KLASIFIKASI ABSTRAK ARTIKEL ILMIAH BIDANG SAINS MENGGUNAKAN METODE *K-NEAREST NEIGHBOR* (KNN) DAN *SUPPORT VECTOR MACHINE* (SVM)

(Skripsi)

Oleh

Nurul Dawati Adawiyah

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS LAMPUNG

BANDAR LAMPUNG

2019

ABSTRAK

CLASSIFICATION ABSTRACT SCIENTIFIC ARTICLES OF SCIENCE USING K-NEAREST NEIGHBOR (KNN) AND SUPPORT VECTOR MACHINE (SVM) METHODS

Oleh

NURUL DAWATI ADAWIYAH

Scientific articles are evidence of papers produced by researchers conducting research on specific topics or fields. The number of scientific articles is increasing very fast especially with the development of the internet. Large amounts of data will be very difficult to classify documents manually because it will require a long time and requires a very high degree of accuracy, so we need techniques in the field of document processing to determine the classification of certain topics automatically. Techniques that can do classification quickly are using text *mining* techniques. In this study, the method used in classification is SVM and KNN. KNN has advantages in terms of considerable training data, as a comparison, in this study also uses SVM because this method is one method that is widely used for data verification, specifically data text. These two methods will be compared to find out the best classification accuracy results. The results of this study that the Linear Kernel SVM, Polynomial Kernel and RBF produce the best classification

accuracy is polynomial kernel when compared with KNN with the results of the accuracy, recall, precision and F-Measure values of 80%, 76%, and 77%.

Keywords: *K-Nearest Neighbor, Support Vector Machines, Text Mining*

ABSTRAK

KLASIFIKASI ABSTRAK ARTIKEL ILMIAH BIDANG SAINS MENGGUNKAN METODE *K-NEAREST NEIGHBOR* (KNN) DAN *SUPPORT VECTOR MACHINE* (SVM)

By

NURUL DAWATI ADAWIYAH

Artikel ilmiah adalah bukti dari karya tulis yang dihasilkan oleh peneliti yang melakukan penelitian pada topik atau bidang tertentu. Jumlah artikel ilmiah meningkat sangat cepat terutama dengan perkembangan internet. Data dalam jumlah besar akan sangat sulit untuk mengklasifikasikan dokumen secara manual karena akan membutuhkan waktu yang lama dan membutuhkan tingkat akurasi yang sangat tinggi, sehingga kami membutuhkan teknik di bidang pemrosesan dokumen untuk menentukan klasifikasi topik tertentu secara otomatis. Teknik yang dapat melakukan klasifikasi dengan cepat adalah dengan menggunakan teknik *text mining*. Penelitian ini menggunakan metode klasifikasi SVM dan KNN. KNN memiliki kelebihan dalam hal data pelatihan yang cukup banyak sebagai komparasi, dalam penelitian ini juga menggunakan SVM karena metode ini merupakan salah satu metode yang banyak digunakan untuk kasifikasi data, khusus nya data text. Kedua metode ini akan dibandingkan untuk mengetahui hasil ketepatan klasifikasi yang paling baik. Hasil dari penelitian ini bahwa SVM *Kernel Linier, Kernel polynomial* dan *RBF* menghasilkan ketepatan klasifikasi

yang paling baik adalah *kernel polynomial* apabila dibandingkan dengan KNN dengan hasil nilai akurasi, *recall*, *precision* dan *F-Measure* sebesar 80%, 76%, dan 77%.

Kata kunci: *K-Nearest Neighbor, Support Vector Machine, Text Mining*

KLASIFIKASI ABSTRAK ARTIKEL ILMIAH BIDANG SAINS MENGGUNAKAN METODE *K-NEAREST NEIGHBOR* (KNN) DAN *SUPPORT VECTOR MACHINE* (SVM)

Oleh

NURUL DAWATI ADAWIYAH

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA KOMPUTER

Pada

Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS LAMPUNG

BANDAR LAMPUNG

2019

Judul Skripsi	:	**KLASIFIKASI ABSTRAK ARTIKEL ILMIAH BIDANG SAINS MENGGUNAKAN METODE *K-NEARS NEIGHBOR* (KNN) DAN *SUPPORT VECTOR MACHINE* (SVM)**
Nama Mahasiswa	:	Nurul Dawati Adawiyah
No. Pokok Mahasiswa	:	1517051038
Jurusan	:	Ilmu Komputer
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Rico Andrian, S.Si., M.Kom.	**Ardiansyah, S.Kom., M.Kom.**
NIP. 19750627 200501 1 001	NIP. 19870128 201803 1 001

2. Mengetahui
Ketua Jurusan Ilmu Komputer
FMIPA Universitas Lampung

Dr. Ir. Kurnia Muludi, M.S.Sc.
NIP. 19640616 198902 1 001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Rico Andrian, S.Si., M.Kom.** .....

   Sekretaris : **Ardiansyah, S.Kom., M.Kom.** .....

   Penguji
   Bukan Pembimbing : **Aristoteles, S.Si., M.Si.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Drs. Suratman, M.Sc.
NIP. 19640604 199003 1 002

Tanggal Lulus Ujian Skripsi : **22 November 2019**

PERNYATAAN

Saya yang bertanda tangan di bawah ini, menyatakan bahwa skripsi saya yang berjudul "Klasifikasi Artikel Ilmiah Bidang Sains Menggunakan Metode *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM)" merupakan karya saya sendiri dan bukan karya orang lain. Semua tulisan yang tertuang di skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti skripsi saya merupakan hasil penjiplakan atau dibuat orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar yang telah saya terima.

Bandar Lampung, 12 Desember 2018

NURUL DAWATI ADAWIYAH

NPM : 1517051038

RIWAYAT HIDUP

Penulis dilahirkan pada tanggal 27 juli 1997 di Menggala merupakan anak ketiga dari empat bersaudara dengan Ibu Rusminar dan ayah Tafrizal. Penulis menyelesaikan pendidikan formal pertama kali di Taman Kanak-kanak (TK) Hidayatullah Menggala pada tahun 2002, menyelesaikan Sekolah Dasar (SD) di SD Islam Hidayatullah Menggala pada tahun 2009, menyelesaikan Sekolah Menengah Pertama (SMP) di MTS Al-izzah Menggala pada tahun 2012, kemudian melanjutkan jenjang Sekolah Menengah Atas (SMA) di SMAN Menggala.

