 5
1.5 Batasan Masalah 5
1.6 Sistematika Penulisan 5
BAB II KAJIAN PUSTAKA 7
2.1 Data Mining 7
2.2. Text Mining 9
2.2.1 Text Preprocessing 11
2.2.2 Pembobotan TF-IDF 14
2.2.3 *Splitting* Data 17
2.3 Text Classification 20
2.3.1 Cosine Similarity 24
2.3.2 K-Nearest Neighbor 28
2.4 Jurnal Penelitian Ilmu Komputer 31
2.5 Penelitian Terkait KNN *Clasification* 32
2.6 Evaluasi Hasil Klasifikasi 33
BAB III METODOLOGI PENELITIAN 36
3.1 Desain Penelitian 36
3.1.1 Alat Penelitian 39
3.1.2 Bahan Penelitian 40
3.2 Metode Penelitian 40
3.2.1 Metode Pengumpulan Data 40
3.2.2 Proses Pengembangan Perangkat Lunak 41
BAB IV HASIL DAN PEMBAHASAN 43
4.1 Pengumpulan Data 43
4.2 Perancangan Data *Preparation* 45
4.2.1 Text Preprocessing 45
4.2.2 Feature selection dengan pembobotan TF-IDF 52
4.3 Klasifikasi K-Nearest Neighbor 53
4.4 Perancangan Alur Sistem 55

4.5 Pengembangan Perangkat Lunak 56
4.5.1 Analisa Kebutuhan 56
4.5.2 Desain 57
4.5.3 Parameter Analisis 58
4.6 Implementasi *Holdout Method* 59
4.6.1 Implementasi Skenario-1 metode *Holdout Method* 60
4.6.2 Implementasi Skenario-2 metode *Holdout Method* 61
4.6.3 Implementasi Skenario-3 metode *Holdout Method* 62
4.6.4 Implementasi Skenario-4 metode *Holdout Method* 63
4.6.5 Implementasi Skenario-5 metode *Holdout Method* 65
4.7 Implementasi 10 Fold Cross Validation 66
4.7.1 Implementasi Skenario-1 metode 10 *Fold Cross Validation* 67
4.7.2 Implementasi Skenario-2 metode 10 *Fold Cross Validation* 69
4.7.3 Implementasi Skenario-3 metode 10 *Fold Cross Validation* 70
4.7.4 Implementasi Skenario-4 metode 10 *Fold Cross Validation* 71
4.7.5 Implementasi Skenario-5 metode 10 *Fold Cross Validation* 73
4.7.6 Implementasi Skenario-6 metode 10 *Fold Cross Validation* 74
4.7.7 Implementasi Skenario-7 metode 10 *Fold Cross Validation* 75
4.7.8 Implementasi Skenario-8 metode 10 *Fold Cross Validation* 77
4.7.9 Implementasi Skenario-9 metode 10 *Fold Cross Validation* 78
4.7.10 Implementasi Skenario-10 metode 10 *Fold Cross Validation* 79
4.8 Analisis Hasil Skenario 81
4.8.1 Analisis terhadap jumlah nilai k 81
4.8.2 Analisis terhadap jumlah data *training* 83
4.8.3 Analisis terhadap waktu eksekusi 85
4.9 Perbandingan dengan penelitian sebelumnya 87
BAB V KESIMPULAN DAN SARAN 92
5.1 Kesimpulan 93
5.2 Saran 94
DAFTAR PUSTAKA 95

