ANALISA PERBANDINGAN ALGORITMA C4.5 DAN NAIVE BAYES DALAM MELAKUKAN KLASIFIKASI TEKS BERITA

Diajukan untuk Menyusun Tugas Akhir

di Jurusan Teknik Informatika Fakultas Ilmu Komputer UNSRI

Oleh :

FARIS HARUN AHMAD

09021281320012

Jurusan Teknik Informatika

FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA

2019

LEMBAR PENGESAHAN USULAN TUGAS AKHIR

ANALISA PERBANDINGAN ALGORITMA C4.5 DAN NAIVE BAYES DALAM MELAKUKAN KLASIFIKASI TEKS BERITA

Oleh :

Faris Harun Ahmad
NIM : 09021281320012

Palembang, September 2019

Pembimbing I,	Pembimbing II,
Yoppy Sazaki, M.T	Danny Matthew Saputra, M.Sc.
NIPUS. 197406062012101201	NIP. 198505102015041002

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, MT
NIP. 197706012009121004

TANDA LULUS SIDANG TUGAS AKHIR

Pada hari Jum'at, 23 Agustus 2019 telah dilaksanakan ujian sidang tugas akhir oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Faris Harun Ahmad

NIM : 09021281320012

Judul : Analisa Perbandingan Algoritma C4.5 dan Naive Bayes dalam Melakukan Klasifikasi Teks Berita

1. Pembimbing I

Yoppy Sazaki, M.T .....
NIPUS. 197406062012101201

2. Pembimbing II

Danny Matthew Saputra, M.Sc. .....
NIP. 198505102015041002

3. Penguji I

Drs. Megah Mulya, M.T. .....
NIP. 196602202006041001

4. Penguji II

Kanda Januar Miraswan, M.T .....
NIP. 199001092019031012

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T
NIP. 197706012009121004

HALAMAN PERNYATAAN BEBAS PLAGIAT

Yang bertanda tangan di bawah ini :

Nama	:	Faris Harun Ahmad
NIM	:	09021281320012
Program Studi	:	Teknik Informatika
Judul Skripsi	:	Analisa Perbandingan Algoritma C4.5 dan Naive Bayes dalam Melakukan Klasifikasi Teks Berita
Hasil Pengecekan Software *iThenticate/Turnitin*	:	19 %

Menyatakan bahwa Laporan Projek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan projek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Palembang, September 2019

Faris Harun Ahmad
NIM. 09021281320012

Motto:

1. Allah is above everything in this world.

2. You can't fight your destiny, but even destiny contains path. It means you can choose paths that you will take to your future.

3. If you can't do something that you currently want to, you are not on the right level yet. Keep levelling up yourself, and you will be capable of doing everything.

4. Love is a condition where your brain and heart is on one way.

I present this work to :

- Allah SWT
- Myself
- Both of my lovely parents
- My fellow friends and closest friends
- My favorite alma mater

ANALISA PERBANDINGAN ALGORITMA C4.5 DAN NAIVE BAYES DALAM MELAKUKAN KLASIFIKASI TEKS BERITA

By:

FARIS HARUN AHMAD

09021281320012

ABSTRACT

Classification is one of the data mining techniques used to predict group membership in data instances. Text classification is a branch of classification that classifies a set of documents into automatically assigned categories. C4.5 and Naive Bayes algorithms are two algorithms that are often compared in the classification tasks because both of them have high accuracy, but generally only with the implementation of numeric datasets. In this study the C4.5 and Naive Bayes algorithms use word weighting techniques and pre-processing to finally predict the classes, and then the performace can be compared to see if they still maintain good performance or not. The C4.5 algorithm has threshold, entropy, info, and gain values which has an important role in building a decision tree, and related to the prediction of each document, variations in the gain value, and the frequency of occurrences for each word in the dataset and the key to making tuples in decision tree. While in the Naïve Bayes Algorithm, predictions depend on the posterior value that can be obtained by multiplying all the word weights for each document and comparing them by the training set. Naive Bayes algorithm with a total of 500 training data text documents resulting a high accuracy on 97.4% and an efficient computing time of 98.45 seconds.

