SISTEM KLASIFIKASI KATEGORI BERITA MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*

SKRIPSI

Oleh :
FAISAL BRILIANSYAH
NIM. 13650056

JURUSAN TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2020

SISTEM KLASIFIKASI KATEGORI BERITA MENGGUNAKAN METODE K-NEAREST NEIGHBOR

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh:
FAISAL BRILIANSYAH
NIM. 13650056

JURUSAN TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2020

LEMBAR PERSETUJUAN

SISTEM KLASIFIKASI BERITA MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*

SKRIPSI

Oleh :
FAISAL BRILIANSYAH
NIM. 13650056

Telah Diperiksa dan Disetujui untuk Diuji
Tanggal : 22 Mei 2020

Dosen Pembimbing I	Dosen Pembimbing II
Dr. Cahyo Crysdian	Fresy Nugroho, M.T
NIP. 19740424 200901 1 008	NIP. 19710722 201101 1 001

Mengetahui,
Ketua Jurusan Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 19740424 200901 1 008

LEMBAR PENGESAHAN

SISTEM KLASIFIKASI KATEGORI BERITA MENGGUNAKAN METODE K-NEAREST NEIGHBOR

SKRIPSI

Oleh :
FAISAL BRILIANSYAH
NIM. 13650056

Telah dipertahankan di Depan Dewan Penguji
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Pada Tanggal : 22 Mei 2020

Susunan Dewan Penguji			Tanda Tangan
Penguji Utama	:	Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )
Ketua Penguji	:	Ainatul Mardhiyah, M.CS NIP. 19860330 20160801 2 075	( )
Sekretaris Penguji	:	Dr. Cahyo Crysdian NIP. 19740424 200901 1 008	( )
Anggota Penguji	:	Fresy Nugroho, M.T NIP. 19710722 201101 1 001	( )

Mengetahui,
Ketua Jurusan Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 19740424 200901 1 008

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama : Faisal Briliansyah

NIM : 13650056

Fakultas/Jurusan : Sains dan Teknologi/Teknik Infomatika

Judul Skripsi : Sistem Klasifikasi Katgeori Berita Menggunakan Metode *K-Nearest Neighbor*

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Skripsi ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 22 Mei 2020
Yang membuat pernyataan,

Faisal Briliansyah
NIM. 13650056

HALAMAN MOTTO

"Belajar untuk hidup dan hidup untuk belajar"

"Allah mengangkat derajat orang-orang yang beriman diantara kalian dan orang-orang yang diberi ilmu"

(QS Al-Mujadalah ayat 11)

"*Give a person an idea, and you enrich their day. Teach a person how to learn, and they can enrich their entire life*."

(Jim Kwik)

"*If you can't explain it simply, you don't understand it well enough*."

(Albert Einstein)

HALAMAN PERSEMBAHAN

Alhamdulillahirobbil Alamin puji syukur ke hadirat Allah SWT, Shalawat serta salam senantiasa kepada nabi Muhammad SAW, yang telah mengantarkan kita dari zaman kegelapan menuju zaman terang benderang.

Terima kasih saya ucapkan kepada kedua orang tua saya, yang selalu mendukung saya dalam menempuh pendidikan S1 di kampus UIN Maulana Malik Ibrahim Malang selama 7 tahun ini.

Terima kasih juga saya ucapkan kepada bapak Dr. Cahyo Crysdian dan Fressy Nugroho, M.T selaku dosen pembimbing. Dan juga tidak lupa terima kasih saya ucapkan kepada segenap dosen informatika UIN Maulana Malik Ibrahim Malang. Semoga apa yang telah diajarkan dapat menjadi ilmu yang bermanfaat.

Terima kasih juga saya ucapkan kepada seluruh teman – teman mahasiswa yang telah memberi dukungan kepada saya. Semoga kita menjadi manusia yang berhasil menjalankan tugasnya di dalam kehidupan ini. Amin.

KATA PENGANTAR

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

Assalamu'alaikum Wr. Wb.

Puji syukur kepada Allah tuhan semesta alam, yang telah mengizinkan diselesaikannya penelitian yang berjudul "Sistem Klasifikasi Kategori berita Menggunakan Metode *K-Nearest Neighbor*." Semoga penelitian ini menjadi penelitian yang bermanfaat.

Penelitian ini tentu tidak dapat diselesaikan tanpa bantuan pihak lain.Oleh karena itu, penulis ingin memberikan ucapan terima kasih kepada :

1. Prof. Dr. Abdul Haris, M.Ag selaku Rektor Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang.
2. Dr. Sri Harini, M.Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang.
3. Dr. Cahyo Crysdian, selaku Ketua Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang serta Pembimbing I yang telah meluangkan waktu untuk membimbing, memberikan masukan dan nasihat kepada penulis hingga akhir penyusunan skripsi.
4. Fresy Nugroho, M.T, selaku Dosen Pembimbing II yang telah membimbing dalam penyusunan skripsi ini.
5. Dr. M. Amin Hariyadi, selaku Dosen wali yang memberikan semangat, motivasi dan saran untuk penulis.

6. Para staff laboran Fakultas Sains dan Teknologi yang telah bersedia melayani kegiatan akademik dan administrasi.
7. Kedua orang tua yang selalu memberikan segala dukungan kepada penulis.
8. Teman – teman yang telah memberikan dukungan dan semangat dalam menjalani pendidikan.
9. Semua pihak yang telah banyak membantu dalam penyusunan skripsi ini yang tidak bisa penulis sebutkan semuanya.

Demikian ucapan terima kasih kepada pihak – pihak yang telah disebutkan. Penulis berharap penelitian ini akan bermanfaat untuk segenap pembaca meskipun di dalamnya terdapat banyak sekali kekurangan, atas kekurangan tersebut penulis ucapkan maaf yang sebesar – besarnya.

Wassalamu'alaikum Wr. Wb.

Malang, 23 Juni 2020

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Briliansyah, Faisal. 2020. **Sistem Klasifikasi Kategori Berita Menggunakan Metode *K-Nearest Neighbor***. Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Pembimbing (I) Dr. Cahyo Crysdian. (II) Fresy Nugroho, M.T

Kata Kunci : *K-Nearest Neighbor*, Klasifikasi Teks, *text mining, web scrapping.*

Informasi menjadi kebutuhan pokok bagi setiap orang, namun tidak semua informasi yang ada dapat menjadi kebutuhan. Dipengaruhi oleh kemajuan teknologi internet sehingga informasi mengalami pelonjakan yang besar, karena kebutuhan masyarakat akan berita semakin meningkat setiap harinya, berbagai pihak mencoba untuk menyajikan sajian berita berbahasa Indonesia yang sesuai dengan kebutuhan masyarakat Indonesia secara cepat, tepat, akurat, dan terpercaya bagi para pembaca berita. Permasalahan yang muncul adalah penggunaan media digital dalam penyampaian informasi menyebabkan jumlah artikel berita digital yang dirilis oleh portal berita tiap harinya menjadi sangat banyak. Hal ini berdampak pada ketersediaan artikel berita yang jumlahnya sangat melimpah. Klasifikasi merupakan salah satu metode dalam text mining yang bertujuan untuk menentukan label atau kategori objek. Terdapat 4 kategori berita yang digunakan dalam penelitian, yaitu Olahraga, Politik, Ekonomi, dan Teknologi. Dalam mendapatkan kategori berita digunakan metode *K-Nearest Neighbor* untuk menentukan tetangga terdekat suatu artikel berita. Pada penelitian ini dilakukan penerapan algoritma tersebut dalam sistem klasifikasi suatu kategori berita menggunakan bantuan web scrapping untuk mengambil berita secara online dari portal berita. Dalam penelitian pengklasifikasian menjadi 4 kategori berita berhasil melakukan pengujian untuk mencari nilai perhitungan akurasi sebesar 74 %, *precision* 35 %, *recall* 35 %, *F-measure* 35%, dan *Specificity* 84 % dan UAC 60% sehingga diperoleh hasil dengan nilai buruk untuk melakukan klasifikasi kategori berita.

ABSTRACT

Briliansyah, Faisal. 2020. **News category Classification system using the K-Nearest Neighbor method**. Undergraduate Thesis. Informatics Engineering Department Faculty of Science and Technology State Islamic University of Malang.
Supervisor (I) Dr. Cahyo Crysdian. (II) Fresy Nugroho, M.T

Keywords : *K-Nearest Neighbor,text classification, text mining, web scrapping.*

Information becomes a basic requirement for everyone, but not all information can be a necessity. Influenced by the advancement of internet technology so that information experiences a big surge, Because the public's needs for news are increasing every day, various parties try to present Indonesian language news content that is suitable for the needs of the Indonesian people in a fast, accurate, accurate and reliable way for news readers . The problem that arises is the use of digital media in the delivery of information causing the number of digital news articles released by the news portal every day to be very large. This has an impact on the availability of abundant news articles. Classification is one method in text mining that aims to determine labels or categories of objects. There are 4 categories of news used in research, namely sports, politics, economics, and technology. In getting the categories of news used the method K-Nearest Neighbor to determine the closest neighbor of a news article. In this research conducted the application of such algorithms in the classification system of a news category using the help of web scrapping to retrieve news online from the news portal. In classifying research into 4 categories of news successfully tested to find an accuracy calculation value of 74%, precision 35%, recall 35%, F-measure 35%, and Specificity %84% and UAC 60% so that the result with bad value to classify the news category.

الملخص

بريليانسياه ، فايسال. ٢٠٢٠. **نظام تصنيف فئة الأخبار باستخدام الطريقة ك–نيرست نيفبور** . أطروحة جامعية. قسم هندسة المعلوماتية لكلية العلوم والتكنولوجيا في جامعة مولانا مالك إبراهيم الإسلامية الحكومية بمالانق.

المشرف : (١) دكتور كاهيو كريسديان . (٢) فرسي نوغروهو ، الماجستير.

الكلمات الرئيسية : ك–نيرست نيفبور ، تحليل النصوص ،تصنيف النص، تجريف على شبكة الإنترنت.

المعلومات هي حاجة أساسية للجميع ، ولكن لا يمكن أن تكون هناك حاجة إلى جميع المعلومات. تأثر بالتقدم في تكنولوجيا الإنترنت بحيث تشهد المعلومات طفرة هائلة ، لأن احتياجات المجتمع للأخبار تتزايد كل يوم ، تحاول أطراف مختلفة تقديم عروض الأخبار الإندونيسية التي تتوافق مع احتياجات الشعب الإندونيسي بسرعة وبدقة ودقة وموثوقية لقراء الأخبار . المشكلة التي تنشأ هي استخدام الوسائط الرقمية في توصيل المعلومات ، مما يؤدي إلى زيادة عدد المقالات الإخبارية الرقمية التي تصدرها البوابة الإخبارية كل يوم. هذا له تأثير على توافر المقالات الإخبارية الوفيرة للغاية. التصنيف هو أحد الأساليب في استخراج النص الذي يهدف إلى تحديد تسمية أو فئة الكائنات. هناك 4 فئات من الأخبار يتم استخدامها في الأبحاث ، وهي الرياضة والسياسة والاقتصاد والتكنولوجيا. في الحصول على فئة الأخبار ، يتم استخدام طريقة ك–نيرست نيفبور لتحديد أقرب جار لمقال إخباري. في هذا البحث ، يستخدم تطبيق الخوارزمية في نظام التصنيف لفئة إخبارية مساعدة إلغاء الويب لاسترداد الأخبار عبر الإنترنت من بوابة الأخبار. في بحث التصنيف إلى 4 فئات إخبارية تم اختبارها بنجاح لإيجاد قيمة حساب دقة 74٪ ، دقة 35٪ ، استدعاء 35٪ ،ف–قياس 35٪ ، خصوصية 84٪ و 60٪ UAC بحيث النتائج التي تم الحصول عليها بقيمة سيئة للقيام بالتصنيف فئة الأخبار

BAB I

PENDAHULUAN

1.1 Latar Belakang

Informasi mengenai berita-betita aktual yang terjadi setiap hari atau yang terjadi setiap menit saat ini bisa dengan mudah didapatkan seperti situs berita online yang sifatnya umum memuat berbagai informasi teraktual, maupun situs berita yang menampilkan rubrik secara khusus, misal tentang politik, ekonomi, pendidikan, olahraga dan lain sebagainya. Hal tersebut bisa didapatkan dengan membuka berbagai media online yang saat ini sangat beragam jenisnya (Bahri, 2010). Berita dapat diperoleh dari media cetak maupun elektronik seperti koran, televisi, radio, dan internet. Berita yang disajikan dalam bentuk teks pada media elektronik, biasanya dikelompokkan berdasarkan isinya seperti berita olahraga, ekonomi, sains, dan lain sebagainya.

Permasalahan yang muncul adalah penggunaan media digital dalam penyampaian informasi menyebabkan jumlah artikel berita digital yang dirilis oleh portal berita tiap harinya menjadi sangat banyak. Hal ini berdampak pada ketersediaan artikel berita yang jumlahnya sangat melimpah. Pada umumnya berita yang disampaikan dalam portal tersebut terdiri dari beberapa kategori seperti berita politik, olahraga, ekonomi, teknologi dan lain - lain (sebagai contoh pada website detik.com dan kompas.com). Namun, dalam membagi berita ke dalam kategori-kategori tersebut untuk saat ini masih dilakukan secara manual, artinya dalam mengunggah berita pengunggah harus terlebih dahulu mengetahui isi dari berita yang akan diunggah secara kesuluruhan untuk selanjutnya dimasukkan ke dalam

kategori yang tepat (Lin, 2014). Hal ini sangat merepotkan bagi para pengunggah berita apabila jumlah berita yang ingin diunggah berjumlah banyak.

Berdasarkan permasalahan tersebut diperlukan adanya pengorganisasian dokumen artikel berita. Salah satu cara yang dapat dilakukan dengan cepat dan dapat dipahami oleh para penerima informasi adalah dengan melakukan klasifikasi dokumen artikel berita berdasarkan kategorinya. Klasifikasi merupakan salah satu metode dalam *text mining* yang bertujuan untuk menentukan label atau kategori objek (Prasetyo, 2012). Sekumpulan objek yang mempunyai kesamaan fitur biasanya dikategorikan ke dalam kelompok tertentu dan kelompok itu kemudian diberi nama atau label.

Dalam ayat alquran surat Al-Hujarat ayat 6 Allah juga telah menegaskan untuk meneliti suatu berita dengan seteliti mungkin, ayat tersebut berbunyi :

يَا أَيُّهَا الَّذِينَ آمَنُوا إِنْ جَاءَكُمْ فَاسِقٌ بِنَبَإٍ فَتَبَيَّنُوا أَنْ تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْبِحُوا عَلَىٰ مَا فَعَلْتُمْ نَادِمِينَ

Artinya : “Hai orang-orang yang beriman, jika datang kepadamu orang fasik membawa suatu berita, maka periksalah dengan teliti agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya yang menyebabkan kamu menyesal atas perbuatanmu itu” (QS. Al- Hujurat:6).

