Pengaruh *N-Gram* pada Klasifikasi Dokumen menggunakan Algoritma *Naïve Bayes Classifier*

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika

Oleh :

Fitria Khoirunnisa
09021181520130

JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2019

LEMBAR PENGESAHAN SKRIPSI

PENGARUH *N-GRAM* PADA KLASIFIKASI DOKUMEN MENGGUNAKAN ALGORITMA *NAÏVE BAYES CALSSIFIER*

Oleh :

FITRIA KHOIRUNNISA
NIM : 09021181520130

Palembang, Desember 2019

Pembimbing I,

Novi Yusliani, M.T.
NIP. 198211082012122001

Pembimbing II,

Desty Rodiah, M.T.
NIK. 1671016112890005

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, MT
NIP. 197706012009121004

TANDA LULUS UJIAN SIDANG SKRIPSI

Pada hari Kamis tanggal 26 Desember 2019 telah dilaksanakan ujian sidang skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya

Nama	:	Fitria Khoirunnisa
NIM	:	09021181520130
Judul	:	Pengaruh *N-Gram* pada Klasifikasi Dokumen Menggunakan Algoritma *Naïve Bayes Classifier*

1. Pembimbing I

   Novi Yusliani, M.T.
   NIP. 198211082012122001 .....

2. Pembimbing II

   Desty Rodiah, M.T.
   NIK. 1671016112890005 .....

3. Penguji I

   Yunita, M.Cs
   NIP. 198306062015042002 .....

4. Penguji II

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

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009122004

HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini :

Nama	:	Fitria Khoirunnisa
NIM	:	09021181520130
Program Studi	:	Teknik Informatika
Judul Skripsi	:	Pengaruh *N-Gram* Pada Klasifikasi Dokumen Menggunakan Algortima *Naïve Bayes Classifier*

Hasil Pengecekan Software *iThenticate/Turnitin* : 14%

Menyatakan bahwa Laporan Proyek 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, Desember 2019

Fitria Khoirunnisa
NIM. 09021181520130

MOTTO DAN PERSEMBAHAN

Motto :

- "Hai orang-orang yang beriman, jadikanlah sabar dan shalat sebagai penolongmu, sesungguhnya Allah beserta orang-orang yang sabar"
  (QS. Al-Baqarah : 153)

- "Sesungguhnya bersama kesulitan ada kemudahan. Maka apabila engkau telah selesai (dari sesuatu urusan), tetaplah bekerja keras (untuk urusan yang lain). Dan hanya kepada Tuhanmulah engkau berharap"
  (QS : Al-Insyirah : 6-8)

Kupersembahkan karya tulis ini kepada :

- Allah Subhanahuwa ta'ala
- Kedua orang tua tercinta
- Ketiga adik tercinta
- Keluarga besarku
- Sahabat dan teman seperjuangan
- Seluruh dosen dan staff pengajar
- Almamater Universitas Sriwijaya

THE EFFECT OF N-GRAM ON DOCUMENT CLASSIFICATION USING NAÏVE BAYES CLASSIFIER ALGORITHM

By :
Fitria Khoirunnisa
09021181520130

ABSTRACT

News has become a major need for everyone, with news we can get the information needed. News can be distributed in the form of print mass media, electronic mass media and online media. The means of spreading the news now has grown very rapidly, making the amount of information being managed more and more and word management classified as not small. Therefore, we need a system in the classification of documents that are not structured. In this study, word processing in a document is done by N-Gram as a feature generation. The document classification process is carried out using the Naïve Bayes Classifier algorithm. This study examines the effect of N-Gram on document classification using the Naïve Bayes Classifier algorithm. The results of the classification accuracy of documents by applying N-Gram of 32.68% and without applying N-Gram of 84.97%. A decrease in the classification results occurs the number of features that result from solving N-Gram that is unique or dominant to another category. The accuracy of the results obtained shows that the application of N-Gram in the classification of documents using the Naïve Bayes Classifier algorithm gives a decreased effect on the performance of the classification.