Pada tahun 2015. Penulis terdaftar sebagai mahasiswi Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SNPTN. Pada bulan januari - maret 2015. Penulis melakukan kerja praktek (KP) di Badan Pusat Statistik (BPS) Provinsi Lampung selama 40 hari. Kemudian pada bulan juli 2018 penulis melakukan Kuliah Kerja Nyata (KKN) selama 32 hari di desa Tanjungan Kabupaten Tulang Bawang Barat. Selama menjadi mahasiswi, penulis aktif dalam beberapa organisasi dan kegiatan kemahasiswaan antara lain:

1. Aktif dalam Organisasi Himpunan Mahasiswa Jurusan Ilmu Komputer (Himakom) Universitas Lampung dengan menjabat sebagai Anggota Bidang kestari pada tahun 2015-2016
2. Aktif dalam Organisasi Rohani Islam (ROIS) FMIPA Universitas Lampung dengan menjabat sebagai Anggota Bidang Informasi dan Komunikasi (Infokom) pada tahun 2015-2017.
3. Aktif dalam Organisasi Badan Eksekutif Mahasiswa (BEM) FMIPA Universitas Lampung dengan menjabat sebagai Anggota Departemen Komunikasi dan Informasi (Kominfo) pada tahun 2016 dan sebagai Bendahara Bidang Pengembangan Sains dan Lingkungan Hidup (PSLH)
4. Peserta Karya Wisata Ilmiah (KWI) di Pekon Batutegi, Kecamatan Air Naningan, Kabupaten Tanggamus pada bulan Januari 2016.

MOTTO

Barang siapa yang hari ini sama dengan hari kemarin, maka ia termasuk orang yang merugi. Dan barang siapa yang hari ini lebih baik dari hari kemarin, maka ia termasuk orang yang beruntung."
(HR. Bukhari)

" Sesungguhnya sesudah kesulitan itu ada kemudahan, maka apabila kamu telah selesai (dari suatu urusan) kerjakan dengan sesungguhnya(urusan) yang lain dan hanya kepada Tuhanmulah hendaknya kamu berharap."

(Al-Insyiroh: 6-8)

Jangan hanya menghindari yang tidak mungkin. Dengan mencoba sesuatuyang tidak mungkin,anda akan bisa mencapai yang terbaik dari yangmungkin anda capai

(Mario Teguh)

"Man Jadda wa Jadda, Man Shobaro Zhafira".

PERSEMBAHAN

Puji dan syukur saya panjatkan kepada Allah SWT atas segala berkah-Nya yang telah diberikan kepada saya sehingga skripsi ini dapat terselesaikan.

Kupersembahkan karya kecil ini kepada :

Teruntuk kedua orang tuaku Bapak Tafrizal dan Ibu Rusminar yang selalu memberikan do'a, nasehat, serta segala dan upayanya demi tercapai harapan dan cita-citaku. Tidak tercapai semua keinginan dan cita-citaku tanpa kehadiran kalian.

Keluarga besar yang telah memberikan apresiasi.

Keluarga Ilmu Komputer 2015

Serta Almamater tercinta

Universitas Lampung.

SANWACANA

Assalamualaikum wr. wb.

Alhamdulillah, segala puji bagi Allah SWT yang telah melimpahkan rahmat, hidayah, kesehatan dan karunia-Nya sehingga penulis dapat menyelesaikan penulisan skripsi yang berjudul "Klasifikasi Artikel Ilmiah Bidang Sains Mengguanakn Metode *K-Nearest Neighbor* (K-NN) dan *Support Vector Machine* (SVM)" dengan baik. Terima kasih penulis ucapkan kepada semua pihak yang telah membantu dan berperan besar dalam menyusun skripsi ini, antara lain:

1. Kedua orangtua tercinta, Ibu Rusminar dan Ayah Tafrizal yang telah memberikan doa, kasih sayang, dukungan dan semangat yang tak terhingga serta memfasilitasi kebutuhan untuk menyelesaikan skripsi ini.
2. Bapak Rico Andrian, S.Si., M.Kom. sebagai pembimbing utama yang telah membimbing, memotivasi serta memberikan ide, kritik dan saran selama masa perkuliahan dan penyusunan skripsi sehingga penulis bisa sampai ditahap ini.
3. Bapak Ardiansyah S.Kom., M.Kom. sebagai pembimbing kedua yang telah memberikan banyak kritik dan saran yang bermanfaat untuk perbaikan dalam penyusunan skripsi ini.

4. Bapak Aristoteles, S.Si., M.Si. sebagai pembahas yang telah memberikan banyak kritik dan saran yang bermanfaat untuk perbaikan dalam penyusunan skripsi ini.
5. Bapak Drs. Suratman, M.Sc. sebagai Dekan FMIPA Universitas Lampung.
6. Bapak Dr. Ir. Kurnia Muludi, M.S.Sc. sebagai Ketua Jurusan Ilmu Komputer FMIPA Universitas Lampung.
7. Bapak Didik Kurniawan, S.Si., M.T. sebagai Sekretaris Jurusan Ilmu Komputer FMIPA Universitas Lampung yang telah banyak membantu penulis selama perkuliahan.
8. Febi Eka Febriansyah, ST. sebagai Pembimbing Akademik yang selalu memberi nasihat dan motivasi selama penulis menjadi mahasiswi.
9. Bapak dan Ibu Dosen Jurusan Ilmu Komputer yang telah memberikan ilmu selama Penulis menjadi mahasiswa.
10. Mas Naufal dan Mas Sam yang telah membukakan MIPA Terpadu dan ruang baca serta menyiapkan ruang seminar.
11. Kakak Pajar sebagai kakak mentor selama penyusunan skripsi ini.
12. Sahabat Tersayang Fatiya Hasanah yang tak kenal lelah dalam memberikan doa, semangat, motivasi dan dukungannya dalam hal apapun.
13. Teman seperjuangan selama kuliah di Jurusan Ilmu Komputer, khususnya Devi Maharani, Silviyah, Eliza, dan kelas A angkatan 2015 yang lainnya yang tidak bisa disebutkan satu persatu namanya yang selalu memberikan doa, semangat, dan kebersamaannya selama ini.

14. Teman Kerja Praktik yang selalu bekerja sama dan Keluarga BPS Provinsi Lampung yang telah memberikan pengalaman berharga selama kerja praktik.
15. Teman-teman Ilmu Komputer Angkatan 2015 yang sama-sama memperjuangkan mimpi dan masa depannya. Keluarga besar yang telah memberikan doa dan *support* yang tiada henti-hentinya.
16. Almamater tercinta, Universitas Lampung.

Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan, akan tetapi sedikit harapan semoga skripsi ini bermanfaat bagi perkembangan ilmu pengetahuan terutama bagi teman-teman Ilmu Komputer.