Allah SWT memerintahkan (orang-orang percaya) untuk memeriksa dengan seksama pesan orang fasik, dan biarkan mereka berhati-hati dalam menerimanya dan tidak menerima begitu saja, yang akibatnya akan membalikkan kenyataan. Orang yang hanya menerima berita darinya, berarti sama dengan mengikuti jejaknya. Sedangkan Allah SWT telah melarang orang percaya untuk mengikuti jalan orang yang rusak.

Untuk mengetahui kategori pada sebuah artikel berita secara otomatis tanpa harus dikategorikan dibaca satu persatu maka perlu dilakukan pengukuran kemiripan dokumen terkait dengan menggunakan metode *K-Nearest Neighbor* yaitu metode untuk menghitung jarak terdekat dari dua buah objek kemudian mengelompokan objek yang berdekatan ke dalam satu kelas.

Web scraping merupakan teknik yang digunakan untuk mengekstrak sejumlah besar dokumen artikel berita dari situs web portal berita dimana data yang sudah diekstraksi disimpan ke sebuah file lokal di komputer atau ke database dalam format tabel (*spreadsheet*). Inilah yang memungkinkan user untuk mengeksplorasi isi dari situs web tanpa mengunjungi situs web yang bersangkutan, sehingga *user* bisa melakukan berbagai bentuk analisis semantik tanpa mengganggu *resource* situs web yang bersangkutan.

Sistem pencarian informasi juga telah diteliti oleh Purwanti (2015), kasus di kombinasikan dalam klasifikasi. Pembobotan kata menggunakan *Term Frequency Inverse Document Frequency* (TF-IDF). Dan perhitungan kemiripan antar dokumen jurnal menggunakan perhitungan *Cosine Similarity*, lalu hasil kesamaan tersebut di klasifikasi dengan mengunakan *K-Nearest Neighbor*.

Pada penelitian yang dilakukan oleh Johanes Widagdho dan Yodha, Achmad Wahid Kurniawan (2014), membuat penelitian dengan judul Pengenalan Motif Batik Menggunakan Deteksi Tepi Canny dan K-Nearest Neighbor. Salah satu budaya ciri khas Indonesia yang telah dikenal dunia adalah batik. Penelitian ini bertujuan untuk mengenali 6 jenis motif batik pada buku karangan H.Santosa Doellah yang berjudul “Batik: Pengaruh Zaman dan Lingkungan”. Proses

klasifikasi akan melalui 3 tahap yaitu preprosesing, feature extraction dan klasifikasi. Preproses mengubah citra warna batik menjadi citra grayscale. Pada tahap feature extraction citra grayscale ditingkatkan kontrasnya dengan histogram equalization dan kemudian menggunakan deteksi tepi Canny untuk memisahkan motif batik dengan backgroundnya dan untuk mendapatkan pola dari motif batik tersebut. Hasil ekstraksi kemudian dikelompokkan dan diberi label sesuai motifnya masing-masing dan kemudian diklasifikasikan menggunakan K-Nearest Neighbor menggunakan pencarian jarak Manhattan. Hasil uji coba diperoleh akurasi tertinggi mencapai 100% pada penggunaan data-testing sama dengan data training (dataset sebanyak 300 image). Pada penggunaan data training yang berbeda dengan data testing diperoleh akurasi tertinggi 66,67%. Kedua akurasi tersebut diperoleh dengan menggunakan lower threshold = 0.010 dan upper threshold = 0.115 dan menggunakan k=1.

Berdasarkan permasalahan dan penelitian sebelumnya maka penulis me nyusun laporan penelitian yang berjudul "Sistem Klasifikasi Kategori Berita Menggunakan Algoritma *K-Nearest Neighbor* (K-NN)".

1.2 Pernyataan Masalah

Berdasarkan latar belakang yang sebelumnya telah diterangkan, maka pernyataan masalah yaitu seberapa akurat klasifikasi berita menggunakan metode K-NN *(K-Nearest Neighbor)* ?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini ialah mengukur tingkat akurasi dari klasifikasi berita menggunakan *K-Nearest Neighbor*.

1.4 Manfaat Penelitian

Dengan dilakukannya penelitian ini diharapkan mampu memberikan manfaat secara keilmuan berupa kontribusi pada klasifikasi berita menggunakan metode *K-Nearest Neighbor*, sedangkan manfaat kepada masyarakat adalah memberikan kemudahan dalam proses klasifikasi berita terutama untuk penyedia media online di Indonesia.

1.5 Batasan Masalah

Berdasarkan rumusan masalah yang telah ditetapkan, maka batasan masalah penelitian ini adalah sebagai berikut.

1. *Dataset* berita pelatihan dari artikel berita bahasa Indonesia disimpan ke dalam format csv secara manual. Masukkan *query* dari *user* memakai Bahasa Indonesia.
2. Berita yang dibutuhkan untuk pelatihan sebanyak 400 data yang diambil dari portal berita *online* Detik dan Kompas dari bulan Januari 2019 sampai Desember 2019.
3. Penentuan kategori dokumen berita yang digunakan berupa olahraga, teknologi, ekonomi, dan politik.

BAB II

STUDI PUSTAKA

2.1 Data Mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. Dalam data mining terdapat dua pendekatan metode pelatihan, yaitu *Unsupervised learning*, metode ini diterapkan tanpa adanya latihan *(training)* dan tanpa ada guru *(teacher)*. Guru di sini adalah label dari data. *Supervised learning*, yaitu metode belajar dengan adanya latihan dan pelatih. Dalam pendekatan ini, untuk menemukan fungsi keputusan, fungsi pemisah atau fungsi regresi, digunakan beberapa contoh data yang mempunyai output atau label selama proses *training*.

Ada beberapa teknik yang dimiliki *data mining* berdasarkan tugas yang bisa dilakukan, setiap teknik memiliki algoritma masing-masing. Teknik dalam data mining terbagi menjadi enam kategori yaitu : Deskripsi, para peneliti biasanya mencoba menemukan cara untuk mendeskripsikan pola dan trend yang tersembunyi dalam data; Estimasi, mirip dengan klasifikasi kecuali variabel tujuan yang lebih kearah numerik dari pada kategori; Prediksi, memiliki kemiripan dengan estimasi dan klasifikasi. Hanya saja, prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi dimasa depan); Klasifikasi, dalam klasifikasi variabel, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan kategori berita dalam empat kelas, yaitu kategori berita olahraga, politik, ekonomi dan

teknologi; Klastering, lebih ke arah pengelompokan *record*, pengamatan, atau kasus dalam kelas yang memiliki kemiripan; Asosiasi, mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu.

2.2 Klasifikasi Teks

Klasifikasi teks merupakan proses menemukan pola baru yang belum terungkap sebelumnya. Klasifikasi teks dilakukan dengan memproses dan menganalisa data dalam jumlah besar. Dalam prosesnya, klasifikasi teks melibatkan struktur yang mungkin terdapat pada teks dan mengekstraks informasi yang relevan pada teks. Dalam menganalisis sebagian atau keseluruhan teks yang tidak terstruktur, klasifikasi teks mencoba mengasosiasikan sebagian atau keseluruhan satu bagian teks dengan yang lainnya berdasarkan aturan-aturan tertentu (Miller, 2015). Tantangan dari klasifikasi teks adalah sifat data yang tidak terstruktur dan sulit untuk menangani, sehingga diperlukan proses *text mining*. Diharapkan melalui proses *text mining*, informasi yang ada dapat dikeluarkan secara jelas di dalam teks tersebut dan dapat dipergunakan dalam proses analisis menggunakan alat bantu komputer (Witten dkk, 2016).

Tahapan praproses ini dilakukan agar dalam klasifikasi dapat diproses dengan baik. Tahapan dalam praproses teks adalah sebagai berikut:

a. *Case Folding*, merupakan proses untuk mengubah semua karakter pada teks menjadi huruf kecil. Karakter yang diproses hanya huruf ‘a’ hingga ‘z’ dan selain karakter tersebut akan dihilangkan seperti tanda baca titik (.), koma (,), dan angka. (Weiss, 2005).

b. *Tokenizing*, merupakan proses memecah yang semula berupa kalimat menjadi kata-kata atau memutus urutan string menjadi potongan-potongan seperti kata-kata berdasarkan tiap kata yang menyusunnya. Sehingga dapat dikatakan mengembalikan kata penghubung .

c. *Stopwords*, yakni kosakata yang bukan merupakan kata unik atau ciri pada suatu dokumen atau tidak menyampaikan pesan apapun secara signifikan pada kalimat (Dragut dkk, 2016). Kosakata yang dimaksudkan tersebut adalah kata penghubung dan kata keterangan yang bukan merupakan kata unik misalnya "sebuah", "oleh", "pada", dan sebagainya.

d. *Stemming*, yakni proses untuk mendapatkan kata dasar dengan cara meng-hilangkan awalan, akhiran, sisipan, dan kombinasi dari awalan dan akhiran.

2.3 *Nearest Neighbor*

Algoritma *Nearest Neighbor* adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru (*testing data*) dengan kasus lama (*training data*), yang didasarkan pada pencocokan bobot dari sejumlah fitur. Ada 2 jenis algoritma *Nearest Neighbor*, yaitu klasifikasi 1-NN dilakukan pada 1 data berlabel terdekat; K-NN, yaitu klasifikasi dilakukan pada data berlabel k terdekat dengan k.

K-Nearest Neighbor (K-NN) adalah metode yang menggunakan algoritma *supervised* dimana hasil dari query instance yang baru diklasifikasikan berdasarkan mayoritas label kelas pada K-NN. Tujuan dari algoritma K-NN adalah mengklasifikasikan objek baru berdasarkan *atribut* dan *training data*.

Purwanti (2015) melakukan penelitian mencari tahu bagaimana cara klasifikasi jurnal menggunakan penerapan sistem pencarian informasi. Data yang diteliti dalam bentuk jurnal, teknik *text mining* digunakan dalam konstruksi, kemudian dilakukan pembobotan setiap token, kemudian kesamaan antara dokumen dihitung menggunakan Vector Space Model (kesamaan *Cosine Similarity*), dan digunakan teknik *K-Nearest Neighbor* sebagai menentukan hasil klasifikasi dokumen. Penelitian tersebut berhasil melakukan klasifikasi sesuai dengan kategori dokumen yang diuji.

Hardiyanto et. al (2016), melakukan penelitian tentang Implementasi K - Nearest Neighbor (K-NN) pada Klasifikasi Artikel Wikipedia Indonesia. Suatu hal yang dibutuhkan seiring dengan perkembangan teknologi informasi dan komunikasi adalah informasi. Salah satu sumber informasi tersebut adalah Wikipedia Bahasa Indonesia. Banyaknya artikel yang masuk dalam beberapa kategori menyebabkan pembaca kesulitan dalam mencari informasi, terutama dalam pencarian berdasarkan kategori. Oleh karena itu diperlukan sebuah klasifikasi untuk artikel Wikipedia agar memiliki tepat satu kategori namun tetap dapat berhubungan dengan kategori lainnya. Diperlukan sistem yang dapat mengklasifikasi artikel Wikipedia Indonesia secara otomatis. Klasifikasi artikel Wikipedia Indonesia adalah sebuah sistem yang berfungsi untuk mengklasifikasi artikel Wikipedia Indonesia yang berupa dokumen teks dengan tahapan text preprocessing dilanjutkan dengan pembobotan TF IDF pada masing-masing artikel Wikipedia Indonesia terbentuk vektor kata. Berdasarkan pembobotan tersebut, artikel-artikel Wikipedia Indonesia tersebut diklasifikasikan dengan metode K

Nearest Neighbor. Perhitungan centroid pada masing-masing sub sub kategori terdiri dari tiga buah artikel yang diambil nilai tengahnya kemudian dihitung jarak kedekatan dengan masing-masing data uji. Berdasarkan hasil pengujian manual menunjukkan akurasi kebenaran sebesar 60%.

Implementasi Metode Klasifikasi K-Nearest Neighbor (K-NN) Untuk Pengenalan Pola Batik Motif Lampung oleh Naufal (2017), Batik merupakan nama terkenal dari suatu kain yang berasal dari pulau Jawa. Batik telah diakui sebagai salah satu Hak Kekayaan Intelektual Indonesia oleh UNESCO sejak 2 Oktober 2009. Seiring dengan perkembangan zaman, Batik telah berkembang ke seluruh nusantara dan menyebabkan banyak motif unik dan berbeda yang tercipta. Batik Lampung merupakan salah satunya. Penelitian ini membahas tentang klasifikasi motif (pola) Batik Lampung menggunakan metode K-Nearest Neighbor. Motif yang digunakan pada penelitian ini adalah Jung Agung, Siger Kembang Cengkih, Siger Ratu Agung dan Sembagi. Sampel gambar asli disimpan dalam RGB (Red Green Blue). Tahap pertama yaitu merubah ukuran gambar menjadi 50 x 50 pixel dan dikonversi menjadi keabu-abuan (Grayscale). Untuk mengenali ciri suatu gambar, digunakan metode Gray Level Co-Occurence Matrix (GLCM). Metode K-Nearest Neighbor pada penelitian ini menggunakan nilai k = 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29. Orientasi sudut yang digunakan 00 , 450 , 900 dan 1350 . Akurasi tertinggi didapatkan pada pengujian di orientasi arah sudut sebesar 450 di nilai k = 17 yaitu sebesar 98,182%.

Algoritma K-NN bekerja berdasarkan jarak terpendek dari *query instance* ke *training data* untuk menentukan K-NN. Salah satu cara untuk menghitung

jarak dekat atau jauhnya tetangga menggunakan metode *Euclidian Distance*, sering digunakan untuk menghitung jarak sehingga berfungsi menguji ukuran yang bisa digunakan sebagai interpretasi kedekatan jarak antara dua obyek. *Euclidian Distance* dirusmuskan dalam persamaan (2.1).

$$d_{ij} = \left(\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2\right)^{1/2} \quad (2.1)$$

Dimana: X_k = nilai X pada *training* data; X_{jk} = nilai X pada *testing* data ; m = batas jumlah banyaknya data; d_{ij}= jarak dari *training* ke *testing*.

Jika hasil nilai dari rumus di atas besar maka akan semakin jauh tingkat keserupaan antara kedua objek dan sebaliknya jika hasil nilainya semakin kecil maka akan semakin dekat tingkat keserupaan antar objek tersebut. Objek yang dimaksud adalah *training* data dan testing data.

