PENERAPAN GINI INDEX DAN K-NEAREST NEIGHBOR UNTUK KLASIFIKASI TINGKAT KOGNITIF SOAL PADA TAKSONOMI BLOOM

Tyas Setiyorini [1], Rizky Tri Asmono [2]
Program Studi Teknik Informatika [12]
STMIK Nusa Mandiri Jakarta [1]; STMIK Swadharma [2]
http://www.nusamandiri.ac.id [1]; http://www.swadharma.ac.id/[2]
tyas.setiyorini@gmail.com[1]; rtriasmono@gmail.com[2]

Abstract — *As a guideline in designing the proper exam, which consists of questions that have varying degrees of cognitive, Bloom's Taxonomy has been applied widely. Currently, the educators identify the cognitive level of the problem in Bloom's Taxonomy still using the manual way. Only slight educators can identify cognitive level correctly, mostly made a mistake in classifying questions. K-Nearest Neighbor (KNN) is an effective method for classifying cognitive level problems on Bloom's taxonomy, however KNN has the disadvantage that is the computational complexity of similarity of data was big when dimensional of data was high. To solve the problem, it required the Gini Index method to reduce the high feature dimension. Several experiments were conducted to obtain the best architecture and produce an accurate classification. The results from 10 experiments on the Question Bank dataset with KNN obtained the greatest accuracy was 59.97% and the highest kappa was 0.496. Then the KNN + Gini Index obtained the greatest accuracy is 66.18% and the highest kappa is 0.574. Based on these results it can be concluded that Gini Index is able to reduce the dimensions of high features, thus improving the performance of KNN and improve the level of accuracy of the cognitive level classification of the problem on Bloom's Taxonomy.*

Intisari — Sebagai pedoman dalam merancang ujian yang layak, yang terdiri dari soal-soal yang memiliki berbagai tingkatan secara kognitif, Taksonomi Bloom telah diterapkan secara luas. Saat ini, kalangan pendidik mengidentifikasi tingkat kognitif soal pada Taksonomi Bloom masih menggunakan cara manual. Hanya sedikit pendidik yang dapat mengidentifikasi tingkat kognitif dengan benar, sebagian besar melakukan kesalahan dalam mengklasifikasikan soal-soal. K-Nearest Neighbor (KNN) adalah metode yang efektif untuk klasifikasi tingkat kognitif soal pada Taksonomi Bloom, tetapi KNN memiliki kelemahan yaitu kompleksitas komputasi kemiripan datanya besar apabila dimensi fitur datanya tinggi. Untuk menyelesaikan kelemahan tersebut diperlukan metode Gini Index untuk mengurangi dimensi fitur yang tinggi. Beberapa percobaan dilakukan untuk memperoleh arsitektur yang terbaik dan menghasilkan klasifikasi yang akurat. Hasil dari 10 percobaan pada dataset *Question Bank* dengan KNN diperoleh akurasi tertinggi yaitu 59,97% dan kappa tertinggi yaitu 0,496. Kemudian pada KNN+Gini Index diperoleh akurasi tertinggi yaitu 66,18% dan kappa tertinggi yaitu 0,574. Berdasarkan hasil tersebut maka dapat disimpulkan bahwa Gini Index mampu mengurangi dimensi fitur yang tinggi, sehingga meningkatkan kinerja KNN dan meningkatkan tingkat akurasi klasifikasi tingkat kognitif soal pada Taksonomi Bloom.

Kata Kunci*: klasifikasi, Taksonomi Bloom, K-Nearest Neighbor, Gini Index*

PENDAHULUAN

Taksonomi Bloom diciptakan pada tahun 1948 oleh seorang psikolog yang bernama Benjamin Bloom dan beberapa rekannya (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). Awalnya dikembangkan sebagai metode klasifikasi sasaran pendidikan untuk mengevaluasi hasil belajar siswa. Taksonomi Bloom telah direvisi selama bertahun-tahun dan masih diterapkan dalam pendidikan saat ini. Tujuan awal dalam pembuatan Taksonomi Bloom adalah memfokuskan pada tiga domain utama dari pembelajaran, yaitu kognitif, afektif, dan psikomotorik. Meskipun tujuannya untuk mengatasi semua domain, Taksonomi Bloom diterapkan hanya untuk memperoleh pengetahuan dalam domain kognitif. Domain kognitif mengandung tingkah laku yang mengutamakan aspek intelektual, seperti kemampuan berpikir, pengetahuan, dan pemahaman