Bandar Lampung 12 Desember 2019

Penulis

Nurul Dawati Adawiyah

DARTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

I. PENDAHULUAN

1.1 Latar Belakang

Artikel ilmiah merupakan salah satu bukti karya tertulis yang dihasilkan oleh seorang peneliti yang melakukan penelitian pada suatu topik atau bidang tertentu. Artikel-artikel tersebut umumnya diterbitkan oleh lembaga-lembaga penyedia informasi ilmiah maupun institusi perguruan tinggi. Jumlah artikel ilmiah meningkat sangat cepat terutama dengan adanya perkembangan internet karena untuk menerbitkan artikel diluar institusi baik dalam maupun luar negeri saat ini lebih mudah dengan proses yang lebih cepat. Artikel penelitian yang telah diterbitkan oleh Elselvier setiap tahunnya adalah 420.000 sehingga sampai sekarang secara total terdapat sekitar lebih dari 25 juta dokumen. Jumlah data yang begitu besar akan sangat sulit untuk melakukan klasifikasi dokumen secara manual karena akan membutuhkan waktu yang lama dan dibutuhkan tingkat ketelitian yang sangat besar, sehingga dibutuhkan suatu teknik dalam bidang pemrosesan dokumen untuk menentukan klasifikasi topik tertentu secara otomatis. Teknik yang dapat melakukan klasifikasi secara cepat adalah dengan *text mining* (Widaningsi, 2018).

Text mining adalah untuk menemukan pola-pola informasi yang dapat digali dari suatu teks yang tidak terstruktur (Hamzah, 2012). Data teks dianalisis menggunakan metode dalam *text mining,* perlu dilakukan *text preprocessing* diantaranya adalah *tokenizing*, *case folding*, *stopwords*, dan *stemming*. *Text mining* merupakan suatu proses mencari informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen, namun kebanyakan dokumen tidak diklasifikasi atau dikelompokan sesuai dengan kelompoknya sehingga dokumen-dokumen yang berhubungan sulit untuk ditemukan, seperti pada dokumen abstrak artikel (Latif, 2018).

Abstrak artikel merupakan bentuk mini dari sebuah tulisan ilmiah yang di dalamnya mengandung hal-hal penting dari tulisan ilmiah. Penggalian informasi pada artikel dapat dilakukan dengan memanfaatkan bagian abstrak untuk menemukan kata kunci dari artikel, sehingga hal ini diperlukannya proses *text mining* dengan mengklasifikasikan abstrak artikel agar mempermudah pengguna untuk mengetahui kategori artikel sesuai dengan kelasnya (Latif, 2018).

Klasifikasi merupakan suatu metode untuk memprediksi kategori atau kelas dari suatu *item* atau data yang telah didefinisikan sebelumnya. Metode klasifikasi banyak digunakan dalam melakukan klasifikasi berupa teks diantaranya adalah *Naïve Bayes Classifier* (NBC), *K-Nearest Neigbour* (KNN), *Artificial Neural Network* (ANN), dan *Support Vector Machines* (SVM) (Han dan Kamber, 2006).

Penelitian menggunakan metode *Support Vector Machines* (SVM) dan *K-Nearest Neigbour* (KNN) pernah dilakukan oleh Asiyah dan Fithriasari (2016) tentang klasifikasi berita *online*. Kedua metode ini dibandingkan untuk mengetahui hasil ketepatan klasifikasi yang paling baik. Hasil dari penelitian ini bahwa SVM lebih baik dari pada KNN dengan hasil nilai akurasi, *recall*, *precision* dan *F-Measure* sebesar 93.2%, 93.2%, 93.63% dan93.14%. Ariadi dan Fithriasari (2015) juga membandingkan 2 metode klasifikasi yaitu metode NBC dan SVM dengan *Confix Stipping Stemmer*. Metode NBC menunjukan akurasi sebesar 82,2% sedangkan metode SVM mencapai akurasi 88,1%.

Penelitian lain dilakukan Claudy, Perdana dan Fauzi, et.al, (2018) tentang klasifikasi dokumen Twitter menggunakan algoritma *K-Nearest Neigbour* (KNN). Data yang digunakan yaitu data dari Twitter para calon karyawan di suatu perusahaan. Data tersebut berjumlah 160 data yang dibagi menjadi dua bagian yaitu 50% dari total data atau 80 data untuk data latih dan 50% dari total data atau 80 data untuk data uji. Kelas atau kelompok yang digunakan dalam klasifikasi terdapat empat jenis yaitu artisan, *Guardian, Idealist*, dan Rasional. Data uji diproses dalam beberapa tahapan antara lain *text preprocessing*, pembobotan kata dan klasifikasi algoritma KNN. Penelitian klasifikasi kepribadian atau karakter calon karyawan dengan algoritma KNN menghasilkan nilai akurasi sebesar 66% dengan jumlah data yang mendapatkan hasil klasifikasi dengan benar yaitu 53 data dan jumlah data yang mendapatkan hasil klasifikasi dengan salah yaitu 27 data.

Algoritma KNN tetap dapat melakukan klasifikasi kepribadian atau karakter dengan baik sehingga menghasilkan nilai akurasai sebesar 66%.

Penelitian ini fokus pada klasifikasi abstrak artikel ilmiah menggunakan metode *Support Vector Machines* (SVM) dan *K-Nearest Neigbour* (KNN). KNN memiliki kelebihan dalam hal data *training* yang cukup banyak. Penelitian ini juga menggunakan metode SVM untuk klasifikasi data, khususnya data teks. Metode SVM memiliki kelebihan yang dapat diimplementasikan *relative* mudah, karena proses penentuan *support vector* dapat dirumuskan dalam QP *problem*. Penelitian ini diharapkan dapat membandingkan metode KNN dan SVM pada klasifikasi Abstrak Artikel Ilmiah Bidang Sains.

1.2 Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah bagaimana kinerja klasifikasi abstrak artikel ilmiah bidang sains menggunakan metode *K-Nearest Neigbour* (KNN) dan *Support Vector Machines* (SVM).

1.3 Batasan Masakah

Batasan masalah pada penelitian ini adalah:

1. Metode klasifikasi menggunakan metode *Nearest Neigbour* (KNN) dan *Support Vector Machines* (SVM).
2. Proses pengklasifikasi dilakukan pada abstrak artikel ilmiah bidang sains berbahasa Indonesia.
3. Abstrak yang digunakan bersumber dari situs jurnal.mipa.unila.ac.id.
4. Jenis *file* abstrak artikel yang digunakan dalam bentuk *TXT*.

1.4 Tujuan

Tujuan dari penelitian ini adalah membandingkan kinerja *K-Nearest Neigbour* (KNN) dengan kinerja *Support Vector Machines* (SVM) pada Abstrak Artikel Ilmiah Bidang Sains.

1.5 Manfaat

Manfaat dari penelitian ini adalah :

1. Memperoleh hasil klasifikasi abstrak dalam banyak dokumen sekaligus dengan akurasi yang baik.
2. Menjadikan penelitian ini sebagai bahan rujukan penelitian lain mengenai klasifikai abstrak artikel ilmiah bidang sains menggunakan metode *K-Nearest Neigbour* (KNN) dan *Support Vector Machines* (SVM).