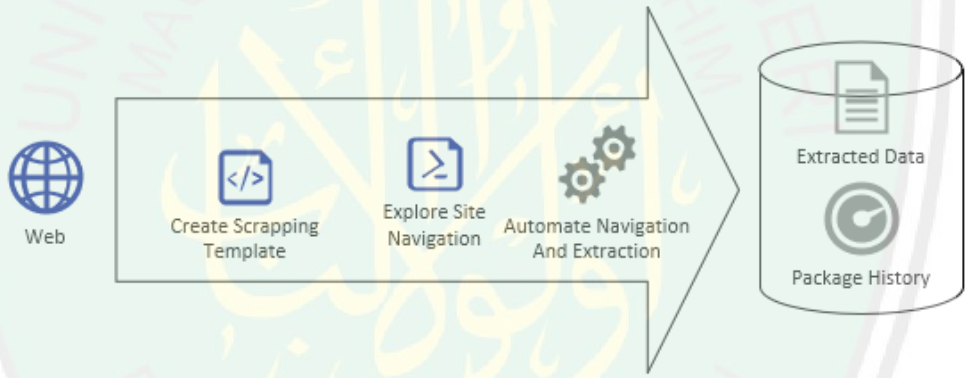
Dalam algoritma ini, nilai k yang terbaik itu tergantung pada jumlah data. Ukuran nilai k yang besar belum tentu menjadi nilai k yang terbaik begitupun juga sebaliknya.

Langkah-langkah untuk menghitung algoritma k-NN yaitu menentukan nilai k; menghitung kuadrat jarak *euclid* (*query instance*) masing-masing objek terhadap *training* data yang diberikan; kemudian mengurutkan objek-objek tersebut kedalam kelompok yang mempunyai jarak euclid terkecil; mengumpulkan label class Y (Klasifikasi *Nearest Neighbor*); dengan menggunakan kategori *Nearest Neighbor*

yang paling mayoritas maka dapat dipredeksikan nilai *query instance* yang telah dihitung.

2.4 Web Scraping

WebScraping Turland (2010) adalah proses pengambilan sebuah dokumensemi-terstruktur dari internet, umumnya berupa halaman-halaman web dalam bahasa markup seperti HTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain. *Webscraping* memiliki sejumlah langkah, sebagai berikut:



Gambar 2. 1 Langkah *Web Scraping*

Create Scraping Template: Pembuat program mempelajari dokumen HTML dari website yang akan diambil informasinya untuk tag HTML yang mengapit informasi yang akan diambil.

Explore Site Navigation: Pembuat program mempelajari teknik navigasi pada website yang akan diambil informasinya untuk ditirukan pada aplikasi *webscraper* yang akan dibuat.

Automate Navigation and Extraction: Berdasarkan informasi yang didapat pada langkah 1 dan 2 di atas, aplikasi web scraper dibuat untuk mengotomatisasi pengambilan informasi dari website yang ditentukan.

Extracted Data and Package History: Informasi yang didapat dari langkah 3 disimpan dalam tabel atau tabel-tabel database.

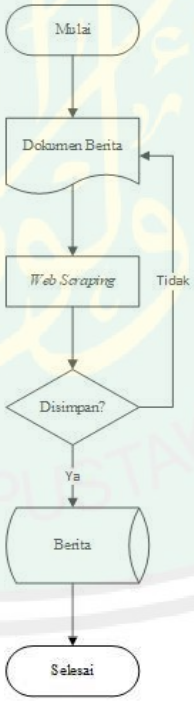
BAB III

METODE PENELITIAN

Dalam bab ini membahas implementasi sistem dan desain yang akan dibuat, dimana kebutuhan – kebutuhan apa yang dibutuhkan dalam membuat sistem klasifikasi kategori berita menggunakan metode *K-Nearest Neighbor*.

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini diambil dari portal berita *online* Detik dan Kompas. Data yang diperoleh merupakan kumpulan berita yang didapatkan dengan menggunakan *web scraping*.



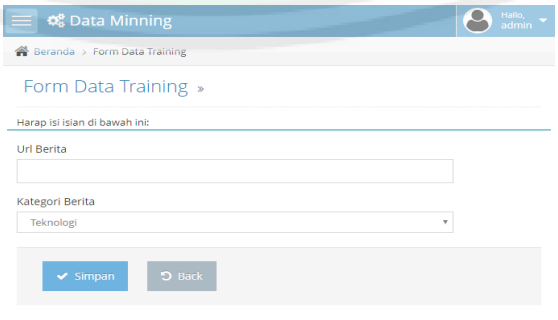
Gambar 3. 1 *Flowchart* Pengumpulan Data

Data yang digunakan dalam penelitian ini diambil dari portal berita Detik dan Kompas dengan menggunakan *web scraping* dan akan disimpan pada sistem kemudian dilakukan penyimpanan data dalam database MySQL.

```
public function getBerita($post){
        $client = new Client();
        $berita_isi = '';
        $crawler = $client->request('GET', $post['datatraining_url']);
        $crawler->filter('#detikdetailtext')->each(function ($node) {
                $berita_isi = $node->html();
                $this->dataMinning['berita_isi'] = $berita_isi;
        });
        // $no = 0;
        $crawler->filter('#detikdetailtext > strong')->each(function ($node) {
                $berita_lokasi = $node->html();
                if ($this->dataMinning['berita_lokasi'] == '') {
                        $this->dataMinning['berita_lokasi'] = $berita_lokasi;
                         // $no++;
                }
        });
        $crawler->filter('.jdl > h1')->each(function ($node) {
                $berita_judul = $node->text();
                $this->dataMinning['berita_judul'] = $berita_judul;
        });
}
```

Gambar 3. 2 Potongan *Source Code web scrapping*

Pada potongan program diatas berfungsi untuk scraping data training dimana dapat memasukkan url berita baru dan juga kategori beritanya. Berikut ini hasil implementasi data training.



Gambar 3. 3 Halaman *Scrapping Data Training*

3.2 Desain Sistem dan Implementasi Sistem

Pengimplementasian dilakukan menggunakan *Code Igniter* sebuah *framework* PHP yang dapat membantu mempercepat developer dalam pengembangan aplikasi web berbasis PHP dibandingkan dengan menulis semua kode dari awal. *Codeigniter* menyediakan banyak library untuk mengerjakan tugas-tugas yang umumnya ada pada sebuah aplikasi berbasis web (kadir, 2015). Desain sistem penelitian dijelaskan sebagai berikut :

1. Pengumpulan dokumen berita disimpan kedalam sistem, dari data yang diperoleh maka dilakukan preprosessing untuk menggambil kata dasar yang ada dalam setiap dokumen.
2. Pelabelan

   Pemberian nilai awal pada data atau kata biasa disebut *groundtruth* dilakukan untuk pelabelan adalah tahapan dimana berita diberi label yang nantinya akan digunakan pada proses *training* di tahap klasifikasi. Terdapat empat label yang disediakan yaitu politik untuk berita yang berisi informasi mengenai politik, olahraga untuk berita yang berisi informasi mengenai olahraga, teknologi untuk berita yang berisi informasi mengenai teknologi, dan ekonomi untuk berita yang hanya berisi informasi mengenai ekonomi. Hasil dari tahapan ini adalah kumpulan berita yang memiliki label.
3. Klasifikasi

   Proses klasifikasi dibedakan menjadi dua proses yaitu:

   a. *Training*, proses ini digunakan untuk melatih algoritma klasifikasi yang digunakan yaitu algoritma *K-Nearest Neighbor* agar mampu melakukan

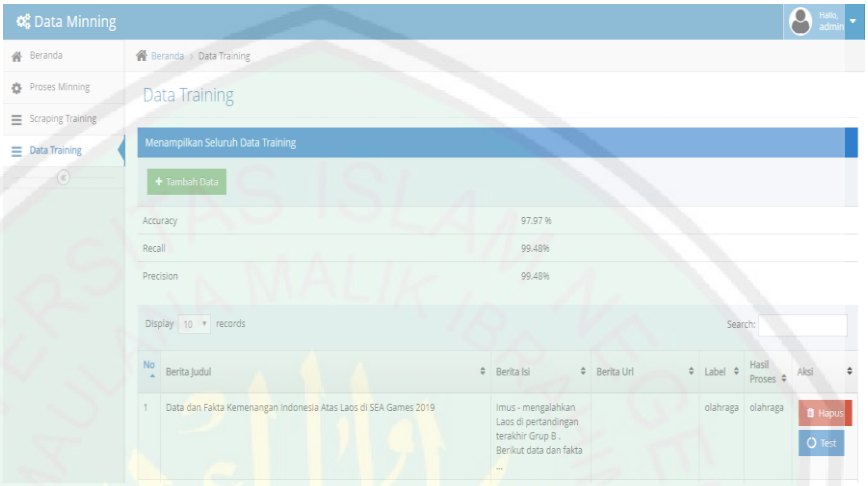
prosesnya sesuai dengan yang diharapkan. Pada tahap ini pertama-tama akan dilakukan proses pembobotan terhadap kumpulan berita hasil pelabelan menggunakan perhitungan TF-IDF. Pada tahap ini data *training* dalam metode K-NN berupa dataset disimpan dalam database tidak membentuk model klasifikasi dan menahan setiap *data training* dataset karena tidak ada pekerjaan yang dilakukan sampai prediksi diperlukan pada tahap testing, K-NN biasa disebut sebagai *Lazy Algorithm* (Algoritma Malas).

```
public function getBeritaKompas($post){
        $client = new Client();
        $berita_isi = '';
        $crawler = $client->request('GET', $post['datatraining_url']);
        $crawler->filter('.read__content')->each(function ($node) {
            $berita_isi = $node->html();
            $this->dataMinning['berita_isi'] = $berita_isi;
        });
        // $no = 0;
        $crawler->filter('.read__content > p > strong')->each(function
        ($node) {
            $berita_lokasi = $node->html();
            if ($this->dataMinning['berita_lokasi'] == '') {
                $this->dataMinning['berita_lokasi'] = $berita_lokasi;
                // $no++;
            }
        });
        $crawler->filter('.read__title')->each(function ($node) {
            $berita_judul = $node->text();
            $this->dataMinning['berita_judul'] = $berita_judul;
        });
}
```

Gambar 3. 4 Potongan *Source Code* Halaman *Data Training*

Pada potongan program diatas berfungsi untuk menampilkan data

training, akurasi, recall dan precision. Berikut ini hasil implementasi data training.



Gambar 3. 5 Halaman Data Training

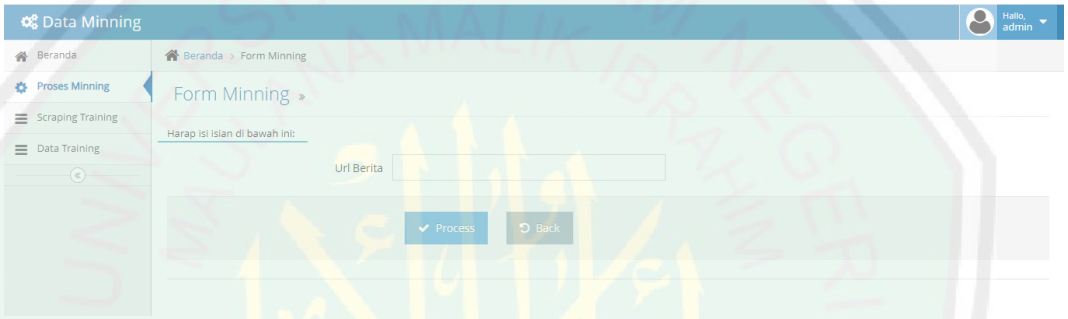
b. *Testing*, proses ini dilakukan dengan memasukkan berita baru yang belum melalui proses *training* sebagai bentuk untuk melakukan pengklasifikasian terhadap dataset dengan memanfaatkan model klasifikasi yang dihasilkan pada proses *training.*

```
class M_minning extends CI_Model
{

    function __construct()
    {
        parent::__construct();
    }

    public function getBerita(){
        $this->db->select('berita_id,berita_judul');
        $this->db->from('tbl_berita');
        $query = $this->db->get();
        $result = $query->result();
        return $result;
    }

    public function getBeritaDetail($id){
        $this->db->select('*');
        $this->db->from('tbl_berita');
        $this->db->where('berita_id' , $id);
```

```
            $query = $this->db->get();
            $result = $query->row();
            return $result;
        }
    }
}
```

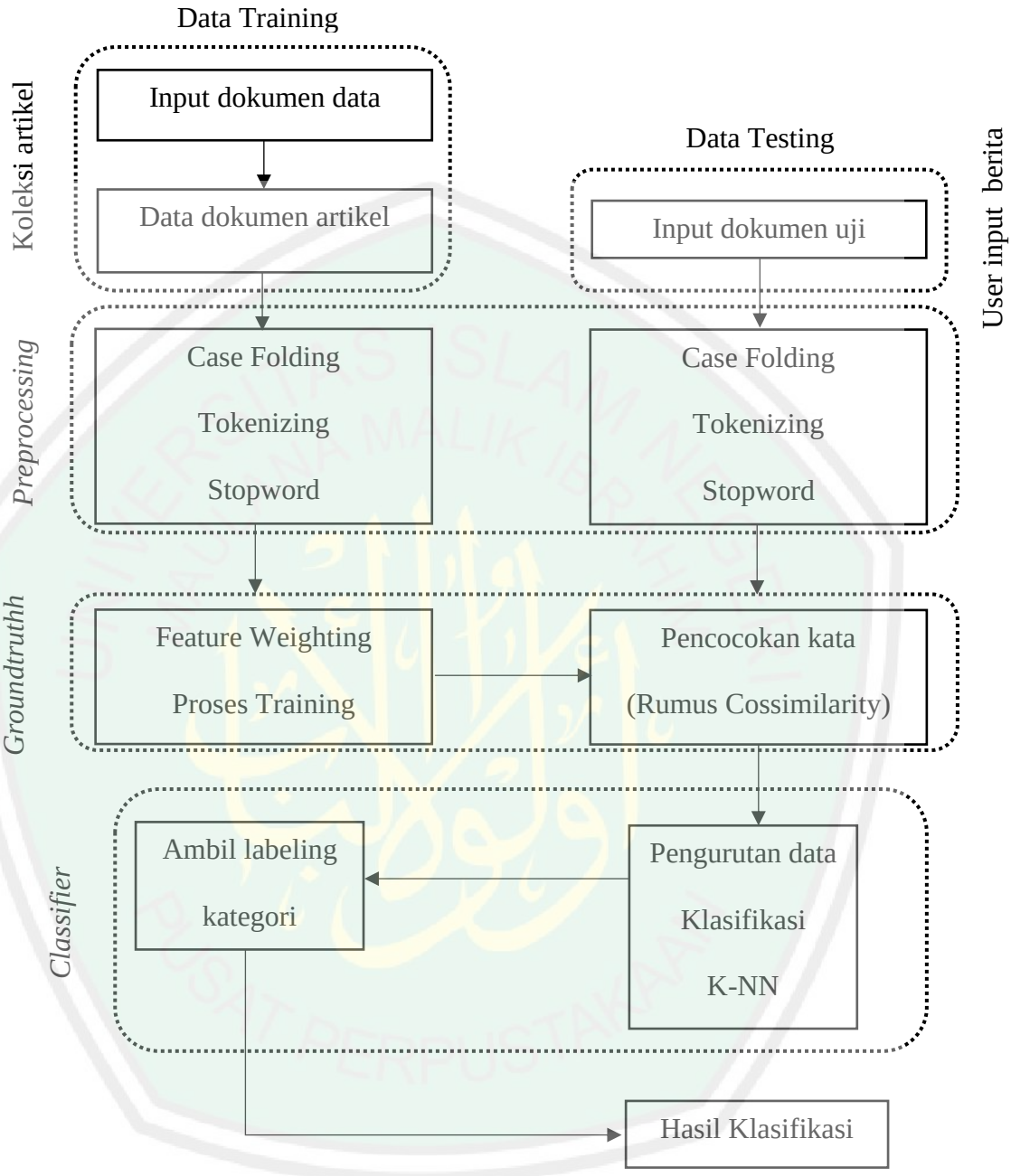
Gambar 3. 6 Potongan *Source Code Proses Mining*

Pada potongan program diatas berfungsi untuk tambah data dimana dapat memasukkan url berita baru. Berikut ini hasil implementasi halaman proses mining.