Tingkat kognitif terdiri dari 6 tingkat. Tingkat pertama adalah tingkat pengetahuan (*knowlegde level*), yaitu kemampuan mengingat kembali materi yang telah dipelajari. Kedua, tingkat pemahaman (*comprehension level*) yaitu kemampuan untuk menjelaskan pengetahuan yang dipelajari. Ketiga, tingkat aplikasi (*application level*), yaitu kemampuan untuk menerapkan pengetahuan yang telah dipelajari

dalam suatu tindakan nyata. Keempat, tingkat analisis (*analysis level*) yaitu kemampuan untuk menyelidiki pengetahuan yang mereka pelajari. Kelima, tingkat sintesis (*synthesis level*), yaitu kemampuan untuk menghubungkan semua pengetahuan dan mengintegrasikan menjadi sesuatu hal yang baru. Tingkat terakhir adalah tingkat evaluasi (*evaluation level*) yaitu kemampuan untuk menilai manfaat dari suatu pengetahuan.

Sebagai pedoman dalam merancang ujian yang layak, yang terdiri dari soal-soal yang memiliki berbagai tingkatan secara kognitif, Taksonomi Bloom telah diterapkan secara luas (Jones, Harland, Reid, & Bartlett, 2009). Taksonomi Bloom telah banyak diterapkan dalam pengukuran dan penilaian proses pembelajaran (Khairuddin & Hashim, 2008)(Thompson, Luxton-Reilly, Whalley, Hu, & Robbins, 2008). Penerapan identifikasi tingkat kognitif soal pada Taksonomi Bloom telah banyak dilakukan oleh kalangan pendidik saat ini, namun cara yang digunakan masih bersifat manual. Hanya sedikit yang dapat mengidentifikasi tingkat kognitif dengan benar, sebagian besar melakukan kesalahan dalam mengklasifikasikan soal-soal (Yusof & Hui, 2010). Tingginya jumlah soal-soal yang harus diklasifikasikan dalam dokumen ujian, sehingga proses penentuan tingkat kognitif Bloom membutuhkan waktu yang lama dan melelahkan. Oleh karena itu dibutuhkan metode yang mampu mengidentifikasi tiap-tiap soal menurut tingkat kognitifnya secara otomatis.

Semakin berkembangnya dokumen, klasifikasi teks menjadi teknologi kunci untuk menangani dan mengatur sejumlah dokumen yang kompleks (Shang, Huang, & Zhu, 2007). Klasifikasi teks merupakan salah satu bagian dari *text mining*. *Text mining* adalah proses menggali pengetahuan dari data teks (Fayyad, 1996). *Data mining* adalah proses penyaringan informasi yang bermanfaat dari fakta yang disimpan di dalam database terstruktur, sedangkan text mining menemukan pola dari teks bahasa alami (Weiss et al., 1999). *Text mining* dimulai dengan memilih koleksi dokumen, seperti pekerjaan pada *data mining* (Cherfi, Napoli, & Toussaint, 2005). *Text mining* terdiri dari 7 jenis yaitu klasifikasi dokumen, pengelompokan dokumen, ekstraksi informasi, pemulihan informasi, web mining, pemrosesan bahasa alami dan ektraksi konsep (Miner, 2012). Klasifikasi teks adalah proses untuk mengklasifikasikan sekumpulan dokumen ke dalam kategori yang telah ditetapkan (Genkin, Lewis, & Madigan, 2007).