Keywords: News classifications, Income Value, Posterior, C4.5 Algorithm, Naïve Bayes Algorithm

Indralaya, September 2019

Pembimbing I,

Yoppy Sazaki, M.T
NIPUS. 197406062012101201

Pembimbing II,

Danny Matthew Saputra, M.Sc.
NIP. 198505102015041002

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T
NIP. 199770601200912100 4

ANALISA PERBANDINGAN ALGORITMA C4.5 DAN NAIVE BAYES DALAM MELAKUKAN KLASIFIKASI TEKS BERITA

Oleh:

FARIS HARUN AHMAD

09021281320012

ABSTRAK

Klasifikasi merupakan salah satu teknik dari *data mining* yang digunakan untuk memprediksi keanggotaan kelompok terhadap *data instances*. Klasifikasi teks merupakan cabang dari klasifikasi yang menggolongkan satu set dokumen ke dalam kategori yang telah ditetapkan secara otomatis. Algoritma C4.5 dan Naive Bayes merupakan dua algoritma yang sering dibandingkan dalam proses klasifikasi karena tingkat akurasi yang tinggi, akan tetapi pada umumnya hanya dengan implementasi dataset numerik. Pada penelitian ini algoritma C4.5 dan Naive Bayes menggunakan teknik pembobotan kata dan melakukan praproses untuk melakukan prediksi sehingga dapat diketahui apakah kedua algoritma tersebut akan tetap memiliki kinerja yang baik. Algoritma C4.5 memiliki nilai *threshold*, entropi, *info*, dan *gain* yang memegang peran penting dalam hal membangun decision tree, yang mengarah pada prediksi setiap dokumen, variasi nilai *gain*, dan frekuensi kemunculan untuk setiap kata pada dataset adalah kunci untuk membuat tuple di decision tree. Sedangkan dalam Algoritma Naïve Bayes, prediksi tergantung pada nilai posterior yang dapat diperoleh dengan mengalikan semua bobot kata untuk setiap dokumen dan membandingkannya dengan set pelatihan. Algoritma Naive Bayes dengan jumlah data latih sebanyak 500 dikumen teks menghasilkan akurasi tertinggi dengan besaran 97,4% dan waktu komputasi yang efisien yaitu 98,45 detik.

Kata kunci: Klasifikasi berita, Gain, Posterior, Algoritma C4.5, Algoritma Naïve Bayes

Indralaya, September 2019

Pembimbing I,

Yoppy Sazaki, M.T
NIPUS. 197406062012101201

Pembimbing II,

Danny Matthew Saputra, M.Sc.
NIP. 198505102015041002

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T
NIP. 197706012009121004

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

BAB 1

PENDAHULUAN

1.1 Pendahuluan

Pada bab ini membahas tentang latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, dan batasan masalah. Bab ini akan menjelaskan keseluruhan penelitian secara umum. Pendahuluan dimulai dengan penjelasan mengenai latar belakang masalah dimana kedua metode yang dibandingkan yaitu C4.5 dan Naïve Bayes dapat menyelesaikan kasus klasifikasi berita secara tepat berdasarkan penilitian yang telah dilakukan sebelumnya.

1.2 Latar Belakang Masalah

Klasifikasi merupakan salah satu teknik dari *data mining* yang digunakan untuk memprediksi keanggotaan kelompok terhadap *data instances*. Penyelesaian masalah dengan klasifikasi merupakan suatu hal yang fundamental, baik dalam data mining maupun *knowldege discovery*. Klasifikasi bertujuan untuk menetapkan setiap target di dalam sebuah set data menjadi sebuah set kelas atau grup yang telah ditentukan (Kesavaraj & Sukumaran, 2013).

Metode *decision tree*, menurut Hssina, Merbouha et al. (2014) sering digunakan untuk melakukan klasifikasi karena merupakan struktur hirarkis yang sederhana bagi pengguna untuk dimengerti dan membantu dalam pengambilan keputusan. *Decision tree* melakukan partisi dari suatu set data ke dalam grup sehomogen mungkin agar variabel kemudian dapat diprediksi, karena pada dasarnya *tree* menerapkan algoritma *greedy* untuk melakukan induksi secara

rekursif, dari atas ke bawah, serta bersifat *divide and conquer* (Sumathi & Esakkirajan, 2007).