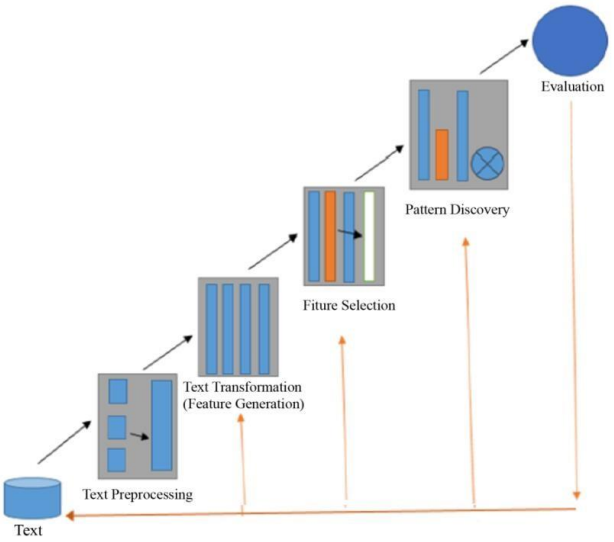
II. TINJAUAN PUSTAKA

2.1 *Text Mining*

Text Mining merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi, penerapan *text mining* pada konsep dari data *mining* yang mencari pola menarik dari sekumpulan data tekstual yang berjumlah besar berguna memberikan informasi yang bermanfaat untuk tujuan tertentu. Tujuan dari pengolahan teks yaitu untuk mendapatkan informasi yang berguna dari sekumpulan dokumen dengan mengekstrak informasi dari data dengan mengidentifikasi dan mencari pola menarik dari sekumpulan data. Hasil yang didapat yaitu kata-kata yang mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen (Kurniawan, dan Efendi, 2012).

Informasi biasanya diperoleh melalui peramalan pola dan kecenderungan pembelajaran pola statistik. *Text mining* yaitu *parsing*, bersama dengan penambahan beberapa fitur linguistik turunan dan penghilangan beberapa diantaranya, dan penyisipan *subsequent* ke dalam database, menentukan pola dalam data terstruktur, dan akhirnya mengevaluasi dan menginterpretasi *output*. *Text mining* biasanya merujuk ke beberapa kombinasi relevansi, pembaharuan, dan *interestingness*, proses *text mining*

yang khas meliputi kategorisasi teks, *text clustering*, ekstraksi konsep/entitas, produksi taksonomi granular, *sentiment analysis*, penyimpulan dokumen, dan pemodelan relasi entitas yaitu, pembelajaran hubungan antara entitas. Proses *text mining* diperlihatkan pada Gambar 1 (Kumar dan Bhatia, 2013).



Gambar 1. Proses *Text Mining* (Kumar dan Bhatia, 2013).

Text mining mencoba untuk mengekstrak informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi dari suatu pola menarik. Sumber data berupa sekumpulan dokumen dan pola menarik yang tidak ditemukan dalam bentuk *database record*, tetapi dalam data teks yang tidak terstruktur (Fildman dan Sanger, 2007).

Tahapan proses pokok dalam *text mining* yaitu pemrosesan awal *text (text preprocessing*), transformasi teks (*text transformation*)/ (*Feature Generation*) pemilihan fitur (*feature selection*), dan penemuan pola *text/data mining* (*pattern discovery*) (Even dan Zohar 2002).

2.1.1 Text

Tahap pertama adalah permasalahan yang dihadapi pada *text mining* sama dengan permasalahan yang terdapat pada data *mining*, yaitu jumlah data yang besar, dimensi yang tinggi, data struktur yang terus berubah, dan data *noise*. Perbedaan di antara keduanya adalah pada data yang digunakan. *Data mining* yang digunakan adalah *structured* data, sedangkan pada *text mining*, data yang digunakan adalah minimal semi *structured*, hal ini menyebabkan adanya tantangan tambahan pada *text mining* yaitu struktur *text* yang komplek dan tidak lengkap, arti yang tidak jelas dan tidak standar dan bahasa yang berbeda ditambah tranlasi yang tidak akurat (Latif, 2018).

2.1.2 Preprocessing Text

Tahap ini melakukan analisis semantik (kebenaran arti) dan statistik (kebenaran susunan) terhadap teks. Tujuan dari pemrosesan awal adalah untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan lebih lanjut. Operasi yang dilakukan pada tahap *Preprocessing Text* Menurut Fildman dan Sanger (2007) yaitu:

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.

b. *Tokenizing*, merupakan proses memecah yang semula berapa kalimat menjadi kata-kata atau memutus urutan string menjadi potongan-potongan seperti kata-kata berdasarkan tiap kata yang menyusunnya.

c. *Stopwords*, merupakan kosa kata yang bukan merupakan kata unik atau ciri pada suatu dokumen atau tidak menyampaikan pesan apapun secara signifikan pada teks atau kalimat. Kosa kata yang dimaksudkan 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 menghilangkan awalan, akhiran, sisipan, dan *confixes* (kombinasi dari awalan dan akhiran).

2.1.3 Text Transformation (Feature Generation)

Dokumen teks diwakili oleh kata-kata (fitur) yang mengandung suatu kejadian. Representasi dokumen merupakan salah satu teknik yang digunakan untuk mengurangi kompleksitas dokumen sehingga lebih mudah untuk ditangani. Dokumen harus ditransformasikan dari versi *text* kedalam bentuk *vector* dokumen. Representasi dokumen yang umum digunakan disebut *vector space model* atau Model Ruang Vektor dimana dokumen direpresentasikan menjadi vektor dari kata-kata, beberapa keterbatasannya adalah representasi dimensi yang tinggi, hilangnya korelasi dengan kata-kata yang berdekatan dan

hilangnya hubungan semantik yang ada di dalam dokumen. Mengatasi masalah ini, metode pembobotan kata digunakan untuk menetapkan bobot yang sesuai dengan kata tersebut. Model ruang vector membuat dokumen diwakili oleh vektor kata-kata karena kata-kata yang membentuk dokumen akan menentukan isi dari dokumen tersebut (Sukhjit, Sehra dan Nayyar, 2013).

2.1.4 Feature Selection

Tahap ini adalah teknik pengurangan dimensi yang efektif untuk menghilangkan fitur *noise.* Fitur *noise* merupakan informasi- informasi yang tidak berguna yang dapat mengganggu hasil penelitian dalam *text mining*, seperti tulisan *copyright*, menu navigasi pada halaman *web*, dan sebagainya. Pemilihan fitur juga dikenal sebagai pemilihan variabel, adalah proses pemilihan *subset* dari fitur yang penting untuk digunakan dalam pembuatan model. Pemilihan fitur ini dilakukan karena data mengandung banyak fitur yang berlebihan atau tidak relevan. Misalkan fitur yang berulang adalah salah satu yang tidak memberikan informasi tambahan. Fitur yang tidak relevan tidak memberikan informasi yang berguna atau relevan dalam konteks apapun. Pemilihan fitur dilakukan dengan menyimpan kata-kata dengan bobot tertinggi sesuai dengan ukuran yang telah ditentukan terhadap pentingnya kata tersebut (Ting dan Tsang, 2011).