Banyak metode yang telah diaplikasikan beberapa tahun terakhir pada klasifikasi teks berdasarkan teori pembelajaran mesin dan statistik. Seperti contoh, K-Nearest Neighbor (KNN)(Cover & Hart, 1967)(Yiming Yang & Liu, 1999)(Tan, 2005), Support Vector Machines (SVM) (Joachims, 1998), Naive Bayes (Lewis, 1998), decision tree (DD & Ringuette, 1994), neural network, linear least squares fit, SWAP-1, dan Rocchio. Tantangan utama dari klasifikasi teks dalam menyelesaikan permasalahan dunia nyata yaitu klasifikasi hirarkis, ketidakseimbangan klasifikasi, dan mengklasifikasikan data teks besar secara efisien (SU, 2006). Ada hubungan yang kompleks antara kategori dalam permasalahan klasifikasi multi-kategori. Untuk lebih memahami hubungan ini, beberapa klasifikasi teks hierarkis telah diteliti (Hao, Chiang, & Tu, 2007)(Esuli, Fagni, & Sebastiani, 2008). Dalam permasalahan klasifikasi teks, kinerja KNN terbukti lebih baik dibandingkan dengan NN, NB dan Rocchio (Yiming Yang & Liu, 1999).

KNN merupakan metode yang efektif untuk klasifikasi teks, namun memiliki kelemahan yaitu ruang penyimpanannya besar, kompleksitas komputasi kemiripan datanya besar serta mudah dipengaruhi oleh data noise (Guo, Wang, Bell, Bi, & Greer, 2003). Bagi banyak algoritma pembelajaran, dimensi fitur yang tinggi tidak diizinkan (Shang et al., 2007). Banyak peneliti mencari cara untuk mengurangi kompleksitas KNN, yang dapat dibagi menjadi 3 metode umum, yaitu mengurangi dimensi fitur yang tinggi (de Vries, Mamoulis, Nes, & Kersten, 2002), mengurangi jumlah data pelatihan (Lu & Fa, 2004), dan mempercepat proses menemukan K tetangga terdekat (Aghbari, 2005).

Dalam mengatasi kelemahan pada KNN, seleksi fitur (*feature selection*) dapat digunakan untuk mengurangi dimensi fitur yang tinggi pada klasifikasi teks tingkat kognitif soal pada Taksonomi Bloom. Dimensi fitur yang tinggi merupakan masalah utama pada klasifikasi teks (Shang et al., 2007). Seleksi fitur didasarkan pada pengurangan fitur yang besar, yaitu dengan menghapus atribut yang tidak relevan (Koncz & Paralic, 2011).

Seleksi fitur adalah salah satu bagian terpenting untuk meningkatkan kinerja *classifier* (Wang, Li, Song, Wei, & Li, 2011). Menggunakan metode seleksi fitur yang tepat dapat meningkatkan akurasi (Xu, Peng, & Cheng, 2012)(Forman, 2003). Saat ini, beberapa metode seleksi fitur yang terkenal yaitu information gain, mutual information, expected cross entropy, term frequency the weight of evidence of text, odds ratio, CHI (Yang & Pedersen, 1997)(Mladenić & Grobelnik, 2003)(Mladenic & Grobelnik, 1999).

Shang et al menyajikan metode seleksi fitur teks baru yaitu Gini Index, yang digunakan dalam pohon keputusan untuk memisahkan atribut dan mendapat ketepatan klasifikasi yang lebih baik

(Shang et al., 2007). Gini Index mampu mengatasi masalah utama pada klasifikasi teks yaitu mengurangi dimensi fitur yang tinggi (Shang et al., 2007). Shankar dan Karypis membahas bagaimana menggunakan Gini Index untuk seleksi fitur teks dan penyesuaian bobot (Shankar & Karypis, 2000). Dari beberapa percobaan yang dilakukan, Gini Index menunjukkan kinerja klasifikasi yang lebih baik dengan metode seleksi fitur lainnya (Shang et al., 2007).

Dari penjelasan di atas tampak bahwa Gini Index memiliki potensi yang lebih baik dalam proses mengurangi dimensi fitur yang tinggi. Untuk itu penelitian ini akan menggunakan kombinasi kedua metode yaitu KNN dan Gini Index tersebut untuk memungkinkan klasifikasi tingkat kognitif soal pada taksonomi bloom dapat dilakukan secara otomatis dengan tingkat akurasi yang terbaik.