C4.5 merupakan algoritma *decision tree* yang merupakan pengembangan dari algoritma ID3. Menurut Masetic et al. (2016), algoritma ini dikenal karena memiliki karakter sederhana, namun efektif. Algoritma C4.5 bekerja dengan melewati *decision tree*, mengunjungi setiap *node* dan memilih *split* yang optimal. *Decision tree* yang dibentuk oleh C4.5 menggunakan entropy untuk menentukan nilai *information gain* terbaik sehingga dapat menentukan *split*. Sebagai suksesor dari *decision tree* sebelumnya, algoritma C4.5 mampu melakukan klasifikasi baik dengan atribut diskrit maupun kontinyu. Akan tetapi, algoritma ini memiliki kelemahan yaitu sering terjadinya *over fitting* apabila terdapat data dengan karakteristik yang tidak biasa (Singh & Gupta, 2014).

Naïve Bayes Classifier (NBC) merupakan salah satu algoritma yang sering digunakan selain *decision tree* dan telah menunjukkan efisiensi tinggi pada data yang bervariasi, karena algoritma ini bekerja dengan mengasumsikan bahwa atribut data bersifat independen atau tidak memiliki ketergantungan, kemudian mengestimasi probabilitas kelas dengan kondisi tertentu. Disamping kesederhanaan yang dimilikinya, menurut penjelasan Taheri, Mammadov et al. (2010) yang diadopsi dari pendapat Chickering (1996) dan Heckerman (2004), asumsi dengan independensi yang tinggi pada *Naïve Bayes* akan mengganggu kinerjanya ketika data yang digunakan memiliki ketergantungan terhadap data lain.

Pengklasifikasian data menggunakan C4.5 dan *Naïve Bayes* telah dilakukan pada beberapa penelitian sebelumnya, seperti seleksi subset fitur menggunakan *Naïve Bayes* dalam klasifikasi teks yang dilakukan oleh Feng, Guo et al., (2015). Penelitian tersebut membandingkan metode *Naïve Bayes* dengan *Support Vector Machine* dan beberapa metode *Naïve Bayes* yang telah dimodifikasi dengan fitur seleksi dan *weighting* yang hasilnya menunjukkan bahwa metode *Naïve Bayes* memiliki akurasi yang baik dari 4 dataset yang digunakan dengan persentase 81% hingga 95%. Penelitian sebelumnya dengan menggunakan C4.5 yaitu penelitian oleh Muniyandi, Rajeswari dan Rajaram pada tahun 2012 yang berjudul "*Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision tree algorithm*" yang menggabungkan K-Means dan C4.5 dengan menjadikan K-Means sebagai pembagi atau pembentuk partisi dari data latih untuk menghasilkan *k cluster*, lalu membentuk *decision tree* untuk melakukan *clustering*, dan berdasarkan tiap *cluster* akurasi yang dihasilkan sebesar 95.8% dengan KDD99 sebagai dataset yang digunakan.

Mengklasifikasikan konten atau topik semantik, merupakan permasalahan komputasi populer yang terkait dengan *information retrieval*, pemrosesan bahasa alami, dan *machine learning* secara luas. Klasifikasi berita dalam praktiknya dapat menandai artikel-artikel berita *online* dan mengumpulkannya berdasarkan topik, serta memberikan dasar untuk sistem rekomendasi berita yang akan sangat berguna untuk pembaca. Lebih luas lagi, mengingat dampak berita dan media terhadap sosial dan politik yang besar, dapat dilakukan analisis data berita secara terprogram untuk mengungkap pola dan bias

dalam produksi berita (Chase et al., 2014). Untuk mengetahui kualitas dari suatu algoritma, secara umum pada penelitian ini dilihat melalui dua aspek utama yaitu efektivitas dan efisiensi yang diukur berdasarkan kinerja algoritma dalam menangani dataset yang diberikan. Pada penelitian ini, kinerja dari algoritma C4.5 dan Naïve Bayes akan dilihat berdasarkan nilai akurasi dan *running time* dari masing-masing algoritma. Nilai tersebut akan didapatkan dengan melakukan perhitungan dengan metode *k-fold cross validation* dengan *k*=10. Data yang digunakan dalam penelitian ini adalah merupakan artikel berita yang diambil dari BBC *News* yang merupakan salah satu media penyedia berita yang terpercaya di dunia.