Gambar 3. 7 Halaman *Proses Mining*

Hasil pada tahap ini adalah kumpulan berita yang telah diklasifikasikan ke dalam kategori berita dengan menghitung setiap bobot kata yang ada dan menghitung nilai kesamaan kata dalam tetangga terdekat dalam data training untuk diuji terhadap data testing . Untuk jelasnya dapat dilihat pada Gambar 3.8.



Gambar 3. 8 Desain Sistem

3.2.1 *Preprocessing*

Data yang digunakan dalam penelitian ini diambil dari portal berita Detik dan Kompas dengan memanfaatkan *web scraping*. Tahap *preprocessing*

dilakukan aplikasi memilih data yang akan melakukan proses pada setiap dokumen. Proses *preprocessing* diantaranya (1) *Case Folding* (menyeragamkan bentuk huruf) (2) *Tokenizing* (pemenggalan suku kata) (3) *Stopword* (menghilangkan kata tidak deskriptif) (4) *Stemming* (merubah suku kata menjadi bentuk kata dasar). Kemudian diberi label, pembobotan dan proses klasifikasi dengan *K-Nearest Neighbor*. Berikut adalah source codenya :

```
public function doProccess(){
		$post = $this->input->post();
		$dataMinning = [];

		$dataBerita = $this->dataMinning;
		$dataBerita = (object) $dataBerita;

		$dataBerita->berita_isi = preg_replace('/(\v|\s)+/', ' ', $dataBerita->berita_isi);

		$dataMinning['dataBerita'] = $dataBerita;

		$dataCaseFolding = $this->caseFolding($dataBerita);
		$dataMinning['dataCaseFolding'] = $dataCaseFolding;

		$dataTokenizing = $this->tokenizing($dataCaseFolding);
		$dataMinning['dataTokenizing'] = $dataTokenizing;

		$dataStopword = $this->stopWord($dataTokenizing);
		$dataMinning['dataStopword'] = $dataStopword;

		$dataStemming = $this->doStemming($dataStopword);
		$dataMinning['dataStemming'] = $dataStemming;

		$dataTraining = $this->db->query('select datatraining_id,datatraining_isi as
berita_isi,datatraining_label from tbl_datatraining')->result();
		foreach ($dataTraining as $key => $value) {
			$dataCaseFolding = $this->caseFolding($value);
			$dataTokenizing = $this->tokenizing($dataCaseFolding);
			$dataStopword = $this->stopWord($dataTokenizing);
			$dataStemming = $this->doStemming($dataStopword);
			$dataTraining[$key] = $dataStemming;
		}

		$k = 3;
		$datatfIdf = $this->tfIdf($dataTraining,$dataMinning['dataStemming']);
		$nilaibanyak = [];
		$hitungSimilairy = [];
		foreach ($datatfIdf['cosSim'] as $key => $value) {
			if ($key<$k) {
				$arrayKe = str_replace("D","",$value['document']);
				$arrayKe = intval($arrayKe)+1;
				if (!empty($dataTraining[$arrayKe])) {
					if(empty($nilaibanyak[$dataTraining[$arrayKe]-
>datatraining_label])) {
						$nilaibanyak[$dataTraining[$arrayKe]-
>datatraining_label] = 0;
```

```
                                        $hitungSimilairy[$dataTraining[$arrayKe]-
>datatraining_label] = 0;
                                        }
                                        $hitungSimilairy[$dataTraining[$arrayKe]-
>datatraining_label] += $value['nilai'];
                                        $nilaibanyak[$dataTraining[$arrayKe]-
>datatraining_label]++;
                                    }
                                }
                            }

                            foreach ($nilaibanyak as $key => $value) {
                                $weight[$key] = 1/(pow($nilaibanyak[$key]/min($nilaibanyak),1)/exp(2));
                                $scoring[$key] = $weight[$key]*$hitungSimilairy[$key];
                            }

                            $dataTest['hitungWeight'] = $weight;
                            $dataTest['hitungScoring'] = $scoring;

                            $keyHeight = array_search(max($scoring), $scoring);

                            $this->m_umum->generatePesan("Hasil data menunjukan ".$keyHeight,"berhasil");
                                redirect('admin/minning');
}
```

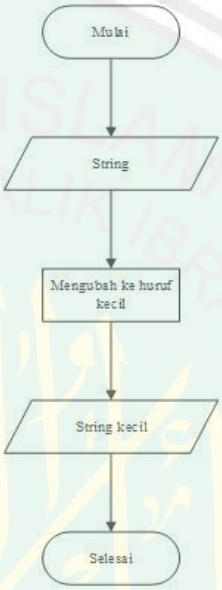
Gambar 3. 9 Potongan *Source Code Process*

3.2.2 *Case Folding*

Dokumen teks biasanya tidak konsisten dalam pemakaian huruf kapital. Oleh sebab itu digunakan *case folding* untuk mengkonversikan seluruh teks dalam dokumen menjadi bentuk standar (huruf kecil atau *lowercase*). Misalnya, pengguna ingin memperoleh informasi tentang "BERITA" dan mengetik "BeRiTa", "BERITA", atau "berita" masih diberi hasil pencarian yang sama dengan "berita". *Case folding* adalah merubah semua huruf dokumen menjadi bentuk huruf kecil. Untuk huruf 'a' hingga 'z' yang diterima. Karakter selain huruf akan dihapus. Lebih detail bisa dilihat pada Gambar 3.9 dan 3.10 *Flowchart case folding*.

```
function caseFolding($berita)
        {
                $new = clone $berita;
                $new->berita_isi = strtolower($new->berita_isi);
                return $new;
        }
```

Gambar 3. 10 Potongan *Source Code Case Folding*



Gambar 3. 11 *Flowchart Case Folding*

3.2.3 *Tokenizing*

Tahap *tokenizing* digunakan untuk memisahkan kalimat dalam *string* menjadi beberapa kata. Contoh penggunaan *tokenizing* dapat dilihat pada Tabel 3.1. Contoh *Tokenizing* sebagai berikut :

Tabel 3. 1 *Tokenizing*

Teks Input	Teks Output
bela anies tak ke bogor, gerindra menyindir ‘orang mau jadi menteri jokowi’	bela | anies | tak | ke | bogor | gerindra | menyindir | orang | mau | jadi | menteri | jokowi

sandi memutuskan untuk mengajukan gugatan ke mahkamah konstitusi	sandi | memutuskan | untuk | mengajukan | gugatan | ke | mahkamah | konstitusi
jokowi pun menyampaikan seluruh rakyat indonesia patut berbangga dan bersyukur	jokowi | pun | menyampaikan | seluruh | rakyat | indonesia | patut | berbangga | dan | bersyukur
sandi mengatakan, saat bertemu dengan relawan, dia membicarakan soal membangkitkan geliat ekonomi	sandi | mengatakan | saat | bertemu | dengan | relawan | dia | membicarakan | soal | membangkitkan | geliat | ekonomi

```
function tokenizing($berita)
{
	$new = clone $berita;
	$return = str_replace("\n", ' ', $new->berita_isi);
	$return = str_replace(' ', '-', $return);
	$return = preg_replace('/[^A-Za-z0-9\-]/', '', $return);
	$return = preg_replace('/\d+/u', '', $return);
	$return = str_replace('-', ' ', $return);
	$return =  str_replace(" ","|",$return);
	$return = explode('|', $return);
	foreach ($return as $keys => $values) {
		if (is_numeric($values)) {
			unset($return[$keys]);
		}
		if (ctype_space($values)) {
			unset($return[$keys]);
		}
	}
	foreach ($return as $keys => $values) {
		if (empty($values)) {
			unset($return[$keys]);
		}
	}
	$return = implode('|', $return);
	$new->berita_isi = $return;
	return $new;
}
```

Gambar 3. 12 Potongan *Source Code Tokenizing*

Tokenizing menguraikan kelompok karakter dalam teks ke dalam unit kata. Misalnya, karakter *whitespace*, seperti *enter*, tabulasi, spasi dianggap sebagai pemisah kata. Tetapi untuk karakter tunggal (‘), titik (.), titik koma (;), titik dua (:), atau lainnya, dapat memiliki peran yang banyak sebagai pemisah tiap kata.

3.2.4 *Stopword*

Pada tahap ini penghapusan kata-kata atau kata - kata yang kurang penting yang sering muncul (*Stopword*), seperti kata sambung dan kata keterangan yang bukan kata - kata unik seperti “sebuah”, “oleh”, “pada”, dan sebagainya. Contoh tahap *stopword* dapat dilihat pada Tabel 3.2. dan source code pada gambar 3.12. Contoh *Stopword* sebagai berikut :

Tabel 3. 2 Contoh *Stopword*

Hasil *Tokenizing*	Hasil *Fitlering*
bela | anies | tak | ke | bogor | gerindra | menyindir | orang | mau | jadi | menteri | jokowi	bela | anies | bogor | gerindra | menyindir | orang | menteri | jokowi
sandia | memutuskan | untuk | mengajukan | gugatan | ke | mahkamah | konstitusi	sandi | memutuskan | mengajukan | gugatan | mahkamah | konstitusi
jokowi | pun | menyampaikan | seluruh | rakyat | indonesia | patut | berbangga | dan | bersyukur	jokowi | pun | rakyat | indonesia | patut | berbangga | bersyukur
sandi | mengatakan | saat | bertemu | dengan | relawan | dia | membicarakan | soal | membangkitkan | geliat | ekonomi	sandi | saat | bertemu | relawan | membicarakan | membangkitkan | geliat | ekonomi

```
function stopWord($berita)
{
        $kataStopWord = 'ada adalah adanya adapun agak agaknya agar akan akankah akhir akhiri akhirnya aku akulah amat amatlah anda andalah antar antara … wah wahai waktu waktunya walau walaupun wong yaitu yakin yakni yang -';
        $kataStopWord = explode(' ', $kataStopWord);
        $new = clone $berita;
        // foreach ($new->berita_isi as $key => $value) {
        $return = preg_replace('/\b('.implode('|',$kataStopWord).')\b/','',$new->berita_isi);
        $return = explode('|', $return);
        foreach ($return as $keys => $values) {
                if (empty($values) || $values == '-') {
                        unset($return[$keys]);
                }
        }
        $new->berita_isi = implode('|', $return);
        // }
        return $new;
```

Gambar 3. 13 Potongan *Source Code Stopword*

3.2.5 *Stemming*

Tahap *Stemming* adalah proses menghapus imbuhan, awalan, akhiran yang bertujuan untuk mengubah kata-kata sesuai dengan kata dasarnya. Contoh dari tahap *stemming* dapat dilihat pada Tabel 3.3 dan *source code stemming* pada gambar 3.13. Contoh *Stemming* sebagai berikut :

Tabel 3. 3 Contoh *Stemming*

Hasil *Filtering*	Hasil *Stemming*
bela | anies | bogor | gerindra | menyindir | orang | menteri | jokowi	bela | anies | bogor | gerindra | sindir | orang | menteri | jokowi
sandi | memutuskan | mengajukan | gugatan | mahkamah | konstitusi	sandi | putus | aju | gugat | mahkamah | konstitusi
jokowi | pun | rakyat | indonesia | patut | berbangga | bersyukur	jokowi | pun | rakyat | indonesia | patut | bangga | syukur
sandi | saat | bertemu | relawan | membicarakan | membangkitkan | geliat | ekonomi	sandi | saat | temu | relawan | bicara | bangkit | geliat | ekonomi

```
function doStemming($berita)
	{
		$new = clone $berita;
		// foreach ($new->berita_isi as $key => $value) {
		$dataBerita = explode('|', $new->berita_isi);
		foreach ($dataBerita as $keys => $values) {
			$dataBerita[$keys] = $this->stemming($values);
		}
		$new->berita_isi = implode('|', $dataBerita);
		// }
		return $new;
	}
```

Gambar 3. 14 Potongan *Source Code Stemming*

3.2.6 *Groundtruth*

Pada tahap ini setiap kata yang telah melalui *preprocessing* akan memiliki nilai *groundtruth* / nilai dasar kebenaran untuk label dari setiap berita yang dimasukkan dari sistem pada saat training.