3.1.5 Pattern Discovery

Pattern discovery merupakan tahap penting untuk menemukan ciri atau pengetahuan (*knowledge*) dari keseluruhan teks. Data atau *text mining* terdapat dua teknik pembelajaran pada tahap *pattern discovery* ini, yaitu *unsupervised* dan *supervised learning.* Perbedaan antara keduanya adalah pada *supervised learning* terdapat label atau nama kelas pada data latih (*supervisi*) dan data baru diklasifikasikan berdasarkan data latih, sedangkan pada *unsupervised learning* tidak terdapat label atau nama kelas pada data latih, data latih dikelompokkan berdasarkan ukuran kemiripan pada suatu kelas (Even dan Zohar, 2002).

Fungsi *supervised learning* dibagi menjadi 2, regresi dan klasifikasi. *Regresi* terjadi jika output dari fungsi merupakan nilai yang kontinyu, sedangkan klasifikasi terjadi jika keluaran dari fungsi adalah nilai tertentu dari suatu atribut tujuan (tidak kontinyu). Tujuan dari *supervised learning* adalah untuk memprediksi nilai dari fungsi untuk sebuah data masukan yang sah setelah melihat sejumlah data latih (Even dan Zohar, 2002).

3.1.6 Evaluation

Evaluation merupakan hasil dari proses *mining* yang akan diinterpretasikan ke dalam bentuk tertentu untuk dilakukan proses evaluasi. Proses masukan awal dari *text mining* adalah suatu data teks dan menghasilkan keluaran berupa pola sebagai hasil evaluasi

Pola ini diharapkan sesuai pada aplikasi, namun jika hasil tidak sesuai maka dilanjutkan evaluasi dengan melakukan 12 literasi ke satu atau beberapa tahap sebelumnya. Hasil interpretasi merupakan tahap akhir dari proses *text mining* dan akan disajikan ke pengguna dalam bentuk visual (Tan, 1997).

2.2 TF-IDF *(Term Frequency-Inverse Document Frequency)*

Term Frequency Inverse Document Frequency (TF-IDF) merupakan cara untuk memberi bobot pada suatu kata (*term)* terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu *Term Frequency* (*tf*) dan *Inverse Document Frequency* (IDF) Term *Frequency* (*tf*) merupakan pembobotan kata untuk perhitungan jumlah kata yang muncul dalam sebuah dokumen tertentu, sedangkan *Document Frequency* (*idf*) merupakan pembobotan kata untuk perhitungan jumlah kata yang muncul pada seluruh dokumen. Menurut Nasyuha Husaini dan Mursyidah (2016) menyatakan bahwa semakin besar jumlah kemunculan suatu *term* dalam suatu dokumen (*tf*) akan memberikan nilai kesesuaian yang semakin besar dan semakin kecil atau tidak penting nilainya jika suatu kata tersebut muncul dalam banyak dokumen (*idf*), sebaliknya jika kata tersebut jarang muncul dalam banyak dokumen harus diperhatikan dalam pemberian bobot

Langkah dalam *TF-IDF* adalah untuk menemukan jumlah kata yang kita ketahui (*tf*) setelah dikalikan dengan berapa banyak artikel ilmiah dimana suatu kata itu muncul (*idf*). Persamaan *TF-IDF* diprepresentasikan sebagai berikut:

$$w_{ij} = tf_{ij}\, x\, idf \quad (1)$$

$$idf = \log\left(\frac{N}{df}\right)$$

Keterangan :

Wij = Bobot dari kata *i* pada artikel ke *j*

N = Jumlah seluruh dokumen

Tfij = Jumlah kemunculan kata i pada dokumen *j*

Df = Jumlah dokumen yang mengandung kata

2.3 *K-Nearest Neighbor*

K-Nearest Neighbor (KNN) adalah salah satu metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat satu sama lain juga akan memiliki karakteristik yang mirip, KNN adalah improvisasi lanjutan dari teknik klasifikasi *Nearest Neighbor*. Hal ini didasarkan pada gagasan bahwa setiap contoh baru dapat diklasifikasikan oleh suara mayoritas dari *k* tetangga, dimana k adalah bilangan bulat positif, dan biasanya dengan jumlah kecil (Khamis *et,al.*, 2014). Algoritma klasifikasi KNN memprediksi kategori tes sampel sesuai dengan sampel pelatihan *k* yang merupakan tetangga terdekat dengan sampel uji, dan memasukkan ke dalam kategori yang memiliki kategori probabilitas terbesar (Suguna dan Thanushkodi, 2010).

Algoritma KNN pada pengenalan pola sering digunakan sebagai metode untuk mengklasifikasikan objek berdasarkan contoh pelatihan terdekat di

ruang fitur. KNN adalah jenis *insctance-based learning*, atau *lazy learning* dimana fungsi ini hanya didekati secara lokal dan semua perhitungan ditangguhkan sampai klasifikasi (Imandoust dan Bolandraftar, 2013).

Metode klasifikasi KNN memiliki beberapa tahap, yang pertama nilai *k* yang merupakan jumlah tetangga terdekat yang akan menentukan *query* baru masuk ke kelas mana ditentukan. Tahap kedua, *k* tetangga terdekat dicari dengan cara menghitung jarak titik *query* dengan titik *training*. Tahap ketiga, setelah mengetahui jarak masing-masing titik *training* dengan titik *query,* kemudian lihat nilai yang paling kecil. Tahap keempat ambil *k* nilai terkecil selanjutnya lihat kelasnya. Kelas yang paling banyak merupakan kelas dari *query* baru (Pramesti, 2013).

Dekat atau jauhnya jarak titik dengan tetangganya bisa dihitung dengan menggunakan *Euclidean distance*. persamaan *Euclidean distance* direpresentasikan sebagai berikut (Pramesti, 2013):

$$d(a,b) = \sqrt{\sum_{k=1}^{K_n}(a_k - b_k)^2} \qquad (2)$$

Keterangan :

d = Jarak

a = Titik yang sudah diketahui kelasnya.

b = Titik yang akan dicari tau kelasnya.

k = Titik training.