BAHAN DAN METODE

K-Nearest Neighbor (KNN)

KNN merupakan metode algoritma *supervised learning*, di mana hasil dari *query instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Kelas yang paling banyak timbul yang akan menjadi kelas hasil klasifikasi. KNN adalah salah satu metode pengklasifikasian data berdasarkan similaritas dengan label data (Larose, 2006a). Algortima KNN merupakan sebuah metode untuk melakukan klasifikasi terhadap objek baru berdasarkan k tetangga terdekatnya. KNN adalah algoritma pembelajaran berbasis instan yang menggunakan jarak terdekat dalam menentukan kategori vektor baru dalam set data *training* (Gorunescu, 2011).

Metode KNN memiliki permasalahan yaitu menemukan tetangga terdekat k pada titik query dari dataset yang digunakan (Liaw, Yi-Ching, Leou Maw-Lin, 2010)(Liaw, Wu, & Leou, 2010). Metode ini banyak digunakan untuk mengatasi masalah dalam bidang ilmiah dan rekayasa perangkat lunak seperti pengenalan pola, pengenalan objek, pengelompokkan data, fungsi approximate, kuantisasi vektor, klasifikasi pola.

Untuk menentukan jumlah data atau tetangga terdekat ditentukan oleh user yang dinyatakan dengan k.

$$d_i=\sqrt{\sum_{i=1}^{p}(x_{1i}-x_{2i})^2} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Di mana x1= (x11, x12,..., x1n) dan x2= (x21, x22,..., x2n). Dengan kata lain, untuk setiap atribut numerik, kita mengambil perbedaan antara yang sesuai nilai-nilai atribut yang dalam vektor x1 dan x2 dari matriks dengan ukuran dimensi. Akar kuadrat diambil dari akumulasi jumlah total jarak. Biasanya, kita menormalkan nilai masing-masing atribut sebelum digunakan (Jain & Richariya, 2012). Prinsip kerja KNN adalah mencari jarak terdekat antara data yang dievaluasi dengan k tetangga terdekatnya dalam data pelatihan. Persamaan penghitungan untuk mencari Euclidean dengan d adalah jarak dan p adalah dimensi data dengan:

$$d_i=\sqrt{\sum_{i=1}^{p}(x_{1i}-x_{2i})^2} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

di mana:
x1: sample data uji
x2: data uji
d: jarak
p: dimensi data

Gini Index

Gini Index dapat dianggap sebagai probabilitas dari dua data yang dipilih secara acak yang memiliki *class* yang berbeda. Gini Index digunakan oleh Breiman (Breiman, Friedman, Stone, & Olshen, 1984) untuk menghasilkan pohon klasifikasi pada decision tree.

Misalkan S adalah 1 set dari sejumlah s data. Data ini memiliki sejumlah m *class* yang berbeda (C_i, i= 1, ..., m). Berdasarkan pada *class* tersebut, kita bisa membagi S ke dalam sejumlah m subset (S_i, i= 1, ..., m)misalkan Si adalah *dataset* yang tergabung di dalam *class* C_i, s_i adalah jumlah data dari S_i, maka Gini Index dapat dirumuskan sebagai berikut:

$$\text{Gini Index } (S) = 1 - \sum_{i=1}^{m}\left(\frac{s_i}{s}\right)^2 \quad \dots\dots\dots\dots\dots\dots (3)$$

Metode penelitian

1. Pengumpulan Data

Dataset yang digunakan pada penelitian ini mengarah pada penelitian yang dilakukan oleh Chai Jing (HUI, 2009). *Dataset* menggunakan bahasa Inggris, sehingga *stopwords* yang digunakan untuk *preprocessing* data juga berbahasa Inggris.

Tabel 1 menampilkan dataset terdiri dari 274 jumlah soal yang secara keseluruhan yang diklasifikasikan dalam 6 tingkat kognitif soal. Klasifikasi *knowledge* terdapat 28 jumlah soal, klasifikasi *comprehension* terdapat 44 jumlah soal, klasifikasi *application* terdapat 41 jumlah soal, klasifikasi *analysis* terdapat 48 jumlah soal, klasifikasi *synthesis* terdapat 59 jumlah soal, dan klasifikasi *evaluation* terdapat 54 jumlah soal.