1.3 Perumusan Masalah

Rumusan masalah dalam penelitian ini adalah bagaimana perbandingan kinerja algoritma C4.5 dan *Naïve Bayes* dalam melakukan klasifikasi terhadap teks berita.

1.4 Tujuan Penelitian

Tujuan Penelitian ini adalah :

1. Menganalisa kinerja metode C4.5 dan *Naïve Bayes* dalam melakukan klasifikasi teks berita
2. Mengetahui algoritma yang lebih tepat untuk digunakan dalam klasifikasi teks berita

1.5 Manfaat Penelitian

Manfaat yang didapat dari penelitian ini adalah:

1. Peneliti dapat mengetahui algoritma mana yang memiliki akurasi dan running time yang lebih tinggi di antara C4.5 dan *Naïve Bayes* berdasarkan kinerja masing-masing algoritma
2. Dapat digunakan sebagai referensi dalam melakukan penelitian terhadap C4.5 dan *Naïve Bayes* dalam menangani data teks

1.6 Batasan Masalah

Batasan Masalah yang ditetapkan dalam penelitian ini adalah:

1. Jenis data yang digunakan adalah teks berita yang diambil dari website BBC news dengan format *.txt
2. Data yang digunakan adalah teks dalam bahasa inggris
3. Jumlah artikel untuk dataset adalah 500 artikel

1.7 Sistematika Penulisan

Sistematika penulisan pada penelitian ini adalah sebagai berikut :

BAB I. PENDAHULUAN

Pada bab ini diuraikan mengenai latar belakang masalah, perumusan masalah, tujuan, dan manfaat penelitian, batasan masalah, dan sistematika penulisan.

BAB II. KAJIAN LITERATUR

Pada bab ini akan dibahas dasar-dasar teori yang digunakan dalam penelitian, seperti definisi-definisi analisis sentimen, praproses, TF-IDF,

pohon keputusan, algoritma C4.5, Naïve Bayes, *cross validation*, *confusion matrix*, dan *rational unified process*. Pada akhir bab akan disertakan penelitian-penelitian lain yang relevan dengan penelitian ini.

BAB III. METODOLOGI PENELITIAN

Pada bab ini diuraikan mengenai tahapan yang akan dilaksanakan pada penelitian ini. Masing-masing rencana tahapan penelitian dideskripsikan dengan rinci dengan mengacu pada suatu kerangka kerja. Pada akhir bab ini berisi perancangan manajemen proyek pada pelaksanaan penelitian.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Pada bab ini akan dibahas mengenai perancangan dan implementasi perangkat lunak dengan metode pemrograman berorientasi objek berdasarkan panduan *rational unified process* yang di dalamnya terdapat fase insepsi, elaborasi, konstruksi, dan transisi.

BAB V. HASIL DAN ANALISIS PENELITIAN

Pada bab ini akan dibahas mengenai hasil klasifikasi algoritma C4.5 dan Naïve Bayes. Pada akhir bab ini berisi analisis dari hasil yang telah didapatkan.

BAB VI. KESIMPULAN DAN SARAN

Pada bab ini akan menjelaskan kesimpulan dan saran berdasarkan hasil analisis dalam membandingkan akurasi algoritma C4.5 dengan Naïve Bayes pada klasifikasi teks berita.

1.8 Kesimpulan

Berdasarkan uraian di atas, pada penelitian ini akan dilakukan klasifikasi untuk melihat akurasi algoritma C4.5 serta Naïve Bayes pada teks berita dengan batasan masalah yang telah ditentukan.