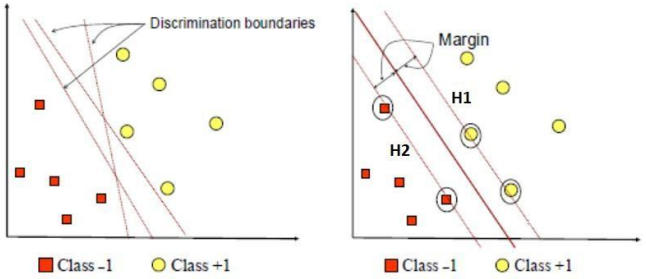
Kn = jumlah *attribute*

d(a,b) merupakan jarak antara titik *a* yang merupakan titik yang telah diketahui kelasnya dan *b* berupa titik baru. Jarak antara titik baru dengan titik-titik *training* dihitung dan diambil *k* buah titik terdekat. Titik baru diprediksi masuk ke kelas dengan klasifikasi terbanyak dari titik-titik tersebut (Pramesti, 2013).

2.4 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah metode klasifikasi yang bekerja dengan cara mencari *hyperplane* dengan *margin* terbesar. *Hyperplane* adalah garis batas pemisah data antar kelas. *Margin* adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas, adapun data terdekat dengan *hyperplane* pada masing-masing kelas inilah yang disebut *support vector* (Yunliang, Qing dan Xiongtao, 2010).

Algoritma SVM pada dasarnya digunakan untuk proses klasifikasi antara dua kelas atau *binary classification.* SVM juga bisa digunakan untuk klasifikasi *multi-class* dengan cara mengkombinasikan beberapa *binary classifier*. Konsep SVM pada dasarnya yaitu upaya pencarian *hyperplane* terbaik sebagai pemisah antara dua *class* dalam *input space* yang memiliki margin atau jarak *hyperplane* terdekat pada setiap kelas (Somantri dan Apriliani, 2018).



Gambar 2. SVM pada kasus *linier* (Nugroho,Witarto dan Handoko, 2003).

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada *input space*. Gambar 2 menjelaskan *hyperplane* optimal H memverifikasi bahwa *hyperplane H1* dan *H2* paralel dan melewati titik terdekat ke H yang disebut *Support Vectors* (SVs), dengan demikian, SVM memilih *hyperplane* optimal yang memaksimalkan margin antara dua kelas yang merupakan jarak antara *H1* dan *H2*. *M*argin terbesar dapat memaksimalkan nilai jarak antara *hyperline* dan titik terdekat, yaitu 1/||w||. Hal ini dapat dirumuskan sebagai *Quadratic Programming* (QP) *problem,* yaitu mencari titik minimal di representasikan pada pada Persamaan 3, dan constrain di representasikan pada Persamaan 4 (Zhang, 2008):

$$min\,\frac{1}{2}\|w\|^2 \tag{3}$$

$$y_i(wx_i + b) - 1 \;\geq = 0 \tag{4}$$

Keterangan :

xi = data masukan.

yi = data keluaran

w & *b* = data yang akan dicari parameternya

Data jika tidak dapat dipisahkan secara *linier* seperti kasus pada Gambar 2, maka data dipetakan ke dalam ruang *non-linier* berdimensi tinggi seperti pada Gambar 3, dimana rangkaian pelatihan dipisahkan secara *linier* oleh fungsi *kernel* yang dipersentasikan dalam Persamaan 5 berikut.

$$min\,\frac{1}{2}\|w\|^2 + C\,(\textstyle\sum_{i-1}^{n}\xi_i) \tag{5}$$

$$\text{dengan } y_i(x_i\,w + b) \;\geq 1 - \xi_i(x_i \;\geq 0) \tag{6}$$

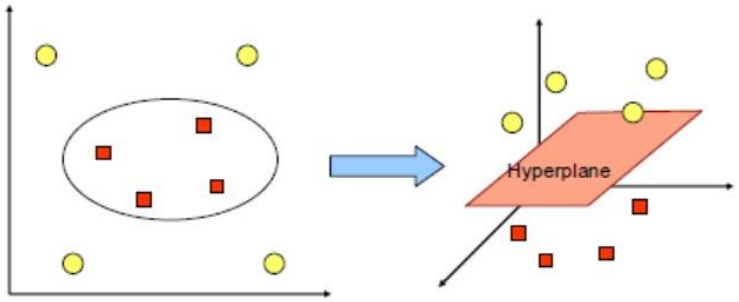
Keterangan :

xi = data masukan.

yi = data keluaran

w & *b* = data yang akan dicari parameternya

ξ = *Soft Margin Hyperline*



Gambar 3. SVM pada kasus non-linier (Nugroho, Witarto dan Handoko 2003).

Kernel trick memberikan berbagai kemudahan, karena dalam proses pembelajaran SVM, untuk menentukan *support vector*, kita hanya cukup mengetahui fungsi *kernel* yang dipakai dan tidak perlu mengetahui wujud dari fungsi *non linear* Φ, selanjutnya hasil klasifikasi dari data *x* diperoleh dari persamaan berikut :

$$\{H(x) = \sum_s \alpha_i y_i x^T x_i + b \quad (7)$$
$$b = y_i - w^T x_i$$

Keputusan pada tiap kelas dapat dibuat dengan aturan sebagai berikut.

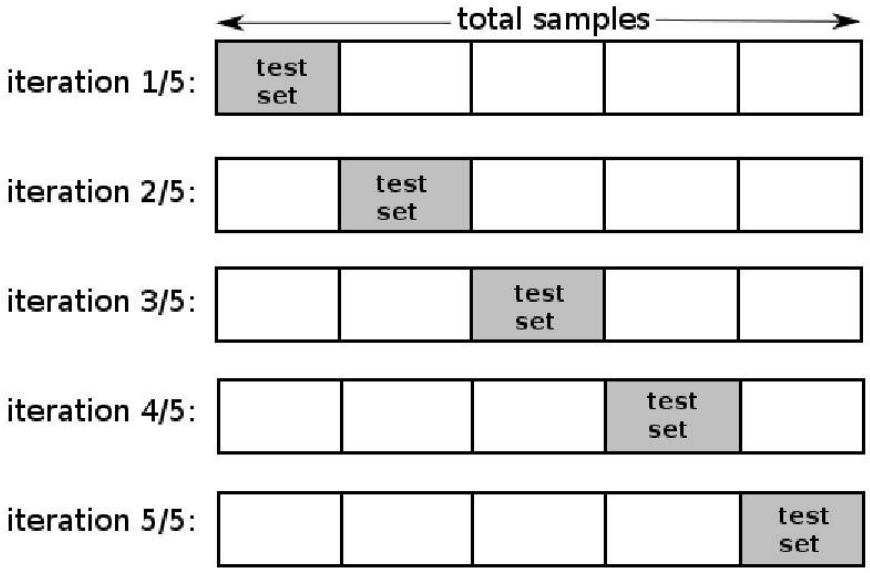
If *H (x)*>0 dimana x ε+1

If *H (x)*<0 dimana x ε -1

If *H (x)*=0 dimana *x* tidak terklasifikasi

2.5 *K-fold Cross Validation*

K-fold cross validation merupakan sebuah teknik yang menggunakan keseluruhan data *set* yang ada sebagai *training* dan *testing* (Bengio, 2004). *K-Fold Cross-validation* adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model klasifikasi dimana data dipisahkan menjadi dua bagian yaitu data proses latihan (*training*) dan data uji (*Testing). K-Fold Cross Validation* digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. Gambar 4 berikut merupakan ilustrasi singkat dari *k-fold cross validation.*