Tabel 1. Klasifikasi Soal

Klasifikasi	Jumlah
Knowledge	28
Coprehension	44
Application	41
Analysis	48
Synthesis	59
Evaluation	54

Sumber: Setyorini & Asmono (2017)

2. Metode yang Diusulkan

Gambar 1 menampilkan metode yang diusulkan yaitu metode Gini Index pada KNN. Pada pengolahan awal, semua pertanyaan pada *Question Bank* diubah menjadi dataset yang diklasifikasikan dengan menggunakan teknik *text processing*.

Sumber: Setyorini & Asmono (2017)

Gambar 1. Metode KNN+Gini Index

Text processing yang digunakan yaitu *tokenizer*, *filter stopwords*, dan *stem*. *Tokenize* adalah proses untuk memisah-misahkan kata. Tahapan *tokenize* dimulai dari memisah-misahkan bagian *review* yang dipisahkan dengan karakter spasi. Bagian yang memiliki karakter non alphabet dan angka akan dibuang. Bagian yang masuk dalam daftar *emoticon* akan dikonversi. Bagian yang memiliki karakter selain alfabet, angka, dan garis bawah akan dipecah sesuai posisi karakter tersebut. Selanjutnya, pada proses filter *stopwords* akan dihilangkan kata-kata yang sering muncul. Terakhir, tahap *stemming* adalah tahap untuk mencari kata dasar dari tiap kata hasil filter.

Fitur dari dataset yang dihasilkan pada tahap *text processing* diseleksi menggunakan metode Gini Index. Setelah diseleksi fitur-fitur tersebut akan digunakan pada proses klasifikasi menggunakan metode KNN, kemudian akan dihitung nilai akurasi dan kappa untuk mendapatkan fitur yang terbaik.

HASIL DAN PEMBAHASAN

Penelitian yang dilakukan menggunakan komputer dengan spesifikasi CPU Intel Core i5 1.6GHz, RAM 4GB, dan sistem operasi Microsoft Windows 10 Professional 64-bit. Aplikasi yang digunakan adalah RapidMiner 7.3. Data penelitian ini menggunakan dataset *Question Bank*. Dataset *Question Bank* mengarah pada penelitian yang dilakukan oleh Chai Jing Hui (HUI, 2009).

Setelah percobaan yang dilakukan dengan KNN dan KNN+Gini Index, kemudian dibandingkan hasil akurasi dan kappa pada metode KNN dengan KNN+Gini Index dari 10 percobaan. Pada Tabel 2 dari 10 percobaan dan rata-rata keseluruhan percobaan secara tetap menunjukkan peningkatan nilai akurasi dan kappa yang signifikan antara KNN dengan KNN+Gini Index.

Tabel 2. Hasil Akurasi dan *Kappa* dengan KNN dan KNN+Gini Index

KNN		KNN + Gini Index	
Akurasi	***Kappa***	**Akurasi**	***Kappa***
53.13%	0.404	68.37%	0.607
47.95%	0.345	63.61%	0.547
47.45%	0.34	65.63%	0.576
53.76%	0.419	60.89%	0.501
53.16%	0.409	62.45%	0.531
59.97%	0.496	66.18%	0.571
58.89%	0.482	62.08%	0.513
59.42%	0.49	57.84%	0.464
55.76%	0.444	62.55%	0.521
56.71%	0.456	57.82%	0.461

Sumber: Setyorini & Asmono (2017)