DAFTAR PUSTAKA

Anjali, Jivani, G., & Anjali, M. (2007). A Comparative Study of *Stemming* Algorithms. *October*, *2*(2004), 1930–1938. https://doi.org/10.1.1.642.7100

Beel, Joeran and Langer, Stefan and Gipp, B. (2017). TF-IDuF: A Novel Term-Weighting Sheme for User Modeling based on Users' Personal Document Collections. *Proceedings of the IConference 2017*, 1–7. Retrieved from http://mr-dlib.org

Feng, G., Guo, J., Jing, B. Y., & Sun, T. (2015). Feature subset selection using naive Bayes for text classification. *Pattern Recognition Letters*, *65*, 109–115. https://doi.org/10.1016/j.patrec.2015.07.028

HSSINA, B., MERBOUHA, A., EZZIKOURI, H., & ERRITALI, M. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13–19. https://doi.org/10.14569/SpecialIssue.2014.040203

Hussien, M. I., Olayah, F., Al-dwan, M., & Shamsan, A. (2011). Arabic Text Classification Using Smo , *Naïve Bayes*ian , J48 Algorithms, *9*(November), 306–316.

Karniol-tambour, O. (n.d.). Learning Multi-Label Topic Classification of News Articles, 1–6. Retrieved from http://cs229.stanford.edu/proj2013/ChaseGenainKarniolTambour LearningMulti-LabelTopicClassificationofNewsArticles.pdf

Kaur, G., & Bajaj, K. (2016). News Classification and Its Techniques: A Review. *IOSR Journal of Computer Engineering Ver. III*, *18*(1), 22–26. https://doi.org/10.9790/0661-18132226

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 1–7. https://doi.org/10.1109/ICCCNT.2013.6726842

Kohavi, R., & Sahami, M. (1996). Error-Based and Entropy-Based Discretization of Continuous Features. *Journal of Microscopy*, *237*(3), 487–496. https://doi.org/10.1.1.80.2847

Kumara, R., & Supriyanto, C. (2014). Klasifikasi Data Mining Untuk Penerimaan Seleksi Calon Pegawai Negeri Sipil 2014 Menggunakan Algoritma Decision Tree C4.5, 1–10.

Masetic, Z., Subasi, A., & Azemovic, J. (2016). Southeast Europe Journal of Soft Computing Available online : http://scjournal.ius.edu.ba Malicious Web Sites Detection using C4 . 5 Decision Tree, *5*(1).

Muniyandi, A. P., Rajeswari, R., & Rajaram, R. (2012). Network anomaly detection by cascading k-Means clustering and C4.5 decision tree algorithm. *Procedia Engineering*, *30*(2011), 174–182. https://doi.org/10.1016/j.proeng.2012.01.849

Ngoc, P. V., Ngoc, C. V. T., Ngoc, T. V. T., & Duy, D. N. (2017). A C4.5 algorithm for english emotional classification. *Evolving Systems*, *0*(0), 0. https://doi.org/10.1007/s12530-017-9180-1

Patil, T. R. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications, ISSN: 0974-1011*, *6*(2), 256–261. https://doi.org/ISSN: 0974-1011

Wong, T., & Yang, N. (2017). Dependency Analysis of Accuracy Estimates in k-fold Cross Validation, 4347(c), 1–12. https://doi.org/10.1109/TKDE.2017.2740926

Pressman, R. S. (2009). *Software Quality Engineering: A Practitioner's Approach*. (F. M. Schilling, Ed.), *Software Quality Engineering: A Practitioner's Approach* (7th ed., Vol. 9781118592). New York: McGraw-Hill. https://doi.org/10.1002/9781118830208

Quinlan, J. R. (1993). J. Ross Quinlan_C4.5_ Programs for Machine Learning.pdf. *Morgan Kaufmann*. https://doi.org/10.1007/BF00993309

Sebastiani, F. (2005). Text Categorization. *Text Mining and Its Applications to Intelligence, CRM and Knowledge Management*, (Ml), 109–123. https://doi.org/10.1.1.105.1540

Singh, S., & Gupta, P. (2014). Comparative study ID3, cart and C4 . 5 Decision tree algorithm: a survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, *27*(27), 97–103. https://doi.org/10.15693/ijaist/2014.v3i7.47-52

Sumathi, S., & Esakkirajan, S. (2007). *Fundamentals of Relational Database Management Systems* (Vol. 47). https://doi.org/10.1007/978-3-540-48399-1

Taheri, S., Mammadov, M., & Bagirov, A. M. (2010). Improving Naive Bayes *Classifier* Using Conditional Probabilities, 63–68.

Wulandini, F., Nugroho, A. S., & Categorization, A. T. (2009). Full-Text.