Gambar 4. Ilustrasi cara kerja *k-fold Cross Validation* (Bengio, 2004)

Ilustrasi yang terdapat pada gambar ini dapat diketahui bahwa percobaan menggunakan *5-fold cross validation*. Hal ini ditandai dengan kotak berwarna abu-abu pada gambar yang merupakan *test set* dan kotak berwarna putih merupakan *training set*, sebagai contoh terdapat 50 *instance* pada ilustrasi gambar ini yang merupakan total *number of*

examples pada percobaan 1 (*experiment* 1). Kotak berwarna abu-abu merupakan *test set* yang berisi 10 *instance* dengan id *instance* dari 1-10 dan kotak berwarna putih merupakan *training set* yang berisi 40 *instance* dengan *id instance* dari 11-50, pada percobaan 1 didapatkan hasil nilai rata-rata eror, kemudian dilanjutkan ke percobaan 2. Kotak berwarna abu-abu pada percobaan 2 merupakan *test* set yang berisi 10 *instance* dengan *id instance* dari 11-20 sedangkan kotak berwarna putih merupakan *training set* yang berisi 40 *instance* dengan *id instance* dari 1-10 dan 21-50, pada percobaan 2 didapatkan hasil nilai rata-rata *eror*. cara yang sama dilakukan pada percobaan 3, 4 dan 5, selanjutnya akan dihitung rata-rata keseluruhan estimasi eror dari setiap percobaan yang telah dilakukan hingga percobaan terakhir.

2.6 Perhitungan Tingkat Akurasi

Perhitungan akurasi merupakan salah satu hal yang penting dalam pengenalan pola. Proses ini dilakukan sebagai salah satu tolak ukur evaluasi dalam suatu sistem. Evaluasi tingkat akurasi sistem merupakan tahapan dalam penelitian dengan tujuan untuk mengetahui pengaruh metode yang digunakan terhadap sistem dan memperoleh parameter terbaik. Pengukuran tingkat akurasi yang sering digunakan adalah dengan menghitung akurasi, *recall, precission* dan *f-measure.* Akurasi merupakan persentase dari total dokumen yang teridentifikasi secara tepat dalam proses klasifikasi. *Recall* adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. *Precision* adalah tingkat

ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. *F-Measure* adalah kompromi dari *recall* dan *precision* untuk mengukur kinerja keseluruan pengklasifikasi. Pengukuran tingkat akurasi dapat dipresentasikan sebagai berikut (Hotho, Andreas dan Augustin, 2005)

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (8)$$

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad (9)$$

$$F= Measure = \frac{2\ x\ Recall\ x\ Precision}{Recall\ x\ Precision}\ x\ 100\% \quad (10)$$

Keterangan :

True Positif (TP) = Jumlah dokumen benar yang terdeteksi benar.

False Positif (FP) = Jumlah dokumen yang salah yang terdeteksi Benar

False Negatif (FN) = Jumlah dokumen yang salah yang terdeteksi salah

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian dilakukan di Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang beralamatkan di Jalan Prof. Dr. Soemantri Brojonegoro No.1 Gedung Meneng, Bandar Lampung. Penelitian ini dilaksanakan pada bulan September-Desember 2018/2019.

3.2 Alat dan Bahan

A. Alat Penelitian

Penelitian ini dilakukan dengan menggunakan alat untuk mendukung dan menunjang pelaksanaan penelitian, antara lain:

1. Perangkat Keras (*Hardware)*

   Perangkat keras yang digunakan dalam klasifikasi Konten Abstrak Artikel ini adalah 1 unit Laptop Asus A442U dengan spesifikasi:

   - *Processor*: Intel® Core™ *i7-7500U CPU @ 2.70GHz 2.90 GHz.*
   - Display : 14 inci,
   - RAM : 8 GB
   - *Storage* : 1 TB HDD

2. Perangkat Lunak (*Software)*

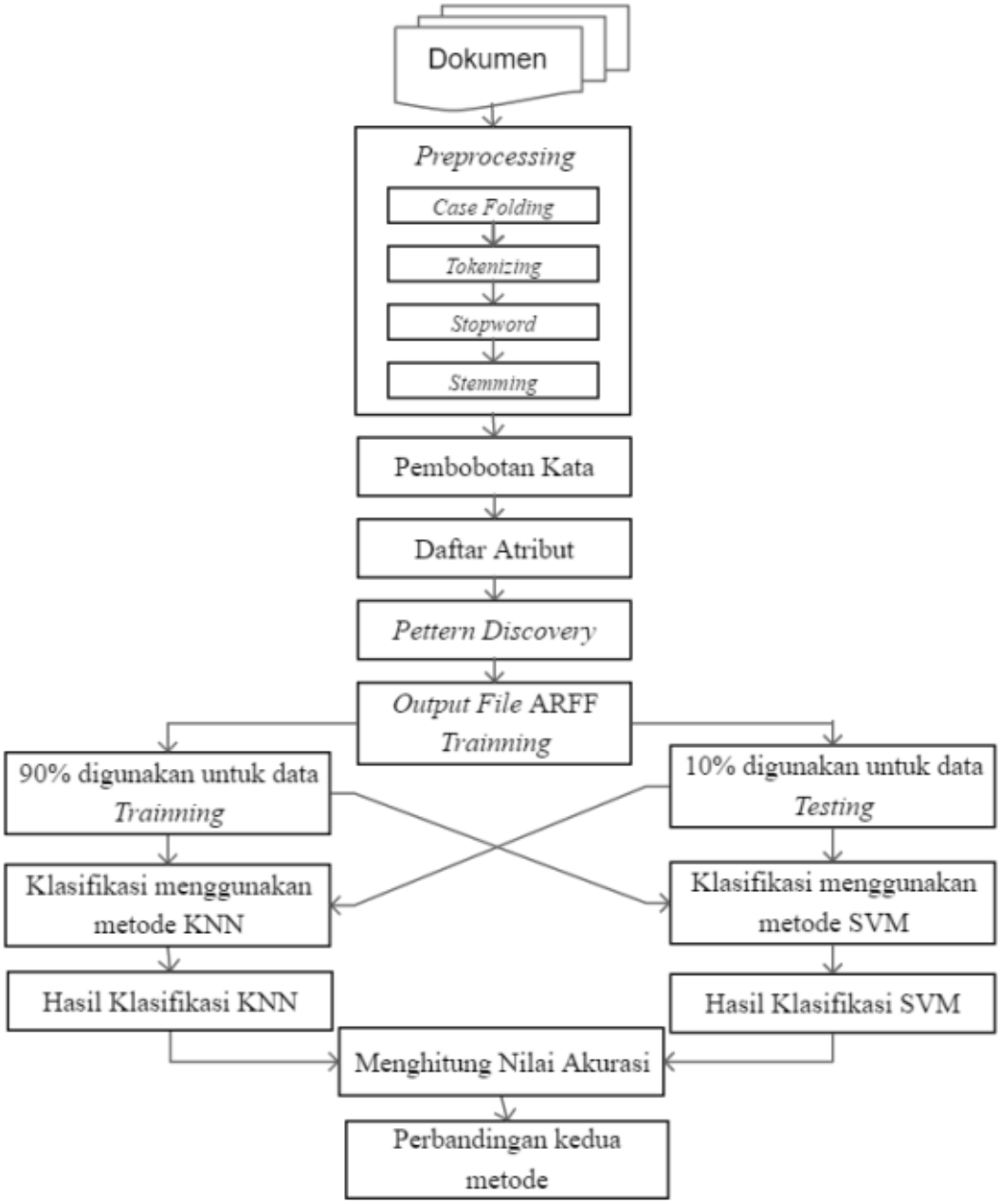
   - Sistem Operasi Windows 10 Pro 64-Bit digunakan untuk menjalankan sistem operasi pada laptop.
   - Weka sebagai aplikasi pengujian untuk membandingkan hasil dari aplikasi yang dibuat.
   - Netbeans sebagai IDE java karena aplikasi ini dibangun dengan bahasa pemrograman Java.
   - Notepad++ sebagai *text editor* untuk melakukan pengeditan *database* karena aplikasi ini menggunakan *database* dalam bentuk txt.