Penggunaan KNN+Gini Index memperoleh nilai akurasi dan kappa yang lebih tinggi dibanding dengan penggunaan KNN. Hal ini dibuktikan dari 10 percobaan yang menunjukkan bahwa peningkatan nilai akurasi dan kappa yang

tetap dan signifikan. Dari hasil percoban tersebut menunjukkan bahwa penggunaan Gini Index pada KNN mampu mengurangi dimensi fitur yang tinggi, sehingga menghasilkan kinerja atau tingkat akurasi klasifikasi tingkat kognitif soal pada taksonomi Bloom yang lebih baik dibanding dengan menggunakan metode KNN. Hal ini membuktikan penelitan Shang et al bahwa Gini Index mampu mengurangi dimensi ruang fitur yang tinggi sehingga mendapat ketepatan klasifikasi yang lebih baik (Shang et al., 2007). Selain itu juga membuktikan penelitan Supriyanto et al. bahwa KNN mampu mengklasifikasi tingkat kognitif soal pada taksonomi Bloom (Supriyanto, Yusof, Nurhadiono, & Sukardi, 2013).

Untuk membuktikan hasil sementara (hipotesis) penelitian ini, diperlukan pengujian dengan cara statistik untuk menguji perbedaan antara penggunaan metode KNN dengan KNN+Gini Index, apakah terdapat perbedaan di antara keduanya. Pengujian hipotesis ini menggunakan metode t-Test. Metode ini termasuk yang paling umum dalam metode statistik tradisional, yaitu t-Test (Maimon, Oded&Rokach, 2010). Ada atau tidaknya perbedaan antara dua model membutuhkan pengujian, salah satunya dengan uji t-Test (Larose, 2006b), dengan melihat nilai P. Jika nilai P < 0,05 maka menunjukkan hipotesis nol ditolak atau disebut hipotesis alternatif (Sumanto, 2014). Hipotesis nol menunjukkan tidak ada perbedaan antara dua buah variabel, sedangkan hipotesis alternatif menunjukkan adanya perbedaan antara dua buah variabel (Sumanto, 2014).

Pada Tabel 3 menampilkan t-Test untuk hasil akurasi KNN dan KNN+Gini Index yang menunjukkan hipotesis hipotesis alternatif yaitu dengan nilai P < 0,05 yaitu 0,0008. Pada Tabel 4 menampilkan t-Test untuk hasil kappa KNN dan KNN+Gini Index juga menunjukkan hipotesis hipotesis alternatif yaitu dengan nilai P < 0,05 yaitu 0,0010.

Tabel 3. Hasil t-Test Akurasi dengan KNN dan KNN+Gini Index

Akurasi		
	KNN	KNN+Gini Index
Mean	0.5462	0.62742
Variance	0.001961691	0.001164824
Observations	10	10
Pearson Correlation	-0.37731238	
Hypothesized Mean Difference	0	
df	9	
t Stat	-3.93178751	
P(T<=t) one-tail	0.001724336	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.003448673	
t Critical two-tail	2.262157163	

Sumber: Setyorini & Asmono (2017)

Tabel 4. Hasil t-Test *Kappa* dengan KNN dan KNN+Gini Index

Kappa		
	KNN	KNN+Gini Index
Mean	0.4285	0.5292
Variance	0.003119167	0.002259733
Observations	10	10
Pearson Correlation	-0.47538963	
Hypothesized Mean Difference	0	
df	9	
t Stat	-3.582037444	
P(T<=t) one-tail	0.002955946	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.005911892	
t Critical two-tail	2.262157163	

Sumber: Setyorini & Asmono (2017)

Hasil t-Test yaitu hipotesis alternatif tersebut menunjukkan bahwa antara penggunaan metode KNN dengan KNN+Gini Index menunjukkan adanya perbedaan yang signifikan. KNN+Gini Index menghasilkan kinerja atau tingkat akurasi yang lebih baik dibanding dengan menggunakan metode KNN. Hal ini membuktikan penelitan Shang et al. bahwa Gini Index mampu mengurangi dimensi fitur yang tinggi sehingga mendapat ketepatan klasifikasi yang lebih baik (Shang et al., 2007). Serta secara tetap menunjukkan kemampuan meningkatkan akurasi klasifikasi pada KNN dalam mengklasifikasi tingkat kognitif soal pada taksonomi Bloom (Supriyanto et al., 2013). Hal ini menunjukkan bahwa seleksi fitur Gini Index yang diusulkan dapat menjadi metode yang efektif untuk meningkatkan kinerja KNN.