Key Word : N-Gram, Naïve Bayes Classifier, Text Mining

PENGARUH *N-GRAM* PADA KLASIFIKASI DOKUMEN MENGGUNAKAN ALGORITMA *NAÏVE BAYES CALSSIFIER*

Oleh :
Fitria Khoirunnisa
09021181520130

ABSTRAK

Berita sudah menjadi kebutuhan utama bagi setiap orang, dengan berita kita dapat memproleh informasi yang dibutuhkan. Berita dapat disebar dalam bentuk media massa cetak, media massa elektronik dan media online. Sarana penyebaran berita sekarang ini telah berkembang sangat pesat, membuat jumlah informasi yang dikelola semakin banyak dan pengelolahan kata tergolong tidak sedikit. Oleh karena itu, dibutuhkan suatu sistem dalam klasifikasi dokumen yang tidak terstruktur. Pada penelitian ini, pemrosesan kata dalam suatu dokumen dilakukan dengan *N-Gram* sebagai pembangkitan fitur. Proses klasifikasi dokumen dilakukan dengan menggunakan algoritma *Naïve Bayes Classifier*. Penelitian ini menguji pengaruh *N-Gram* pada klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*. Hasil akurasi klasifikasi dokumen dengan menerapkan *N-Gram* sebesar 32,68% dan tanpa menerapkan *N-Gram* sebesar 84,97%. Penurunan hasil klasifikasi terjadi banyaknya fitur yang dihasilkan dari pemecahan *N-Gram* yang unik atau dominan ke kategori lain. Dari hasil akurasi yang didapatkan menunjukkan bahwa penerapan *N-Gram* pada klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier* memberikan pengaruh yang menurun terhadap kinerja klasifikasi.

Kata Kunci : *N-Gram, Naïve Bayes Classifier, Text Mining*

KATA PENGANTAR

Puji dan syukur atas kehadirat Allah ta'ala atas segala nikmat, rahmat, dan karunia-Nya sehingga penulis dapat menyelesaikan Tugas Akhir ini. Tugas Akhir yang berjudul "**Pengaruh *N-Gram* pada Klasifikasi Dokumen menggunakan Algoritma *Naïve Bayes Classifier***" disusun untuk memenuhi salah satu persyaratan kelulusan tingkat sarjana (S1) pada Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Pada kesempatan ini, penulis mengucapkan banyak terima kasih kepada pihak-pihak yang telah memberikan bantuan dan dukungan baik secara langsung maupun secara tidak langsung dalam penyelesaian Tugas Akhir ini, yaitu :

1. Bapak Joharudin dan ibu Sobatini selaku orang tua penulis, Andre Cevtiansyah, Iffat Vajriansyah, dan Siti Ai'syah selaku saudara penulis serta keluarga besar yang telah memberikan dukungan baik moril maupun materil kepada penulis.
2. Bapak Jaidan Jauhari, M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya, Bapak Rifkie Primartha, M.T. selaku Ketua Jurusan Teknik Informatika Universitas Sriwijaya dan Ibu Hardini Novianti, M.T. selaku Sekretaris Jurusan Teknik Informatika Universitas Sriwijaya.
3. Bapak Hadipurnawan Satria, Ph.D dan Ibu Rifka Widyastuti S.Kom., M.Ti, M.Im., selaku pembimbing akademik, yang telah membimbing, mengarahkan dan memberikan motivasi penulis dalam proses perkuliahan dan pengerjaan Tugas Akhir.
4. Ibu Novi Yusliani, M.T. selaku pembimbing I Tugas Akhir dan ibu Desty Rodiah, M.T. selaku pembimbing II Tugas Akhir yang telah memberikan pengarahan, bimbingan, bantuan, serta masukan kepada penulis sehingga Tugas Akhir ini dapat diselesaikan dengan baik.
5. Ibu Yunita, M.Cs. selaku penguji I Tugas Akhir dan Bapak Kanda Januar Miraswan, M.T. selaku penguji II Tugas Akhir yang telah memberikan saran dan kritik serta arahan selama penulisan Tugas Akhir.

6. Seluruh bapak dan ibu Dosen Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya yang telah memberikan ilmu pengetahuan kepada penulis selama perkuliahan.
7. Seluruh karyawan dan karyawati Fakultas Ilmu Komputer Universitas Sriwijaya yang telah membantu penulis dalam urusan administrasi selama kegiatan akademik.
8. Mbak Mida dan Lupina sahabat penulis dari awal perkuliahan yang selalu ada sebagai tempat berbagi cerita, curhat, memberi semangat, do'a dan dukungan kepada penulis.
9. Keluarga Besar Asrama Palembang dan Rumah Kosan43 Bapak Feri, Ibu Linda, Septi, Sandira, Alma, Okta, Yulik, Teteh, Amel, Indah, Mak Tika, Marina, dan Dolor-Dolor lainnya yang selalu memberikan asupan semangat dan doa kepada penulis.
10. Teman seperjuangan penulis seluruh mahasiswa/mahasiswi BEM, DPM, FASCO, group bimbingan TA dan Teknik Informatika Reguler/Billingual terutama Reguler Angkatan 2015.
11. Seluruh pihak yang telah membantu dalam penyusuan dan penyempurnaan tugas akhir ini yang tidak dapat disebutkan satu persatu.