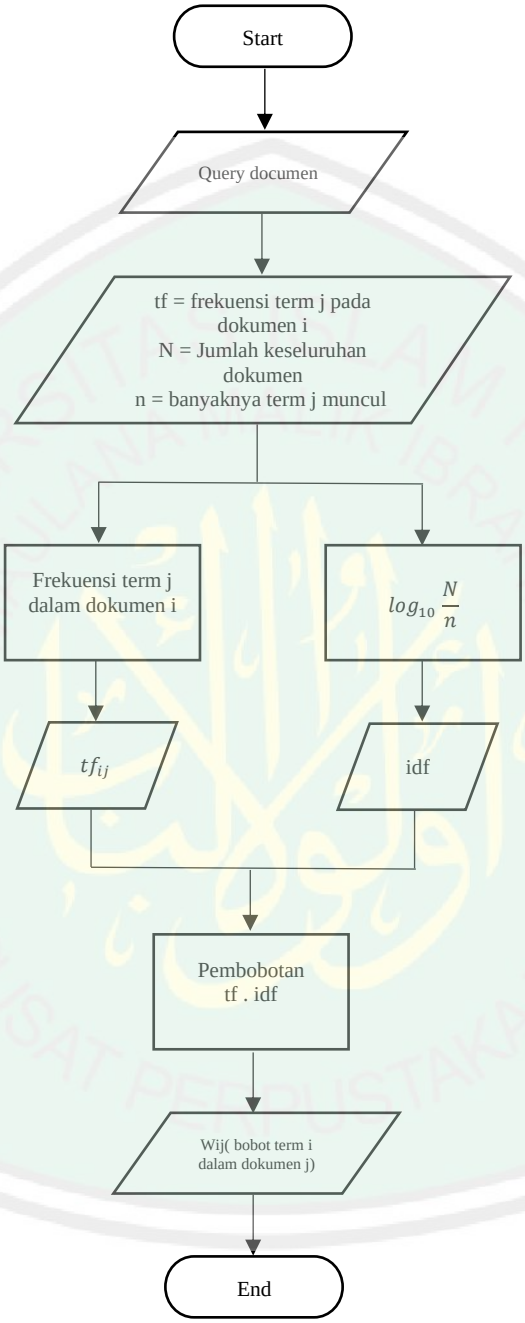
3.2.7 Pembobotan Kata

Dalam klasifikasi berita, pembobotan kata digunakan untuk mendapatkan suatu kategori. Salah satu metode pembobotan adalah TF-IDF (*Term Frequency – Inverse Document Frequency*). Metode *Term Frequency-Inverse Document Frequency* (*tf-idf*) adalah cara pemberian bobot hubungan suatu kata (*term*) terhadap dokumen. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu *term frequency* (*tf*) merupakan frekuensi kemunculan *term* j pada dokumen i. *Document frequency* (*df*) adalah banyaknya kalimat dimana suatu term j muncul. Frekuensi kemunculan *term* di dalam dokumen yang diberikan menunjukkan

seberapa penting *term* itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung *term* tersebut menunjukkan seberapa umum *term* tersebut. Bobot *term* semakin besar jika sering muncul dalam suatu dokumen dan semakin kecil jika muncul dalam banyak dokumen. Pada algoritma *tf-idf* digunakan rumus untuk menghitung bobot term masing masing dokumen dapat dituliskan dalam persamaan (3.1).

$$tf.idf_{ij} = tf_{ij}.log_{10}\frac{N}{n} \quad (3.1)$$

Dimana tf_{ij} merupakan frekuensi kemunculan *term* j dalam dokumen i; idf jumlah kata pada semua dokumen; N merupakan jumlah keseluruhan dokumen; dan n merupakan jumlah dokumen dimana *term* j muncul. Berikut adalah algoritma pembobotan dengan metode TF-IDF pada gambar 3.15 :



Gambar 3. 15 Algoritma Pembobotan TF-IDF

Hal yang perlu diperhatikan dalam pencarian informasi dari koleksi dokumen yang heterogen adalah pembobotan *term*. *Term* dapat berupa kata, frase atau unit hasil *indexing* lainnya dalam suatu dokumen yang dapat digunakan untuk mengetahui konteks dari dokumen tersebut, maka untuk setiap kata tersebut diberikan indikator, yaitu *term weight*.

Inverse Document Frequency (IDF) digunakan dalam meningkatkan fungsi *precision*. penggunaan IDF akan menghasilkan performa yang lebih efektif jika dibandingkan dengan penggunaan frekuensi *term* saja. Kemudian dalam penelitian oleh Salton (1989), untuk mengkombinasikan metode *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF), dengan mempertimbangkan frekuensi antar dokumen dan frekuensi intradokumen dari suatu *term*.

Term Frequency (TF) adalah algoritma yang menunjukkan berapa banyak kata yang muncul dalam setiap dokumen. Sedangkan *Inverse Document Frequency* (IDF) menunjukkan jumlah dokumen yang memiliki kata dalam satu segmen publikasi. Jadi algortima TF-IDF adalah algoritma yang didasarkan pada nilai statistik yang menunjukkan tampilan kata dalam dokumen. TF-IDF adalah hasil dari penggandaan antara TF dengan IDF.

Nilai bobot suatu kata (*term*) menyatakan kepentingan bobot tersebut dalam merepresentasikan judul. Pada pembobotan TF-IDF, bobot akan semakin besar jika

frekuensi kemunculan kata semakin tinggi, tetapi bobot akan berkurang jika kata tersebut semakin sering muncul pada berita lainnya.

Dari persamaan tersebut, diketahui:

$$idf = \log(\frac{N}{df}) \quad (3.2)$$

N = Berita

df = Banyaknya berita dimana suatu kata (*term*) muncul.

Nilai bobot suatu kata (*term*) menyatakan kepentingan bobot tersebut dalam merepresentasikan judul. Pada pembobotan TF-IDF, bobot akan semakin besar jika frekuensi kemunculan kata semakin tinggi, tetapi bobot akan berkurang jika kata tersebut semakin sering muncul pada berita lainnya.

Contoh: Terdapat empat berita (sudah melewati *preprocessing*) seperti berikut:

a. bela anies bogor gerindra sindir orang menteri jokowi
b. sandi putus aju gugat mahkamah konstitusi
c. jokowi pun rakyat indonesia patut bangga syukur
d. sandi saat temu relawan bicara bangkit geliat ekonomi

Doc1 adalah kategori olahraga.

Doc2 adalah kategori teknologi.

Doc3 adalah kategori politik.

Doc4 adalah kategori ekonomi.

Untuk *query* yang diujikan adalah sebagai berikut:

a. anies jokowi orang indonesia

Tabel 3. 4 Perhitungan TF-IDF

No.	Kata	Doc 1	Doc 2	Doc 3	Doc 4	df	Idf	Tf.idf			
								Doc 1	Doc 2	Doc 3	Doc 4
1	bela	1	0	0	0	1	Log(4/1)=0.60205	0.60205	0	0	0
2	anies	1	0	0	0	1	Log(4/1)=0.60205	0.60205	0	0	0
3	bogor	1	0	0	0	1	Log(4/1)=0.60205	0.60205	0	0	0
4	gerindra	1	0	0	0	1	Log(4/1)=0.60205	0.60205	0	0	0
5	sindir	1	0	0	0	1	Log(4/1)=0.60205	0.60205	0	0	0
6	orang	1	0	0	0	1	Log(4/1)=0.60205	0.60205	0	0	0
7	menteri	1	0	0	0	1	Log(4/1)=0.60205	0.60205	0	0	0
8	jokowi	1	0	1	0	2	Log(4/2)=0.30102	0.30102	0	0.30102	0
9	sandi	0	1	0	1	2	Log(4/2)=0.30102	0	0.30102	0	0.30102
10	putus	0	1	0	0	1	Log(4/1)=0.60205	0	0.60205	0	0
11	aju	0	1	0	0	1	Log(4/1)=0.60205	0	0.60205	0	0
12	gugat	0	1	0	0	1	Log(4/1)=0.60205	0	0.60205	0	0
13	mahkamah	0	1	0	0	1	Log(4/1)=0.60205	0	0.60205	0	0
14	konstitusi	0	1	0	0	1	Log(4/1)=0.60205	0	0.60205	0	0
15	pun	0	0	1	0	1	Log(4/1)=0.60205	0	0	0.60205	0
16	rakyat	0	0	1	0	1	Log(4/1)=0.60205	0	0	0.60205	0
17	indonesia	0	0	1	0	1	Log(4/1)=0.60205	0	0	0.60205	0
18	patut	0	0	1	0	1	Log(4/1)=0.60205	0	0	0.60205	0
19	bangga	0	0	1	0	1	Log(4/1)=0.60205	0	0	0.60205	0
20	syukur	0	0	1	0	1	Log(4/1)=0.60205	0	0	0.60205	0
21	saat	0	0	0	1	1	Log(4/1)=0.60205	0	0	0	0.60205
22	temu	0	0	0	1	1	Log(4/1)=0.60205	0	0	0	0.60205
23	relawan	0	0	0	1	1	Log(4/1)=0.60205	0	0	0	0.60205
24	bicara	0	0	0	1	1	Log(4/1)=0.60205	0	0	0	0.60205
25	bangkit	0	0	0	1	1	Log(4/1)=0.60205	0	0	0	0.60205
26	geliat	0	0	0	1	1	Log(4/1)=0.60205	0	0	0	0.60205
27	ekonomi	0	0	0	1	1	Log(4/1)=0.60205	0	0	0	0.60205

Setelah dilakukan pembobotan *term* terhadap tiap-tiap dokumen, maka selanjutnya dilakukan perhitungan *term* terhadap *query*. Dengan cara mengalikan jumlah masing-masing *term query* dengan bobot IDF dari *term* dokumen, maka diperoleh TF.IDF (q_i, d_j).

Tabel 3. 5 Perhitungan TF-IDF pada *term* query uji terhadap tiap dokumen

Term	**TF**				**IDF**	**TF.IDF (q_i, d_j)**
	D1	**D2**	**D3**	**D4**		
anies	1	0	0	0	0.60205	0.60205
jokowi	1	0	1	0	0.30102	0.30102
orang	1	0	0	0	0.60205	0.60205
indonesia	0	0	1	0	0.60205	0.60205

3.2.8 *K-Nearest Neighbor*

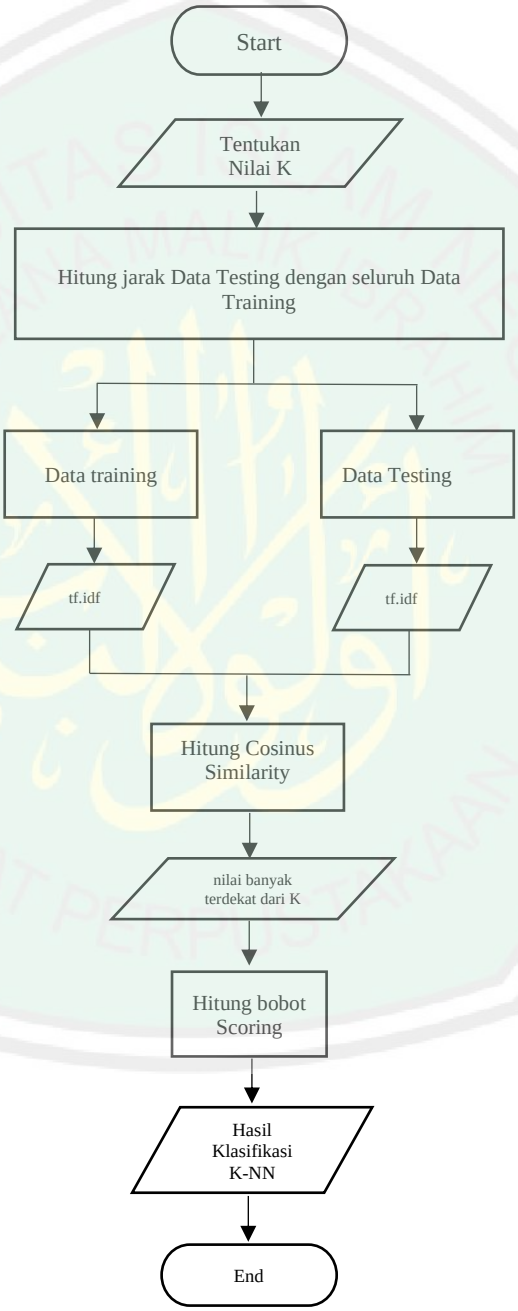
Setelah masing-masing *term* mempunyai bobot dalam bentuk *vector*, maka selanjutnya adalah menghitung kemiripan antara *vector query* dengan *vector* masing-masing dokumen. Perhitungan kemiripan ini menggunakan teknik *Cosine Similarity*.

Langkahnya adalah dengan menghitung perkalian skalar antara vector *query* / data uji dengan vector masing-masing dokumen pada data training, hasil perkalian kemudian dijumlahkan. Dengan menghitung panjang vector masing-masing dokumen, termasuk vector *query*, dengan cara mengkuadratkan bobot setiap *term* terhadap masing-masing dokumen, kemudian menjumlahkan nilai kuadrat dan diakarkan. Kemudian nilai perkalian scalar antara *vector query* dengan *vector* masing-masing dokumen dibagi dengan panjang *vector* sehingga diperoleh nilai kosinus antara *vector query* dengan *vector* masing-masing dokumen. Perhitungan nilai *Cosine Similarity* untuk pembobotan TF.IDF ditunjukkan pada tabel 3.6.

```
k = 3;
$datatfIdf = $this->tfIdf($dataTraining,$dataMinning['dataStemming']);
$nilaibanyak = [];
$hitungSimilairy = [];
foreach ($datatfIdf['cosSim'] as $key => $value) {
        if ($key<$k) {
                $arrayKe = str_replace("D","",$value['document']);
                $arrayKe = intval($arrayKe)+1;
                if (!empty($dataTraining[$arrayKe])) {
                        if (empty($nilaibanyak[$dataTraining[$arrayKe]->datatraining_label])) {
                                $nilaibanyak[$dataTraining[$arrayKe]->datatraining_label] = 0;
                                $hitungSimilairy[$dataTraining[$arrayKe]->datatraining_label] = 0;
                        }
                        $hitungSimilairy[$dataTraining[$arrayKe]->datatraining_label] += $value['nilai'];
                        $nilaibanyak[$dataTraining[$arrayKe]->datatraining_label]++;
                }
        }
}

foreach ($nilaibanyak as $key => $value) {
        $weight[$key] = 1/(pow($nilaibanyak[$key]/min($nilaibanyak),1)/exp(2));
        $scoring[$key] = $weight[$key]*$hitungSimilairy[$key];
}

$dataTest['hitungWeight'] = $weight;
$dataTest['hitungScoring'] = $scoring;
```

```
$keyHeight = array_search(max($scoring), $scoring);


$this->db->update('tbl_datatraining',['datatraining_hasil'=>$keyHeight],['datatraining_id'=>$id]);
// $this->m_umum->generatePesan("Hasil data menunjukan ".$keyHeight,"berhasil");
// redirect('admin/datatraining/daftar');
```

Gambar 3. 16 Potongan *Source Code K-NN*



Gambar 3. 17 Algoritma K-NN

Tabel 3. 6 Perhitungan Manual

No	Kata	W (TF.IDF)					$W_Q * W_{D(i)}$			
		W_Q	W_{D1}	W_{D2}	W_{D3}	W_{D4}	W_Q*W_{D1}	W_Q*W_{D2}	W_Q*W_{D3}	W_Q*W_{D4}
1	Bela	0	0.60205	0	0	0	0	0	0	0
2	anies	0.60205	0.60205	0	0	0	0.36246	0	0	0
3	bogor	0	0.60205	0	0	0	0	0	0	0
4	gerindra	0	0.60205	0	0	0	0	0	0	0
5	sindir	0	0.60205	0	0	0	0	0	0	0
6	orang	0.60205	0.60205	0	0	0	0.36246	0	0	0
7	menteri	0	0.60205	0	0	0	0	0	0	0
8	jokowi	0.30102	0.30102	0	0.30102	0	0.09061	0	0.090610	0
9	sandi	0	0	0.30102	0	0.30102	0	0	0	0
10	putus	0	0	0.60205	0	0	0	0	0	0
11	Aju	0	0	0.60205	0	0	0	0	0	0
12	gugat	0	0	0.60205	0	0	0	0	0	0
13	mahkamah	0	0	0.60205	0	0	0	0	0	0
14	konstitusi	0	0	0.60205	0	0	0	0	0	0
15	Pun	0	0	0	0.60205	0	0	0	0	0
16	rakyat	0	0	0	0.60205	0	0	0	0	0
17	indonesia	0.60205	0	0	0.60205	0	0	0	0.36246	0
18	patut	0	0	0	0.60205	0	0	0	0	0
19	bangga	0	0	0	0.60205	0	0	0	0	0
20	syukur	0	0	0	0.60205	0	0	0	0	0
21	Saat	0	0	0	0	0.60205	0	0	0	0
22	temu	0	0	0	0	0.60205	0	0	0	0
23	relawan	0	0	0	0	0.60205	0	0	0	0
24	bicara	0	0	0	0	0.60205	0	0	0	0
25	bangkit	0	0	0	0	0.60205	0	0	0	0
26	geliat	0	0	0	0	0.60205	0	0	0	0
27	ekonomi	0	0	0	0	0.60205	0	0	0	0
Jumlah		2.10717	4.51537	3.31127	3.6123	4.21435	1.33981	0	0.45307	0
Panjang vektor		1.08536	1.72925	1.50512	1.50512	1.72925				
Cosine simlarity			0.4760	0	0.1748	0				

Jumlah perkalian bobot *query* (W_Q) dengan bobot tiap-tiap dokumen ($W_{(D(i))}$) didapat dari perkalian bobot vector *query* dengan bobot *vector* masing-masing dokumen, kemudian di jumlahkan. Contoh pada D3, dimana:

$W_Q * W_{D3}$ = (0.30102 *0.30102) + (0.60205 *0.60205) = 0.45307

Dilakukan juga, perhitungan jumlah bobot *query* (W_Q) dengan bobot D1, hingga D5.