KESIMPULAN

Hasil dari 10 percobaan pada *dataset Question Bank* dengan KNN diperoleh akurasi tertinggi adalah 59,97% dan kappa tertinggi adalah 0,496. Kemudian pada KNN+Gini Index diperoleh akurasi tertinggi adalah 68,37% dan kappa tertinggi adalah 0,607.

Berdasarkan hasil tersebut maka dapat disimpulkan bahwa Gini Index mampu mengurangi dimensi fitur yang tinggi, sehingga meningkatkan kinerja KNN dan meningkatkan tingkat akurasi klasifikasi tingkat kognitif soal pada Taksonomi Bloom.

REFERENSI

Aghbari, Z. Al. (2005). Array-index : a plug & search K nearest neighbors method for high-dimensional data. *Data & Knowledge Engineering*, *52*, 333–352. https://doi.org/10.1016/j.datak.2004.06.015

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals.* Longman Group.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. https://doi.org/10.1002/widm.8

Cherfi, H., Napoli, A., & Toussaint, Y. (2005). Towards a text mining methodology using association rule extraction. *Soft Computing*, *10*(5), 431–441. https://doi.org/10.1007/s00500-005-0504-x

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory IT*, (11), 21–27.

DD, L., & Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval ({SDAIR?4)}*.

de Vries, A. P., Mamoulis, N., Nes, N., & Kersten, M. (2002). Efficient k-NN search on vertically decomposed data. *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data - SIGMOD '02*, 322. https://doi.org/10.1145/564728.564729

Esuli, A., Fagni, T., & Sebastiani, F. (2008). Boosting multi-label hierarchical text categorization. *Information Retrieval*, *11*(4), 287–313. https://doi.org/10.1007/s10791-008-9047-y

Fayyad, U. et al. (1996). From Data Mining to Knowledge Discovery in Databases. *The Computer Journal*, *58*(1), 1–6. https://doi.org/10.1093/comjnl/bxt107

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J Mach Learn Res*, *3*, 1289–1305.

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, *49*(3), 291–304. https://doi.org/10.1198/004017007000000245

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (2011th ed.). Springer. Retrieved from http://books.google.com/books?hl=en&lr=&id=yJvKY-sB6zkC&oi=fnd&pg=PP2&dq=Data+Mining:+Concepts,+Models+and+Techniques&ots=prPCrasSvA&sig=BQoeQzA2JgBjjkngCL1_XpvpfDY

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. *On the Move to Meaningful Internet …*, 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

Hao, P.-Y., Chiang, J.-H., & Tu, Y.-K. (2007). Hierarchically SVM classification based on support vector clustering method and its application to document categorization. *Expert Systems with Applications*, *33*(3), 627–635. https://doi.org/10.1016/j.eswa.2006.06.009

HUI, C. J. (2009). Feature Reduction For Neural Network In Determining The Bloom's Cognitive Level Of Question Items, (October).

Jain, M. M., & Richariya, P. V. (2012). An Improved Techniques Based on Naive Bayesian for Attack Detection. *International Journal of Emerging Technology and Advanced Engineering Website:*, *2*(1), 324–331.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceeedings of the 10th European Conference on Machine Learning*, 137–142. https://doi.org/10.1007/BFb0026683

Jones, K. O., Harland, J., Reid, J. M. V, & Bartlett, R. (2009). Relationship between examination questions and bloom's taxonomy. In *Proceedings - Frontiers in Education Conference, FIE*. https://doi.org/10.1109/FIE.2009.5350598

Khairuddin, N. N., & Hashim, K. (2008). Application of Bloom's taxonomy in software engineering assessments. *Proceedings of the 8th Conference on Applied Computer Science*, 66–69. Retrieved from http://portal.acm.org/citation.cfm?id=1504034.1504048

Koncz, P., & Paralic, J. (2011). An approach to

feature selection for sentiment analysis. *2011 15th IEEE International Conference on Intelligent Engineering Systems*. https://doi.org/10.1109/INES.2011.5954773