Penulis menyadari bahwa Tugas Akhir ini jauh dari kata sempurna. Untuk itu, penulis mengharapkan kritik dan sarab yang membangun untun kesempurnaan Tugas Akhir ini dimasa mendatang, harapannya semoga Tugas Akhir ini dapat bermanfaat untuk semua.

Palembang, Desember 2019

Penyusun,

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

Halaman

BAB I

PENDAHULUAN

1.1 Pendahuluan

Bab ini membahas latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian dan batasan masalah yang menjadi dasar dalam penelitian ini serta memberikan penjelasan umum mengenai keseluruhan penelitian. Pendahuluan dimulai dengan penjelasan singkat tentang dokumen yang diteliti. Selanjutnya, bab ini akan membahas kebutuhan dan tantangan dalam melakukan klasifikasi dokumen berdasarkan kategori yang menjadi latar belakang masalah penelitian ini, serta pembangkit fitur sebagai solusinya.

1.2 Latar Belakang Masalah

Berita sudah menjadi kebutuhan utama bagi setiap orang. Hal ini dikarenakan berita dapat menghasilkan informasi yang dibutuhkan setiap orang. Dengan adanya berita, suatu informasi tentang fakta atau sedang terjadi dapat diketahui. Berita adalah bentuk laporan tentang suatu kejadian yang sedang terjadi baru-baru ini atau keterangan terbaru dari suatu peristiwa (Juditha, 2013). Berita dapat disebar dalam bentuk media massa cetak seperti surat kabar, *tabloid*, *newsletter*, majalah, *bulletin*, dan buku. Media massa elektronik seperti radio, televisi, dan film. Maupun media online yakni website internet yang berisikan informasi aktual layaknya media cetak.

Sarana penyebaran berita sekarang ini telah berkembang sangat pesat, membuat jumlah informasi yang disebar semakin meningkat. Maka jumlah informasi yang dikelola semakin banyak dan akan membuat pengelolahan kata tergolong tidak sedikit. Sehingga, pengelolahan kata yang dibutuhkan dalam mengategorikan berita perlu mengetahui inti dari isi dokumen. Proses yang dipakai dalam mengategorikan dokumen yang tidak terstruktur adalah proses klasifikasi dokumen dengan algoritma *Naïve Bayes Classifier* (Indriani, 2014).

Algoritma *Naïve Bayes Classifier* adalah algoritma yang digunakan dalam penelitian *text mining*. Algoritma *Naïve Bayes Classifier* merupakan metode klasifikasi berdasarkan probabilitas dan *Teorema Bayesian*. Tahapan klasifikasi ditentukan berdasarkan nilai kategori dari suatu dokumen yakni *term* yang muncul dalam dokumen yang diklasifikasi (Hamzah, 2012).

Salah satu kelebihan algoritma *Naïve Bayes Classifier* dibandingkan algoritma lain menurut Setiawan dan Nursantika (2017) adalah untuk mengetahui kategori pada sebuah dokumen dengan pengukuran kemiripan dokumen terkait yakni melalui proses pengenalan teks dan dokumen. Algoritma *Naïve Bayes Classifier* merupakan klasifikasi statistik yang bisa memprediksi probabilitas sebuah kelas dan tingkat akurasi yang tinggi. Menurut Narayanan, Arora, dan Bhatia (2013) mengatakan algoritma *Naïve Bayes Classifier* dapat meningkatkan hasil akurasi klasifikasi dan menghilangkan *noise* dengan cara memilih jenis fitur yang tepat dan sesuai. Salah satu model yang tepat dalam pengelolahan data berupa teks di *data mining* dan pemrosesan kata (Kumar, Sc, Phil, & Ph, 2018) untuk membangkitkan fitur adalah *N-Gram* (Sugianto, Liliana, dan Rostianingsih, 2013).