Panjang *vector* diperoleh dari nilai akar dari total bobot *query* (W_Q) kuadrat oleh bobot masing-masing dokumen ($W_{(D(i))}$) yang telah dikuadratkan. Misalnya pada D3, di mana :

Panjang Vector D3 = $(0.30102^2 + 0.60205^2 + 0.60205^2 + 0.60205^2 + 0.60205^2 + 0.60205^2)^{1/2}$ = 1.505124

Dilakukan juga, Perhitungan panjang *vector* pada bobot *query* (W_Q) dan bobot D1, hingga D4.

Cosine similarity adalah ukuran kesamaan yang lebih umum digunakan dalam *information*. Setiap vektor mewakili setiap kata dalam setiap dokumen (teks). Ukuran ini menghitung nilai cosinus sudut antara dua vector.

Perhitungan kesamaan menghasilkan bobot dokumen yang mendekati nilai 1 atau menghasilkan bobot dokumen yang lebih besar dari nilai yang dihasilkan dari perhitungan *inner product*. *Cosine Similarity* dapat dirumuskan dalam persamaan (3.3).

$$sim\ (q,d) = \frac{q.d}{|q| * |d|} = \frac{\sum_{i=1}^{t} W_{iq} + W_{ij}}{\sqrt{\sum_{j=1}^{t}(W_{iq})^2} + \sqrt{\sum_{j=1}^{t}(W_{ij})^2}} \qquad (3.\ 2)$$

Similarity atau *sim* (q,d_j) antara *query* dan dokumen berbanding lurus dengan jumlah bobot query (q) dikalikan dengan bobot dokumen (d_j)

dan berbanding terbalik dengan jumlah akar kuadrat *q* *(|q|)* dikalikan dengan jumlah akar kuadrat dokumen *(|dj|)*.

Perhitungan *Cosine Similarity* diperoleh dengan membagi jumlah kali bobot *query* (WQ) dengan bobot masing-masing dokumen (W(D(i)) dengan perkalian antara panjang *vector query* dan panjang *vector* untuk setiap dokumen. Seperti berikut ini :

$$Cos\ (Q, D1) = \frac{1.33981}{1.08536 + 1.7292561} = 0.4760$$

$$Cos\ (Q, D2) = \frac{0}{1.08536 + 1.505124} = 0$$

$$Cos\ (Q, D3) = \frac{0.45307}{1.08536 + 1.505124} = 0.1748$$

$$Cos\ (Q, D4) = \frac{0}{1.08536 + 1.7292561} = 0$$

Berdasarkan seluruh kombinasi perhitungan antar kalimat, diperoleh *Cosine Similarity* untuk setiap kalimat yang dipresentasikan pada tabel berikut.

Tabel 3. 7 *Cosine Similarity*

Dokumen	***Cosine Similarity***
D1	0.4760
D2	0
D3	0.1748
D4	0
Rata - rata	0.1627

Dari hasil perhitungan di atas, dilanjutkan tahap perangkingan dokumen, dapat diketahui bahwa dokumen yang memiliki nilai paling tinggi adalah dokumen 1, yakni 0.4760 dan terendah adalah dokumen 2 dan dokumen 4, yakni 0. Tabel perankingannya adalah sebagai berikut :

Tabel 3. 8 Perangkingan dokumen

Ranking	**Dokumen**	**Kategori**	***Cosine Similarity***
1	D1	Olahraga	0.4760
2	D3	Politik	0.1748
3	D4	Ekonomi	0
4	D2	Teknologi	0

Menggunakan nilai *Cosine Similarity* tabel 3.8 dapat ditentukan relevansi awal untuk setiap kalimat. Kalimat dengan bobot di atas rata-rata akan dianggap relevan, sementara sebaliknya dianggap tidak relevan. Dengan rata-rata sebesar 0.1627, maka diperoleh 2 dokumen yang relevan.

Tabel 3. 9 Relevansi dokumen

Dokumen	***Cosine Similarity***	**Relevansi**
D1	0.4760	RELEVAN
D3	0.1748	RELEVAN
D4	0	TIDAK RELEVAN
D2	0	TIDAK RELEVAN

Relevansi dari suatu kalimat, dapat dilihat dari seberapa relevan tetangga yang ada di sekitarnya. Dengan mempertimbangkan 4 contoh dokumen training dan 1 data uji kita hanya bisa menentukan tetangga terdekat yaitu dari akar jumlah dokumen yaitu sebanyak 2 tetangga dan untuk menghindari nilai yang sama maka dipilih angka ganjil yaitu 1 tetangga yang berdekatan untuk kasus contoh ini (k = 3), data yang memiliki tetangga terdekat dengan dokumen uji di D1 dengan probabilitas relevansi terbesar pada nilai *Cosine Similarity* 0.4760 pada kategori Olahraga. Maka dapat disimpulkan kategori pada data uji adalah Olahraga. Berikut adalah *source code* dari pembobotan *tfidf* hingga proses perhitungan *cosimilarity* :

```
function tfIdf($berita,$testing)
{
	$dataPerkata = array();
	array_unshift($berita,$testing);
	$new = $berita;
	$jumlah_berita = count($berita);
	$checking = [];

	$docnya = [];
	foreach ($new as $key => $value) {
		$docnya["D".$key] = 0;
	}

	foreach ($new as $key => $value) {
		$dataBerita = explode('|', $value->berita_isi);
		foreach ($dataBerita as $keys => $values) {
			if (array_search($values,$checking) == false) {
				$checking[] = $values;
				$perkata = array('kata'=>
$values,'doc'=>$docnya,'df'=>0,'idf'=>0,'jumlah_doc'=>$jumlah_berita,'tfidf'=>array(),'sim'=>array(),'pVector'=>array(
));
				if (!empty($values)) {
					$dataPerkata[] = $perkata;
				}
			}
		}
	}

	foreach ($dataPerkata as $key => $value) {
		foreach ($new as $keys => $values) {
			$dataPerkata[$key]['doc']["D".$keys] += substr_count($values->berita_isi,
$value['kata']);
			$dataPerkata[$key]['df'] += substr_count($values->berita_isi, $value['kata']);
		}
		$dataPerkata[$key]['idf'] =
round(log(($dataPerkata[$key]['jumlah_doc']/$dataPerkata[$key]['df']),2),5);
	}
```

```
		$datajumlah = array('w'=>array(),'perkalianp'=>array(),'pdoc'=>1/$jumlah_berita,'perkalianprior'=>array(),'maxPrior'=>0);
		$jumlahtotalIdf = 0;

		foreach ($dataPerkata as $key => $value) {
			foreach ($value['doc'] as $keys => $values) {
				if (empty($datajumlah['w'][$keys])) {
					$datajumlah['w'][$keys] = 0;
				}
				if ($values > 0) {
					$datajumlah['w'][$keys] += $dataPerkata[$key]['idf'];
					$dataPerkata[$key]['tfidf'][$keys] = $dataPerkata[$key]['idf'];
				}else{
					$datajumlah['w'][$keys] += 0;
					$dataPerkata[$key]['tfidf'][$keys] = 0;
				}
			}
		}
		$totalAll = [];
		$totalSim = $docnya;
		$totalpVec = $docnya;
		$totalAkrpVec = $docnya;

		foreach ($dataPerkata as $key => $value) {
			foreach ($value['doc'] as $keys => $values) {
				if ($keys != 'D0' && $value['doc']['D0'] > 0) {
					$dataPerkata[$key]['sim'][$keys] = round($value['tfidf'][$keys]*$value['tfidf']['D0'],3);
				}else{
					$dataPerkata[$key]['sim'][$keys] = 0;
				}
				$dataPerkata[$key]['pVector'][$keys] = round(pow($value['tfidf'][$keys],2),3);
				$totalpVec[$keys] += $dataPerkata[$key]['pVector'][$keys];
				$totalSim[$keys] += $dataPerkata[$key]['sim'][$keys];
			}
		}

		foreach ($totalpVec as $key => $value) {
			$totalAkrpVec[$key] = round(sqrt($value),2);
		}

		$totalAll = ['sim'=>$totalSim,'pVec'=>$totalpVec,'AkrpVec'=>$totalAkrpVec];

		foreach ($totalAll['sim'] as $key => $value) {
			if (($totalAll['AkrpVec']['D0']*$totalAll['AkrpVec'][$key]) == 0) {
				$totalAll['cosSim'][$key] =  ['nilai'=>0,'document'=>$key];
			}else{
				$totalAll['cosSim'][$key] = ['nilai'=>$totalAll['sim'][$key]/($totalAll['AkrpVec']['D0']*$totalAll['AkrpVec'][$key]),'document'=>$key];
			}
		}

		if (!function_exists('cmp')) {
			function cmp($a, $b) {
			   return $a['nilai']<$b['nilai'];
			}
		}

		usort($totalAll['cosSim'], "cmp");
		// $new->perkata = $dataPerkata;
		// $new->datajumlah = $datajumlah;
		$newmax = $totalAll;
		return $newmax;
}
```

Gambar 3. 18 *Source Code* Pembobotan

BAB IV
UJI COBA DAN PEMBAHASAN

4.1 Langkah-langkah Uji Coba

Berikut ini adalah langkah-langkah uji coba sistem termasuk:

1. Pengumpulan data dokumen berita *online*

Pengumpulan dokumen berita terbatas pada 4 kategori yaitu: olahraga, politik, ekonomi dan teknologi. Dokumen berita yang di ambil untuk uji berjumlah 20 berita, yaitu dari detik.com dan kompas.com untuk per kategori 4 berita dan ada 4 berita di luar kategori/ kategori lain.

2. *Prepocessing* dokumen berita

Prepocessing dokumen berita dilakukan secara otomatis oleh sistem, durasi proses *preprocessing* ditentukan dari jumlah dokumen berita yang terdapat pada database, semakin banyak dokumen, semakin lama prosesnya.

3. Training Data.

Setelah *preprocessing* selesai maka proses selanjutnya adalah mentraining data, dengan memasukan url berita pada portal berita detik.com atau kompas.com maka berita akan otomatis tertraining dan menghasilkan data dokumen berita berdasarkan kategori / dengan pemberian nilai *groundtruth* sebagai label pada proses ini memasukkan *data training* sebanyak 400 data dengan keterangan 100 berita per katgeori.

4. Hasil *Training*

Proses terakhir dari sistem adalah menampilkan hasil *training* yaitu berupa bobot kata di setiap berita yang sudah di kategorikan bedasarkan kategori yang telah di tentukan seperti politik, teknologi, olahraga dan ekonomi / dengan *groundtruth*.

5. Evaluasi Hasil

Langkah terakhir dari percobaan ini adalah mengevaluasi hasil, evaluasi hasil dalam penelitian ini menggunakan *confusion matrix* untuk melihat *recall*, *precision* dan akurasi. *Confusion matrix* merupakan alat pengukuran yang dapat digunakan untuk menghitung kinerja atau tingkat kebenaran proses klasifikasi. Dengan *confusion matrix* dapat dianalisa seberapa baik *classifier* dapat mengenali *record* dari kelas-kelas yang berbeda. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya (Prasetyo, 2012).

Berdasarkan jumlah keluaran kelasnya, sistem klasifikasi dapat dibagi menjadi 4 (empat) jenis yaitu klasifikasi *binary*, *multi-class*, *multi-label* dan *hierarchical* (Sokolova, 2009).

Pada klasifikasi *binary*, data masukan dikelompokkan ke dalam salah satu dari dua kelas. Jenis klasifikasi ini merupakan bentuk klasifikasi yang paling sederhana dan banyak digunakan. Contoh penggunaannya antara lain dalam sistem yang melakukan deteksi berita olahraga atau bukan, sistem deteksi berita ekonomi atau bukan, dan sistem deteksi berita teknologi atau bukan. Tabel *confusion matrix* ditunjukkan pada tabel 4.1.

Tabel 4. 1 *Confusion Matrix*

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

TP (*True Positive*) merupakan jumlah data yang kelas aktualnya positif dengan kelas prediksi merupakan kelas positif; FN (*False Negative*) merupakan jumlah data yang kelas aktualnya adalah kelas positif dengan kelas prediksi merupakan kelas negatif; FP (*False Positive*) merupakan jumlah data yang kelas aktualnya adalah kelas negatif dengan kelas prediksinya merupakan kelas positif; TN (*True Negative*) merupakan banyaknya data yang kelas aktualnya adalah kelas negatif dengan kelas prediksinya merupakan kelas negatif.

Akurasi merupakan metode pengujian berdasarkan tingkat kedekatan antara nilai prediksi dengan nilai aktual. Dengan mengetahui jumlah data yang diklasifikasikan secara benar maka dapat diketahui akurasi hasil prediksi. Persamaan akurasi seperti pada persamaan (4.1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x100\% \qquad (4.\ 1)$$

Presisi merupakan metode pengujian dengan melakukan perbandingan jumlah informasi relevan yang didapatkan sistem dengan jumlah seluruh informasi yang terambil oleh sistem baik yang relevan maupun tidak. Persamaan presisi ditunjukkan pada persamaan (4.2).

$$Precision = \frac{TP}{TP+FP} x100\% \qquad (4.\ 2)$$

Recall merupakan metode pengujian yang membandingkan jumlah informasi relevan yang didapatkan sistem dengan jumlah seluruh informasi relevan yang ada dalam koleksi informasi (baik yang terambil atau tidak terambil oleh sistem). Persamaan *recall* ditunjukkan pada persamaan (4.3).

$$Recall = \frac{TP}{TP+FN} x100\% \qquad (4.\ 3)$$

4.2 Hasil Uji Coba

Hasil dari akuisi data lewat *preprocessing* data ada sebanyak 20 data yang dibagi kedalam 2 *sentiment* yaitu sesuai prediksi dan tidak sesuai prediksi. Data yang sudah dinormalisasi sebelum dimasukan ke mesin klasifikasi. Hasil uji coba manggunakan *confusion matrix* untuk membandingkan hasil prediksi dengan algoritma dengan data *testing* yang sebenarnya. Variabel yang digunakan dalam mengevaluasi algoritma K-NN adalah *precision*, *recall*, dan akurasi. Pengujian menggunakan *confusion matrix* untuk perbandingan sebanyak 20 *data testing* dengan 400 *data training* yang diambil dari portal berita yang sudah berlabel dengan pemberian nilai *groundtruth*.