DAFTAR TABEL

Tabel 2.1 Contoh dataset term frequency (tf) 15
Tabel 2.2 Contoh dataset document frequency (df) 16
Tabel 2.3 Contoh dataset inverse document frequency (idf) 16
Tabel 2.4 Contoh document frequency (df) dengan jumlah dokumen=N 16
Tabel 2.5 Contoh document frequency (df) dengan N=1000 17
Tabel 2.6 Contoh hasil perhitungan rumus TF-IDF 17
Tabel 2.7 Contoh ilustrasi 10 fold cross validation 19
Tabel 2.8 Contoh dataset untuk klasifikasi 22
Tabel 2.9 Contoh dataset untuk sample query 25
Tabel 2.10 Contoh tabel perhitungan TF-IDF untuk cosine similarity 26
Tabel 2.11 Contoh tabel perhitungan TF-IDF (2) 26
Tabel 2.12 Hasil perkalian scalar setiap D terhadap Q 27
Tabel 2.13 Hasil perhitungan cosine similarity 27
Tabel 2.14 Contoh dataset untuk KNN 30
Tabel 2.15 Contoh hasil klasifikasi dengan algoritma KNN 30
Tabel 2.16 Jurnal penelitian terdahulu 32
Tabel 2.17 Confusion matrix 2 kelas 33
Tabel 4.1 Skenario data dengan holdout method 44
Tabel 4.2 Skenario data dengan 10 fold cross validation 44
Tabel 4.3 Skenario data training holdout method 59
Tabel 4.4 Skenario data testing holdout method 59
Tabel 4.5 Data training holdout method skenario 1 60
Tabel 4.6 Data testing holdout method skenario 1 60
Tabel 4.7 Confusion matrix holdout method skenario 1 60
Tabel 4.8 Data training holdout method skenario 2 61
Tabel 4.9 Data testing holdout method skenario 2 61
Tabel 4.10 Confusion matrix holdout method skenario 2 61
Tabel 4.11 Data training holdout method skenario 3 62
Tabel 4.12 Data testing holdout method skenario 3 63
Tabel 4.13 Confusion matrix holdout method skenario 3 63
Tabel 4.14 Data training holdout method skenario 4 64
Tabel 4.15 Data testing holdout method skenario 4 64
Tabel 4.16 Confusion matrix holdout method skenario 4 64
Tabel 4.17 Data training holdout method skenario 5 65
Tabel 4.18 Data testing holdout method skenario 5 65
Tabel 4.19 Confusion matrix holdout method skenario 5 65
Tabel 4.20 Skenario data training 10 fold cross validation 66
Tabel 4.21 Skenario data testing 10 fold cross validation 67
Tabel 4.22 Data training 10 fold cross validation skenario 1 67
Tabel 4.23 Data testing 10 fold cross validation skenario 1 68
Tabel 4.24 Confusion matrix 10 fold cross validation skenario 1 68

Tabel 4.25 Data training 10 fold cross validation skenario 2 69
Tabel 4.26 Data testing 10 fold cross validation skenario 2 69
Tabel 4.27 Confusion matrix 10 fold cross validation skenario 2 69
Tabel 4.28Data training 10 fold cross validation skenario 3 70
Tabel 4.29 Data testing 10 fold cross validation skenario 3 70
Tabel 4.30 Confusion matrix 10 fold cross validation skenario 3 71
Tabel 4.31 Data training 10 fold cross validation skenario 4 72
Tabel 4.32 Data testing 10 fold cross validation skenario 4 72
Tabel 4.33 Confusion matrix 10 fold cross validation skenario 4 72
Tabel 4.34 Data training 10 fold cross validation skenario 5 73
Tabel 4.35 Data testing 10 fold cross validation skenario 5 73
Tabel 4.36 Confusion matrix 10 fold cross validation skenario 5 73
Tabel 4.37 Data training 10 fold cross validation skenario 6 74
Tabel 4.38 Data testing 10 fold cross validation skenario 6 74
Tabel 4.39Confusion matrix 10 fold cross validation skenario 6 75
Tabel 4.40 Data training 10 fold cross validation skenario 7 76
Tabel 4.41 Data testing 10 fold cross validation skenario 7 76
Tabel 4.42 Confusion matrix 10 fold cross validation skenario 7 76
Tabel 4.43 Data training 10 fold cross validation skenario 8 77
Tabel 4.44 Data testing 10 fold cross validation skenario 8 77
Tabel 4.45 Confusion matrix 10 fold cross validation skenario 8 77
Tabel 4.46 Data training 10 fold cross validation skenario 9 78
Tabel 4.47 Data testing 10 fold cross validation skenario 9 78
Tabel 4.48 Confusion matrix 10 fold cross validation skenario 9 78
Tabel 4.49 Data training 10 fold cross validation skenario 10 79
Tabel 4.50 Data testing 10 fold cross validation skenario 10 79
Tabel 4.51 Confusion matrix 10 fold cross validation skenario 10 80
Tabel 4.52 Hasil F1-Measure holdout method terhadap nilai k..... 81
Tabel 4.53 Hasil F1-Measure 10 fold cross validation terhadap nilai k 82
Tabel 4.54 Analisis hasil waktu eksekusi holdout method 85
Tabel 4.55 Analisis hasil waktu eksekusi 10 fold cross validation..... 86
Tabel 4.56 Perbandingan dengan beberapa penelitian sebelumnya..... 89