N-Gram adalah model probabilistik yang dapat digunakan untuk membangkitkan karakter dan kata serta memprediksi kata berikutnya dalam urutan kata tertentu. Dalam pembangkitan karakter, *N-Gram* terdiri dari *substring* sepanjang *n* karakter dari sebuah string. *N-Gram* digunakan untuk mengambil potongan-potongan karakter huruf sejumlah *n* dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen (Sugianto et al., 2013). Semakin besar nilai n dari sebuah kata maka berbanding terbalik dengan jumlah frekuensi keluar yang didapat, yaitu semakin kecil atau lebih jarang keluar. Penggunaan model Bi-Gram dan Tri-Gram untuk model bahasa masih memungkinkan, karena hasil dari jumlah frekuensi keluar pada suku *N-Gram*-nya masih cukup besar dan datanya masih valid apabila diproses lebih lanjut (Rostianingsih, Sugianto, dan Pustaka, 2012).

Penelitian yang menerapkan proses *N-Gram* adalah Pang, Lee, Rd, dan Jose (2002). *N-Gram* yang digunakan yakni Uni-Gram, yang dipakai dalam melakukan klasifikasi sentimen terhadap *review* film dengan menggunakan berbagai teknik pembelajaran mesin, dimana hasil penelitian ini menunjukkan bahwa metode *N-Gram* sebagai pendekatan untuk melakukan ekstraksi fitur. Dimana hasil klasifikasi dengan Uni-Gram dalam penelitian ini menghasilkan akurasi sebanyak 82,9%. Penelitian selanjutnya menerapkan prediksi kata untuk membantu mempercepat proses pengetikan dengan menerapkan *N-Gram* adalah Sugianto et al. (2013). Hasil penelitian ini menunjukkan bahwa *N-Gram* sebagai metode dasar dalam proses prediksi sangatlah membantu pemilahan kata. Sehingga proses prediksi menjadi

lebih efektif, mampu menghasilkan prediksi efektif di atas 20% dari total prediksi yang terjadi.

Berdasarkan dari hasil uraian tersebut, maka penelitian ini akan menguji pengaruh *N-Gram* pada hasil akurasi klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*.

1.3 Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah bagaimana pengaruh *N-Gram* pada klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*. Untuk menyelesaikan permasalahan tersebut, maka penelitian ini dibagi dalam beberapa *research question* antara lain:

1. Bagaimana mekanisme *N-Gram* dalam membangkitkan fitur?
2. Bagaimana mekanisme algoritma *Naïve Bayes Classifier* dalam mengklasifikasikan dokumen?
3. Bagaimana hasil akurasi klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*?
4. Bagaimana pengaruh *N-Gram* pada klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*?

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mengetahui mekanisme *N-Gram* dalam membangkitkan fitur.

2. Mengetahui mekanisme algoritma *Naïve Bayes Classifier* dalam mengklasifikasikan dokumen.
3. Mengetahui hasil akurasi klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*.
4. Mengetahui pengaruh *N-Gram* pada klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*.

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah:

1. Memahami mekanisme pembangkitan fitur pada *N-Gram* dan algoritma *Naïve Bayes Classifier* dalam proses klasifikasi dokumen.
2. Hasil penelitian dapat digunakan sebagai rujukan penelitian dalam pengaruh *N-Gram* pada klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*.

1.6 Batasan Masalah

Adapun batasan masalah penelitian ini adalah sebagai berikut:

1. Metode *N-Gram* yang digunakan adalah *Bi-Gram.*
2. Dokumen yang digunakan merupakan dokumen berita online.
3. Berita online yang digunakan dalam klasifikasi yakni berita bahasa Indonesia.
4. Sumber data yang akan digunakan berasal dari situs http://news.kompas.com dan http://www.republika.co.id.

5. Sistem menerima masukkan berupa file dokumen bertipe teks dengan format teks (*.txt).
6. Banyak kata dalam setiap file dokumen yakni lebih kurang 300 buah kata.
7. Kategori terdiri dari 5 kategori yakni berita olahraga, berita ekonomi, berita kesehatan, berita teknologi, dan berita politik.

1.7 Sistematika Penulisan

Sistematika penulisan proposal skripsi ini adalah sebagai berikut:

BAB I. PENDAHULUAN

Bab I menguraikan latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, dan sistematika penulisan penelitian ini.