Pada tahap awal untuk melihat seberapa akurat sistem dilakukan uji coba terhadap *data training* yang telah dilatih dan sistem mencoba untuk melakukan prediksi kategori yang sesuai dengan algoritma K-NN pada data latih tersebut sebanyak 400 berita, pada tahap evaluasi akhir data aktual merupakan data yang telah duji oleh pakar sebanyak 20 berita dan akan diprediksi sistem untuk

perhitungan akurasi dengan membandingan nilai pakar dan prediksi sistem yang ada.

Berikut adalah hasil uji coba dalam sistem dengan pengujian *data training* dengan menggunakan *cofusion matrix* 5 kelas karena perhitungan data aktual memerlukan perbandingan terhadap data yang disimpan yang diperlukan utuk proses klasifikasi.

Tabel 4. 2 Hasil ujicoba

No	Asal	*Link*	Pakar	Sistem	TP	FP	FN	TN
1	Detik	https://finance.detik.com/berita-ekonomi-bisnis/d-5002809/waduh-ekonomi-ri-cuma-tumbuh-297-di-kuartal-i-2020	Ekonomi	Olaharaga	0	1	1	3
2	Detik	https://finance.detik.com/berita-ekonomi-bisnis/d-5002466/inflasi-rendah-bukti-daya-beli-lesu	Ekonomi	Olahraga	0	1	1	3
3	Detik	https://news.detik.com/berita/d-5001634/wakil-ketua-dprd-kota-surabaya-minta-protokol-covid-19 diperketat?_ga=2.191066668.745140099.1588655040-467314476.1569770611	Politik	Politik	1	0	0	4
4	Detik	https://news.detik.com/berita/d-5002880/zulhas-buka-rakernas-i-pan-2020-pertama-kali-dilakukan-virtual?_ga=2.191066668.745140099.1588655040-467314476.1569770611	Politik	Politik	1	0	0	4
5	Detik	https://inet.detik.com/cyberlife/d-4995745/fitur-baru-gojek-ini-bikin-masyarakat-makin-siap-hadapi-covid-19?tag_from=wp_nhl_7	Teknologi	Olahraga	0	1	1	3
6	Detik	https://inet.detik.com/science/d-5002771/robot-pepper-ingatkan-orang-orang-pakai-masker?tag_from=wp_nhl_12	Teknologi	Olahraga	0	1	1	3
7	Detik	https://sport.detik.com/raket/d-5002927/kata-juara-bertahan-soal-jadwal-baru-kejuaraan-dunia-bulutangkis-2021	Olahraga	Olahraga	1	0	0	4
8	Detik	https://sport.detik.com/moto-gp/d-5002452/rossi-masih-bisa-balapan-sampai-empat-tahun-lagi	Olahraga	Olahraga	1	0	0	4
9	Detik	https://health.detik.com/berita-detikhealth/d-5003294/872-meninggal-dari-12071-kasus-tingkat-kematian-corona-ri-722-persen?tag_from=wp_nhl_59&_ga=2.203691286.745140099.1588655040-467314476.1569770611	Kesehatan	Olahraga	0	1	1	3

10	Detik	https://news.detik.com/berita/d-5003308/mahasiswa-uin-iain-yang-terdampak-corona-bisa-ajukan-keringanan-ukt?tag_from=wp_nhl_41&_ga=2.203691286.745140099.1588655040-467314476.1569770611	Pendidikan	Politik	0	1	1	3
11	Kompas	https://money.kompas.com/read/2020/05/05/093039726/rupiah-dan-ihsg-pagi-ini-menguat	Ekonomi	Teknologi	0	1	1	3
12	Kompas	https://money.kompas.com/read/2020/05/05/063900726/harga-minyak-dunia-menguat-ini-penyebabnya	Ekonomi	Olahraga	0	1	1	3
13	Kompas	https://nasional.kompas.com/read/2020/04/16/15145571/wapres-maruf-amin-tak-masalah-putrinya-jadi-wasekjen-demokrat	Politik	Politik	1	0	0	4
14	Kompas	https://nasional.kompas.com/read/2020/04/07/06225801/pengamat-blunder-pemerintah-terkait-covid-19-karena-faksi-politik-yang	Politik	Politik	1	0	0	4
15	Kompas	https://tekno.kompas.com/read/2020/05/05/14060017/facebook-bisa-transfer-foto-dan-video-ke-google-photos-begini-caranya	Teknologi	Politik	0	1	1	3
16	Kompas	https://tekno.kompas.com/read/2020/05/05/10034587/duka-warganet-atas-kepergian-didi-kempot-teratas-di-trending-topic-dunia	Teknologi	Teknologi	1	0	0	4
17	Kompas	https://www.kompas.com/sports/read/2020/05/05/10400008/lagu-pamer-bojo-didi-kempot-pernah-temani-perjalanan-garuda-select-di	Olahraga	Teknologi	0	1	1	3
18	Kompas	https://www.kompas.com/sports/read/2020/05/05/11000088/carlo-ancelotti-diyakini-bisa-bawa-everton-tampil-di-liga-champions?page=2	Olahraga	Politik	0	1	1	3
19	Kompas	https://nasional.kompas.com/read/2020/05/05/15512021/update-kini-ada-12071-kasus-covid-19-di-indonesia-bertambah-484	Kesehatan	Ekonomi	0	1	1	3
20	Kompas	https://nasional.kompas.com/read/2020/05/05/15512021/update-kini-ada-12071-kasus-covid-19-di-indonesia-bertambah-484	Kesehatan	Teknologi	0	1	1	3
				Total	7	13	13	67

++ adalah data aktual kelas positif diprediksi kelas positif memiliki nilai benar = TP, -+ adalah data aktual kelas negatif diprediksi kelas positif memiliki nilai salah = FP, +- adalah data aktual positif diprediksi kelas negatif = FN, -- adalah data aktual negatif diprediksi kelas negatif dengan nilai benar = TN.

Tabel 4. 3 *Confusion matrix* hasil uji coba

		Prediksi	
		Positif	Negatif
Aktual	Positif	7	13
	Negatif	13	67

Hasil yang didapatkan dari perhitungan data uji sebanyak 20 data dengan TP adalah data yang diprediski benar merupakan aktual data kelas positif sehingga diperoleh nilai TP = 7. Nilai presisi dapat dicari dengan menentukan nilai FP, dengan ketentuan data yang diambil berdasarkan informasi yang kurang atau salah atau tidak tepat. FP adalah data yang bernilai prediksi kelas positif dalam suatu kelas aktual negatif. Nilai FP = 13 diperoleh dari data aktual kelas negatif diprediksi kelas positif bernilai salah. Sedangkan FN = 13 diperoleh dari kelas aktual positif diprediksi dengan salah. Dapat diketahui nilai akurasi dengan presentase dari total data yang diidentifikasi dan dinilai benar dengan perhitungan sebagai berikut,

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

$$Akurasi = \frac{7 + 67}{7 + 67 + 13 + 13} * 100\% = \frac{74}{100} * 100\% = 74\ \%$$

$$Precision = \frac{TP}{(TP + FP)}\ x\ 100\%$$

$$Precision = \frac{7}{(7 + 13)}\ x\ 100\% = \frac{7}{20}\ x\ 100\% = 35\ \%$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$Recall = \frac{7}{(7 + 13)} x\ 100\ \% = \frac{7}{20} x\ 100\ \% = 35\%$$

$$\text{F-}Measure = 2\ x\ \frac{(Recall\ x\ Precision)}{Recall + Precision} x\ 100\% = 2\ x\ \frac{(0.35\ x\ 0{,}35)}{0{,}35 + 0{,}35} x\ 100\% = 35\%$$

$$Specificity = \frac{TN}{(TN + FP)}$$

$$Specificity = \frac{67}{(67 + 13)} = 0,\ 84$$

Tabel 4. 4 Klasifikasi Nilai AUC

AUC	HASIL
0.90 - 1. 00	Sangat Baik
0.80 - 0.90	Baik
0.70 - 0.80	Sedang
0.60 - 0.70	Buruk
<0.60	Gagal

$$AUC = \frac{Recall + Specificity}{2} = \frac{0{,}35 + 0{,}84}{2} = 0{,}60$$

Setelah diketahui nilai *specificity* dan *recall* maka dihitung nilai AUC (*Area Under Curve*) untuk menentukan nilai AUC. Dengan diperoleh nilai AUC sebesar 0,60 maka hasil klasifikasi menggunakan K-NN terhadap 5 kelas memiliki nilai buruk tapi masih bisa digunakan untuk klasifikasi kategori berita.

4.3 Pembahasan

Dari penelitian yang dilakukan kita bisa mengetahui confusion matrix adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep mining atau Sistem Pendukung Keputusan. Untuk mengukur peforma dari suatu algoritma (*Performance Measure*) diperlukan nilai *precision*, *recall*/ *sensitivity*, *F-Measure*, *specifity*, akurasi, dan AUC (*area under curve*). *Precision* merupakan data yang diambil berdasarkan informasi yang kurang atau salah atau tidak tepat. *Recall / sensitivity* adalah data yang tidak mampu diprediksi dengan benar. *Specificity* mengukur proporsi negatif aktual yang diidentifikasi dengan benar. Akurasi adalah presentase dari total data yang diidentifikasi dan dinilai. AUC adalah alat ukur *performance* untuk *classification problem* dalam menentukan *threshold* dari suatu model.

Jika model evaluasi memberikan nilai 100% maka ada masalah pada model yang dibuat atau data yang digunakan. Sehingga model yang dibuat tidak boleh memberikan nilai 0 *false positive* dan 0 *false negative*. Dalam beberapa kasus *false postive* (FP) atau biasa disebut *Error Type I* merupakan data aktual negatif namun diprediksi sebagai data positif bernilai salah. Contohnya dalam kasus kelas olahraga data aktual berita ekonomi, politik, teknologi (negatif) memiliki nilai prediksi olahraga (positif) diprediksi secara salah masuk kelas positif dalam hal ini nilai *Error Type I* / FP mempengaruhi nilai *precision*. Ada juga kasus *Error Type II* / FN yang mempengaruhi nlai *recall* lebih berbahaya misalnya data aktual berita olahraga (positif) namun prediksi sistem menunjukkna bahwa berita tersebut termasuk ekonomi, politik, dan teknologi (negatif) sehingga prediksi bernilai salah.

Dalam hal ini dapat diketahui bahwa semakin tinggi nilai FP pada sistem dimana data aktual bernilai negatif namun hasil prediksi sebagai data positif semakin kecil nilai precision dan akurasi, begitu pula pada sistem semakin tinggi nilai FN dimana data aktual bernilai positif namun diprediksi sebagai data negatif mempengaruhi nilai recall dan akurasi menjadi semakin kecil. Dalam hal ini TP data aktual positif diprediksi positif bernilai benar dan TN merupakan data negatif yang diprediksi negatif memiliki nilai benar memepengaruhi nilai akurasi. Dalam pelatihan suatu data kategori yang memiliki nilai *overfitting* menghasilkan nilai prediksi yang kurang tepat dalam hal ini nilai TP dan TN sehingga tidak dianjurkan untuk memasukkan data training baru yang memiliki kesamaan dengan data uji yang membentuk nilai berat yang bisa membuat berita uji yang dimasukkan memiliki hasil prediksi yang benar sehingga data latih sebaiknya merata untuk setiap kategorinya.

Dalam penelitian ini diperoleh nilai akurasi sebesar 74 %, *precision* 35 % *recall* 35 %, *F-Measure* 35 %, *specificity* 84 % dan AUC sebesar 60 % sehingga menunjukkan sistem klasifikasi menggunakan metode K-NN memiliki nilai yang buruk. Jika sistem memiliki nilai AUC diatas 60 % menunjukkan suatu sistem dinyatakan berhasil dalam mengklasifikasi maka dinyatakan gagal apabila di bawah 60 %.

Berikut merupakan ayat-ayat yang terkait dengan penelitian yaitu perintah meneliti berita yang datang adalah sebagai berikut:

1. An-Nisa: 83

وَإِذَا جَآءَهُمْ أَمْرٌ مِّنَ ٱلْأَمْنِ أَوِ ٱلْخَوْفِ أَذَاعُوا۟ بِهِۦ ۖ وَلَوْ رَدُّوهُ إِلَى ٱلرَّسُولِ وَإِلَىٰٓ أُو۟لِى ٱلْأَمْرِ مِنْهُمْ لَعَلِمَهُ ٱلَّذِينَ يَسْتَنۢبِطُونَهُۥ مِنْهُمْ ۗ وَلَوْلَا فَضْلُ ٱللَّهِ عَلَيْكُمْ وَرَحْمَتُهُۥ لَٱتَّبَعْتُمُ ٱلشَّيْطَـٰنَ إِلَّا قَلِيلًا

"Dan apabila datang kepada mereka suatu berita tentang keamanan ataupun ketakutan, mereka lalu menyiarkannya. Dan kalau mereka menyerahkannya kepada Rasul dan Ulil Amri di antara mereka, tentulah orang-orang yang ingin mengetahui kebenarannya [akan dapat] mengetahuinya dari mereka [Rasul dan Ulil Amri]. Kalau tidaklah karena karunia dan rahmat Allah kepada kamu, tentulah kamu mengikut syaitan, kecuali sebahagian kecil saja [di antaramu]". (83).