Larose, D. T. (2006a). *Data Mining Methodes And Model*. https://doi.org/10.1002/0471756482

Larose, D. T. (2006b). *Data Mining Methods and Model*. New Jersey: John Willey & Sons, Inc.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval (pp. 4–15). https://doi.org/10.1007/BFb0026666

Liaw, Yi-Ching, Leou Maw-Lin, W. C.-M. (2010). Fast exact k nearest neighbors search using anorthogonal search tree. *Pattern Recognition*, *43*(6), 2351–2358. https://doi.org/10.1016/j.patcog.2010.01.003

Liaw, Y. C., Wu, C. M., & Leou, M. L. (2010). Fast k-nearest neighbors search using modified principal axis search tree. *Digital Signal Processing: A Review Journal*, *20*(5), 1494–1501. https://doi.org/10.1016/j.dsp.2010.01.009

Lu, L. R., & Fa, H. Y. (2004). A Density-Based Method for Reducing the Amount of Training Data in kNN Text Classification [J]. *Journal of Computer Research and Development*, *4*, 3.

Maimon, Oded&Rokach, L. (2010). *Data mining and knowledge discovey handbook*. New York: Springer.

Miner, G. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. https://doi.org/10.1016/B978-0-12-386979-1.09002-2

Mladenic, D., & Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive bayes. In *Proceedings of the Sixteenth International Conference (ICML 1999)* (pp. 258–267). https://doi.org/10.1214/aoms/1177705148.

Mladenić, D., & Grobelnik, M. (2003). Feature selection on hierarchy of web documents. *Decision Support Systems*, *35*(1), 45–87. https://doi.org/10.1016/S0167-9236(02)00097-0

Setiyorini, T, Asmono, R, T. (2017). Laporan Akhir Penelitian Mandiri. Jakarta: STMIK Nusa Mandiri

Shang, W., Huang, H., & Zhu, H. (2007). A novel feature selection algorithm for text categorization, *33*, 1–5. https://doi.org/10.1016/j.eswa.2006.04.001

Shankar, S., & Karypis, G. (2000). A Feature Weight Adjustment Algorithm for Document Categorization.

SU, J.-S. (2006). Advances in Machine Learning Based Text Categorization. *Journal of Software*, *17*(9), 1848. https://doi.org/10.1360/jos171848

Sumanto. (2014). *Statistika Deskriptif*. Yogyakarta: Center of Academic Publishing Service.

Supriyanto, C., Yusof, N., Nurhadiono, B., & Sukardi. (2013). Two-level feature selection for naive bayes with kernel density estimation in question classification based on Bloom's cognitive levels. In *2013 International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 237–241). IEEE. https://doi.org/10.1109/ICITEED.2013.6676245

Tan, S. (2005). Neighbor-weighted K-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, *28*(4), 667–671. https://doi.org/10.1016/j.eswa.2004.12.023

Thompson, E., Luxton-Reilly, A., Whalley, J. L., Hu, M., & Robbins, P. (2008). Bloom's taxonomy for CS assessment. *Conferences in Research and Practice in Information Technology Series*, *78*, 155–161.

Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications*, *38*(7), 8696–8702. https://doi.org/10.1016/j.eswa.2011.01.077

Weiss, S. M., Apte, C., Damerau, F. J., Johnson, D. E., Oles, F. J., Goetz, T., & Hampp, T. (1999). Maximizing text-mining performance. *Intelligent Systems and Their Applications,*

IEEE, *14*(4), 63–69. https://doi.org/10.1109/5254.784086

Xu, T., Peng, Q., & Cheng, Y. (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowledge-Based Systems*, *35*, 279–289. https://doi.org/10.1016/j.knosys.2012.04.011

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99* (pp. 42–49). https://doi.org/10.1145/312624.312647

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14 th International Conference on Machine Learning.* (pp. 412–420).

Yusof, N., & Hui, C. J. (2010). Determination of Bloom's cognitive level of question items using artificial neural network. In *Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, ISDA'10* (pp. 866–870). https://doi.org/10.1109/ISDA.2010.5687152