BAB II. KAJIAN LITERATUR

Bab II berisi landasan teori yang digunakan pada penelitian ini, antara lain definisi-definisi *text mining*, *text preprocessing*, *Naïve Bayes Classifier*, *N-Gram*, *term frequency*, berita dan *k-fold cross validation*. Selain itu bab II juga membahas penelitian-penelitian lain yang relevan dengan penelitian ini.

BAB III. METODOLOGI PENELITIAN

Bab III berisi pembahasan mengenai tahapan yang akan dilaksanakan pada penelitian ini. Rencana tahapan penelitian akan dideskripsikan dengan rinci dengan mengacu pada suatu kerangka kerja.

Di akhir bab III akan dijabarkan perancangan manajemen proyek untuk pelaksanaan penelitian ini.

1.8 Kesimpulan

Bab ini telah dibahas latar belakang masalah penelitian dalam melakukan klasifikasi dokumen. Karena itu, penelitian ini akan menguji pengaruh *N-Gram* pada klasifikasi dokumen menggunakan algoritma *Naïve Bayes Classifier*. Dan hasil klasifikasinya akan dibandingkan, yakni hasil pengujian pada algoritma *Naïve Bayes Classifier* dengan *N-Gram* dan algoritma *Naïve Bayes Classifier* tanpa *N-Gram* dalam klasifikasi dokumen.

DAFTAR PUSTAKA

Fathan Hidayatullah, A., & Rifqi Ma, M. (2016). Penerapan Text Mining dalam Klasifikasi Judul Skripsi. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 1907–5022.

Hamzah, A. (2012). Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis Amir. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III*, 3(2011), 269–277. https://doi.org/1979-911X

Indriani, A. (2014). Klasifikasi Data Forum dengan menggunakan Metode Naive Bayes Classifier. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 28-33. https://doi.org/10.1155/2011/172853

Juditha, C. (2013). News Accuracy in Online Journalism (News of Alleged Corruption The Constitutional Court in Detiknews). *Jurnal Pekomnas*. 16(3): 145–154.

Kumar, S. S., & Rajini, A. (2018). An Efficent Sentimental Analysis for Twitter Using Neural Network based on Rmsprop. *IOSR Journal of Engineering (IOSRJEN)*, (Iccids), 17-25.

Ginting, S., L., Br. & Trinanda, R. P. (2016). Teknik Data Mining Menggunakan Metode Bayes Classifier pada Aplikasi Perpustakaan. *Jurnal Teknologi dan Informasi (JATI)*. 3(2).

Maghfira, T. N., Cholissodin, I., & Widodo, A. W. (2017). Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 1(6): 498–506.

Mooney, J. 2006. CS 391L : Machine Learning Text Categorization. *Austin : University of Texas*.

Musthafa, A.. 2009. Klasifikasi Otomatis Dokumen Berita Kejadian Berbahasa Indonesia. Skripsi. Fakultas Sains dan Teknologi universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang

Narayanan, V., Arora, I., & Bhatia, A. (2013b). Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model. H. *Yin et al. (Eds.): IDEAL 2013*, LNCS 8206, pp. 194–201.

Sammut, C., & Webb, G. I. (2016). Encyclopedia of Machine Learning and Data Mining. S*pringer Publishing Company, Incorporated*. https://doi.org/10.1007/978-1-4899-7502-7

Setiawan, I., & Nursantika, D. (2017). Klasifikasi Artikel Berita Menggunakan

Metode Text Mining Dan Naive Bayes Classifier. *Prosiding SENIATI*. 1–6.

Sugianto, S. A., Liliana, L., & Rostianingsih, S. (2013). Pembuatan Aplikasi Predictive Text Menggunakan Metode N-gram-based. *Jurnal Infra*. 1(2). Retrieved from https://www.neliti.com/id/publications/105718/pembuatan-aplikasi-predictive-text-menggunakan-metode-n-gram-based

Wibowo, A. T., & Septiana, G. (2015). Pembobotan Fitur Ekstraksi Pada Peringkasan Teks Bahasa Indonesia Menggunakan Algoritma Genetika. *E-Proceeding of Engineering*. 2(2):6481–6489.

Wijaya, A. P. (2012). Klasifikasi Dokumen dengan Naive Bayes Classification (NBC) Untuk Mengatahui Konten. *Jurnal Datamining Indonesia*. 222D, 1–6.

Zaman, B., Hariyanti, E., Purwanti, E., & Bahasa, A. D. (2015). Sistem Deteksi Bahasa pada Dokumen menggunakan N-Gram. *Jurnal Multinetics*. 1(2): 21–26.