Dalam kitab Tafsir Ibnu Katsir oleh Ismail bin Umar Al-Quraisyi bin Katsir Al-Bashri Ad-Dimasyqi menerangkan, "Dan apabila datang kepada mereka suatu berita tentang keamanan atau ketakutan, mereka lalu menyiarkannya". Hal ini merupakan penginkaran terhadap orang yang tergesa-gesa dalam menanggapi berbagai urusan sebelum meneliti kebenarannya, lalu ia memberitakan dan menyiarkannya, padahal belum tentu hal itu benar. Imam Muslim mengatakan di dalam mukadimah (pendahuluan) kitab sahihnya, "telah menceritakan kepada kami Abu Bakar ibnu Abu Syaibah, telah menceritakan kepada kami Ali ibnu Hafs, telah menceritakan kepada kami Syu'bah, dari Habib ibnu Abdur Rahman, dari Hafs ibnu Asim, dari Abu Hurairah, dari Nabi Saw. yang telah bersabda: Cukuplah kedustaan bagi seseorang bila dia menceritakan semua apa yang didengarnya". Di dalam kitab

Sahihain disebutkan dari Al-Mugirah ibnu Syu'bah hadis berikut, bahwa Rasulullah Saw. telah melarang perbuatan qil dan qal. Makna yang dimaksud ialah melarang perbuatan banyak bercerita tentang apa yang dibicarakan oleh orang-orang tanpa meneliti kebenarannya, tanpa menyeleksi terlebih dahulu, dan tanpa membuktikannya. Di dalam kitab Sunan Abu Daud disebutkan bahwa Rasulullah Saw. telah bersabda, "Seburuk-buruk lisan seseorang ialah (mengatakan) bahwa mereka menduga (anu dan anu)". Di dalam kitab sahih disebutkan hadis berikut, "Barang siapa yang menceritakan suatu kisah, sedangkan ia menganggap bahwa kisahnya itu dusta, maka dia termasuk salah seorang yang berdusta.". Dalam kesempatan ini kami ketengahkan sebuah hadis dari Umar ibnul Khattab yang telah disepakati kesahihannya, "yaitu ketika ia mendengar berita bahwa Nabi Saw. menceraikan istri-istrinya. Maka ia datang dari rumahnya, lalu masuk ke dalam masjid, dan ia menjumpai banyak orang yang sedang memperbincangkan berita itu. Umar tidak sabar menunggu, lalu ia meminta izin menemui Nabi Saw. dan menanyakan kepadanya apakah memang benar beliau menceraikan semua istrinya? Ternyata jawaban Rasulullah Saw. negatif (yakni tidak). Maka ia berkata, "Allahu Akbar (Allah Mahabesar)", hingga akhir hadis". Abdur-Razzak mengatakan, dari Ma'mar, dari Qatadah, bahawa firman Allah berikut, "Tentulah kalian mengikuti setan, kecuali sebagian kecil saja (di anatara kalian). (An-Nisa:83)". Makna yang dimaksud ialah kalian semuanya niscaya mengikuti langkah setan. Orang yang mendukung pendapat ini (yakni yang mengartikan semuanya) memeperkuat alasannya dengan ucapan At-Tirmah ibnu Haki dalam salah satu bait syairnya ketika memuji Yazid ibnul Muhallab yaitu, "Aku mencium keharuman nama orang

yang sangat dermawan, tiada cela dan tiada kekurangan baginya". Makna yang dimaksud ialah tidak ada cela dan tidak ada kekurangannya, sekalipun diungkapkan dengan kata sedikit cela dan kekurangannya.

2. Al-Ahzab: 60-62

لَّئِن لَّمْ يَنتَهِ ٱلْمُنَٰفِقُونَ وَٱلَّذِينَ فِى قُلُوبِهِم مَّرَضٌ وَٱلْمُرْجِفُونَ فِى ٱلْمَدِينَةِ لَنُغْرِيَنَّكَ بِهِمْ ثُمَّ لَا يُجَاوِرُونَكَ فِيهَآ إِلَّا قَلِيلًا ٦٠ (مَّلْعُونِينَ ۖ أَيْنَمَا ثُقِفُوٓا۟ أُخِذُوا۟ وَقُتِّلُوا۟ تَقْتِيلًا ٦١ (سُنَّةَ ٱللَّهِ فِى ٱلَّذِينَ خَلَوْا۟ مِن قَبْلُ ۖ وَلَن تَجِدَ لِسُنَّةِ ٱللَّهِ تَبْدِيلًا

"Sesungguhnya jika tidak berhenti orang-orang munafik, orang-orang yang berpenyakit dalam hatinya dan orang-orang yang menyebarkan kabar bohong di Madinah [dari menyakitimu], niscaya Kami perintahkan kamu [untuk memerangi] mereka, kemudian mereka tidak menjadi tetanggamu [di Madinah] melainkan dalam waktu yang sebentar, (60) dalam keadaan terla'nat. Di mana saja mereka dijumpai, mereka ditangkap dan dibunuh dengan sehebat-hebatnya. (61) Sebagai sunnah Allah yang berlaku atas orang-orang yang telah terdahulu sebelum [mu], dan kamu sekali-kali tiada akan mendapati perubahan pada sunnah Allah. (62)".

Dijelaskan pada kitab Tafsir Ibnu Katsir untuk Al-Azhab 60-62: "dan orang-orang yang menyebarkan kabar bohong di Madinah." Yaitu orang-orang yang mengatakan kepada Nabi dan kaum muslimin, bahwa musuh dalam jumlah yang sangat besar akan datang menyerang dan sebentar lagi akan terjadi perang dhasyat, padahal berita itu dusta dan buat-buatan belaka. Jika mereka tidak mau berhenti dari melakukan perbuatan-perbuatan tersebut (mengganggu Nabi Saw. dan

menyakitinya) dan tidak mau kembali ke jalan yang benar. “niscaya Kami perintahkan kamu (untuk memerangi) mereka”. Ali ibnu Abu Talhah telah meriwayatkan dari Ibnu Abbas, bahwa makna yang dimaksud ialah Kami benar-benar akan menjadikanmu berkuasa atas mereka. Menurut Qatadah, sesungguhnya Kami akan perintahkan kamu untuk memerangi mereka. As-Saddi mengatakan bahwa sesungguhnya Kami memeberikan pelajaran kepada mereka melaluimu. “kemudian mereka tidak menjadi tetanggamu (di Madinah) melainkan dalam waktu yang sebentar, dalam keadaan terlaknat”. Lafaz *mal’unina* berkedudukan menjadi hal atau kata keterangan keadaan bagi mereka. Yakni masa tinggal mereka di Madinah sebentar lagi karena dalam waktu yang dekat mereka akan diusir darinya dalam keadaan terlaknat, yaitu dijauhkan dari rahmat Allah. “Di mana saja mereka dijumpai, mereka ditangkap”. Maksudnya, dimanapun mereka ditemukan, mereka ditangkap karena hina dan jumlah mereka sedikit. “dan dibunuh dengan sehebat-hebatnya”. Kemudian Allah Swt. berfirman, “Sebagai sunnah Allah yang berlaku atas orang-orang yang telah terdahulu sebelum(mu)”. Demikianlah ketetapan Allah terhadap orang-orang munafik. Apabila mereka tetap bersikeras dengan kemunafikan dan kekafirannya serta tidak mau menghentikan perbuatannya, lalu kembali ke jalan yang benar, orang-orang yang beriman akan menguasai mereka dan mengalahkan mereka. “dan kamu sekali-kali tidak akan mendapati perubahan pada sunnah Allah”. Yakni ketetapan Allah dalam hal ini tidak dapat diganti dan tidak pula dapat diubah.

3. Al- Hujurat: 6

يَـٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُوٓاْ إِن جَآءَكُمۡ فَاسِقُۢ بِنَبَإٖ فَتَبَيَّنُوٓاْ أَن تُصِيبُواْ قَوۡمَۢا بِجَهَـٰلَةٖ فَتُصۡبِحُواْ عَلَىٰ مَا فَعَلۡتُمۡ نَـٰدِمِينَ

"Hai orang-orang yang beriman, jika datang kepadamu orang fasik membawa suatu berita, maka periksalah dengan teliti, agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya yang menyebabkan kamu menyesal atas perbuatanmu itu. (6)".

Dalam kitab Tafsir Ibnu Katsir oleh Ismail bin Umar Al-Quraisyi bin Katsir Al-Bashri Ad-Dimasyqi: Allah Swt. memerintahkan (kaum mukmin) untuk memeriksa dengan teliti berita dari orang fasik, dan hendaklah mereka bersikap hati-hati dalam menerimanya dan jangan menerimanya dengan begitu saja, yang akibatnya akan membalikkan kenyataan. Orang yang menerima dengan begitu saja berita darinya, berarti sama dengan mengikuti jejaknya. Sedangkan Allah Swt. telah melarang kaum mukmin mengikuti jalan orang-orang yang rusak. Berangkat dari pengertian inilah ada sejumlah ulama yang melarang kita menerima berita (riwayat) dari orang yang tidak dikenal, karena barangkali dia adalah orang yang fasik. Tetapi sebagian ulama lainnya mau menerimanya dengan alasan bahwa kami hanya diperintahkan untuk meneliti kebenaran berita orang fasik, sedangkan orang yang tidak dikenal (mahjul) masih belum terbukti kefasikannya karena dia tidak diketahui keadaannya. Mujahid dan Qatadah menceritakan bahwa Rasulullah Saw. mengirimkan Al-Walid ibnu Uqbah kepada Banil Mustaliq untuk mengambil harta zakat mereka. Lalu Banil Mustaliq menyambut kedatangannya dengan membawa

zakat (yakni berupa ternak), tetapi Al-Walid kembali lagi dan melaporkan bahwa sesungguhnya Banil Mustaliq telah menghimpun kekuatan untuk memerangi Rasulullah. Menurut riwayat Qatadah, disebutkan bahwa selain itu mereka murtad dari Islam. Maka Rasulullah Saw. mengirim Khalid ibnul Walid r.a. kepada mereka, tetapi beliau Saw. berpesan kepada Khalid agar meneliti dahulu kebenaran berita tersebut dan jangan cepat-cepat mengambil keputusan sebelum cukup buktinya. Khalid berangkat menuju ke tempat Banil Mustaliq, ia sampai di dekat tempat mereka di malam hari. Maka Khalid megirim mata-matanya untuk melihat keadaan mereka; ketika mata-mata Khalid kembali kepadanya, mereka menceritakan kepadanya bahwa Banil Mustaliq masih berpegang teguh pada Islam, dan mereka mendengar suara azan di kalangan Banil Mustaliq serta suara salat mereka, Maka pada keesokan harinya Khalid r.a. mendatangi mereka dan melihat hal yang menakjubkan dirinya di kalangan mereka, lalu ia kembali kepada Rasulullah Saw. dan menceritakan semua apa yang disaksikannya, lalu tidak lama kemudian Allah Swt. menurunkan ayat ini. Interpretasi dari ketiga surat dapat mengintegrasikan penelitian ini ke dalam Islam karena pentingnya menyampaikan suatu kebenaran dari berita, belajar untuk mebuktikan keaslian suatu berita sehingga memberikan manfaat dari berita yang kita sampaikan, tidak membuat ragu dan merugikan orang lain atas berita yang telah disampaikan.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan dari hasil dilakukannya implementasi dan uji coba, dapat disimpulkan tentang klasifikasi kategori berita menggunakan metode *K-Nearest Neighbor* menunjukkan dengan pengujian sebanyak dua puluh data berita didapatkan hasil presentasi nilai akurasi sebesar 74 %, *precision* 35 %, *recall* 35 %, *F- measure* 35%, *Specificity* 84 % dan UAC 60% . Penggunaan metode *K-Nearest Neighbor* pada klasifikasi kategori berita yang telah dibangun dapat dikatakan bisa digunakan untuk klasifikasi namun tidak efektif karena dengan presentase nilai UAC yang rendah sehingga sistem dinilai buruk dalam melakukan klasifikasi kategori berita.

5.2 Saran

Berdasarkan hasil implementasi dari aplikasi, ditemukan saran-saran pengembangan aplikasi yang selanjutnya dapat dilakukan pada penelitian adalah melakukan hal berikut :

1. Sistem dapat dikembangkan menjadi aplikasi *mobile*, sehingga user dapat melakukan klasifikasi berita pada perangkat *smartphone* yang dimiliki.
2. Untuk penelitian selanjutnya dapat menerapkan algoritma klasifikasi seperti *naïve bayes*, SVM, dan *K-Means* dan menerapkan jumlah data uji lebih banyak sehingga hasil yang diperoleh untuk prediksi klasifikasi lebih akurat.

DAFTAR PUSTAKA

Bahri, Syaiful D., Zain, Aswan. 2010. Strategi Belajar. Jakarta: Renikia Cipta.

Dragut, E., Fang, F., Sistla, P., Yu, C., & Meng, W. 2009. *Stop Word And Related Problem in Web Interface Integration*. VLDB Endowment.

Hardiyanto, Erik, Rahutomo, Faisal. 2016. Studi Awal Klasifikasi Artikel Wikipedia Bahasa Indonesia Dengan Menggunakan Metoda K Nearest Neighbor. Prosiding Sentrinov (Seminar Nasional Terapan Riset Inovatif), [S.l.], v. 2, n. 1, p. 158-165, oct. ISSN 2477-2097. Available at: <http://proceeding.sentrinov.org/index.php/sentrinov/article/view/95>. Date accessed: 30 april 2020.

Kadir, Abdul. 2015. Pengenalan Sistem Informasi. Yogyakarta: Andi.

Lin, S. 2008. *A document classification and retrival system for R&D in semiconductor industry-A hybrid approach*. Expert System 18, 2:4753-4764.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. 2014. *Twitter spammer detection using data stream clustering. Information Sciences*, 260, 64-73.

Naufal, M. Adib. 2017. Implementasi Metode Klasifikasi K-Nearest Neighbor (K-Nn) Untuk Pengenalan Pola Batik Motif Lampung. Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung. 1–46.

Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta : Andi.

Purwanti, E. 2015. Klasifikasi Dokumen Temu Kembali Informasi dengan K-Nearest Neghbour *Information Retrieval Document Classified with K-Nearest Neighbor*. Record and Library Journal. Vol. 1, No. 2, Juli-Desember.

Salton. G. 1989. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. Boston, USA: Addison-Wesley Longman Publishing Co.

Sokolova, M., Lapalme, G. 2009. *A systematic analysis of performance measures for classification tasks, Inf. Process. Manag.*, vol. 45, no. 4, hal. 427–437.

Triana, A., Saptono, R., Sulistyo, M. E. 2014. Pemanfaatan Metode Vector Space Model Dan Cosine Similarity Pada Fitur Deteksi Hama Dan Penyakit Tanaman Padi. *Jurnal ITSMART*, 1–6.

Turban, E., Aronson, J. E., & Liang, T.-P. 2005. Decision Support Systems and Intelligent Systems (Edisi Bahasa Indonesia). (D. Prabandini, Ed.) (7th ed., p. 697). New Jersey: ANDI.

Turland, Matthew. 2010. *Php | architect's guide to Web Scraping with PHP*. Web Scraping Defined, Introduction- 2.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. 2005. *Text mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer. doi:10.1007/978-0-387-34555-0.

Witten, Ian H, Frank, Eibe, & Hal, M.A. 2016. *Data Mining: Pratical Machine Learning Tools and Techniques*, Third Edition. Burlington: Morgan Kaufmann Publishers.

LAMPIRAN

Lampiran 1

Profil Penguji Ahli

Nama	Achwan S.Pd., M.Pd.
Alamat	Jl Sumargo gang Anggrek 51, Lamongan
Tempat, Tanggal Lahir	Lamongan, 16 Oktober 1972
Pendidikan	S2 IPS
Profesi	Guru