DAFTAR GAMBAR

Gambar 2.1 Langkah - langkah penemuan pengetahuan (Han, 2012) 8
Gambar 2.2 Tahapan Text Mining (Tutkan et al, 2016) 10
Gambar 2.3 Proses Case Folding (Nugraha, 2014) 12
Gambar 2.4 Proses Character Removal (Nugraha, 2014) 12
Gambar 2.5 Proses Tokenizing (Nugraha, 2014) 13
Gambar 2.6 Proses Stopword Removal (Nugraha, 2014) 13
Gambar 2.7 Estimating accuracy with the holdout method 19
Gambar 2.8 Ilustrasi pembentukan klasifikasi (Sartono, 2015) 21
Gambar 2.9 Ilustrasi pembentukan model klasifikasi (Tedja, 2015) 23
Gambar 2.10 Pendekatan umum untuk model klasifikasi (Tedja, 2015) 24
Gambar 2.11 Ilustrasi konsep KNN (Informatikalogi, 2017) 29
Gambar 3.1 Desain Penelitian 38
Gambar 3.2 Metode Waterfall (Sukamto dan Shalahuddin, 2014) 41
Gambar 4.1 File dataset pada database mysql 43
Gambar 4.2 Diagram alir data preparation 45
Gambar 4.3 Diagram alir text preprocessing 46
Gambar 4.4 Diagram alir case folding dan character removal 47
Gambar 4.5 Contoh proses case folding dan character removal 47
Gambar 4.6 Diagram alir tokenizing 48
Gambar 4.7 Contoh proses tokenizing 48
Gambar 4.8 Contoh tabel t_tokenisasi pada database mysql 49
Gambar 4.9 Diagram alir proses stopword removal 50
Gambar 4.10 Contoh proses stopword removal 50
Gambar 4.11 Contoh tabel stopword list pada database MySql 51
Gambar 4.12 Contoh tabel t_stopword pada database MySql 51
Gambar 4.13 Diagram alir proses pembobotan TF-IDF 52
Gambar 4.14 Contoh tabel hasil TF-IDF 53
Gambar 4.15 Diagram alir proses klasifikasi KNN 53
Gambar 4.16 Contoh tabel hasil cosine similarity 54
Gambar 4.17 Contoh tabel hasil testing 54
Gambar 4.18 Diagram alir sistem 55
Gambar 4.19 Desain antar muka 58
Gambar 4.20 Grafik hasil F1-Measure holdout method terhadap nilai k 81
Gambar 4.21 Grafik hasil F1-Measure 10 fold cross validation terhadap nilai k 82
Gambar 4.22 Grafik F1-Measure holdout method terhadap data training 83
Gambar 4.23 Grafik F1-Measure 10 fold cross validation terhadap data training 84
Gambar 4.24 Grafik perbedaan jumlah dataset penelitian terkait 87
Gambar 4.25 Grafik perbedaan hasil penelitian dengan penelitian terkait 88

DAFTAR LAMPIRAN

Lampiran 1. Contoh *dataset* abstrak 3 kategori 98
Lampiran 2. Kumpulan *stopwords* bahasa Inggris..... 126
Lampiran 3. Kumpulan *psedocode* fungsi 133