B. Bahan Penelitian

Bahan yang digunakan dalam penelitian ini adalah dokumen abstrak artikel ilmiah bidang sains berbahasa indonesia yang diambil pada situs jurnal.mipa.unila.ac.id yang terdiri dari 5 kategori. Kategori tersebut adalah biologi, fisika, ilmu komputer, kimia dan matematika. Dokumen yang digunakan sebanyak 250 dokumen yang terdiri dari 50 abstak artikel Biologi, 50 abstrak artikel fisika, 50 abstrak artikel kimia, 50 abstrak artikel Matematika dan 50 abstrak artikel Ilmu Komputer.

3.3 Tahapan Penelitian

Tahapan penelitian yang dilakukan dalam klasifikasi artikel ilmiah bidang sains menggunakan metode *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM) ditunjukkan pada Gambar 5.



Gambar 5. Tahapan Penelitian

Tahapan–tahapan yang dilakukan pada penelitian ini adalah sebagai berikut :

A. Pengumpulan Dokumen Artikel

Pengumpulan dokumen artikel yang didapatkan dari situs jurnal.mipa.unila.ac.id. Dokumen artikel dalam penelitian ini menggunakan format *txt*. Penamaan *file* harus sesuai dengan kelasnya, karena proses penentuan kelas diambil dari nama dokumen.

B. Tahap *Preprocessing*

Preprocessing proses yang dilakukan agar dapat mengubah teks dalam bentuk kalimat menjadi kata-kata dan siap untuk diproses ke proses selanjutnya yaitu proses *case folding*, *tokenizing*, *stopword*, dan terakhir *stemming*. Tujuan dilakukan tahapan *Preprocessing* untuk mempersiapkan data agar formatnya sesuai dengan kebutuhan sistem. Proses yang dilakukan pada tahapan ini adalah *case folding*, *tokenizing*, *stopwords*, dan *stemming*, berikut penjelasannya:

1. *Case Folding*

   Tahapan ini yang dilakukan untuk mengubah semua huruf yang ada pada dokumen menjadi huruf kecil. Contoh proses *case folding*, dapat dilihat pada Gambar 6.

Gambar 6. Contoh Proses *Case Folding.*

2. *Tokenizing*

Tahap *tokenizing* dilakukan untuk proses pemotongan tiap-tiap kata pada dokumen yang semula berupa kalimat menjadi kata berdasarkan pembatas berupa spasi dan menghilangkan simbol-simbol, serta karakter angka. Contoh proses *tokenizing*, dapat dilihat pada Gambar 7.

Gambar 7. Contoh Proses *Tokenizing*.

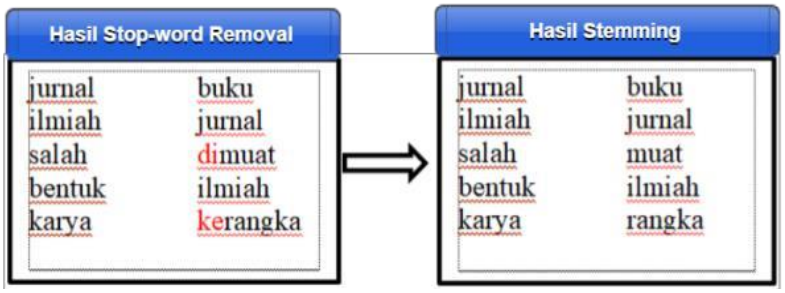
3. *Stopwords Removal*

Tahap *stopword* dilakukan untuk mengambil kata-kata penting dari hasil *tokenizing* dengan cara menghilangkan kata hubung. Daftar *stopwords* diambil dari tesis F.Tala yang Berjudul "*A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia*" Contoh proses *stopwords removal*, dapat dilihat pada Gambar 8.

Gambar 8. Contoh Proses *Stop-word Removal*.

4. *Stemming*

Tahap *stemming* dilakukan untuk proses mencari *root* kata (kata dasar) dari hasil *stop-words. Root* kata (kata dasar) menggunakan *library* Java *(Lucene Bahasa Indonesia)*. Contoh proses *stemming* dapat dilihat pada Gambar 9.



Gambar 9. Contoh Proses *Stemming*

C. Tahap Pembobotan Kata

Pembobotan Kata (*term*) merupakan *term documents matrix* yang representasi kumpulan dokumen yang digunakan untuk melakukan proses klasifikasi dokumen teks. Metode *TF-IDF* adalah metode yang digunakan untuk proses pembobotan pada proses ini, kemudian akan dilakukan pembobotan pada tiap *term* berdasarkan tingkat kepentingan tersebut di dalam sekumpulan dokumen masukan.

D. Tahapan *pattern discovery*

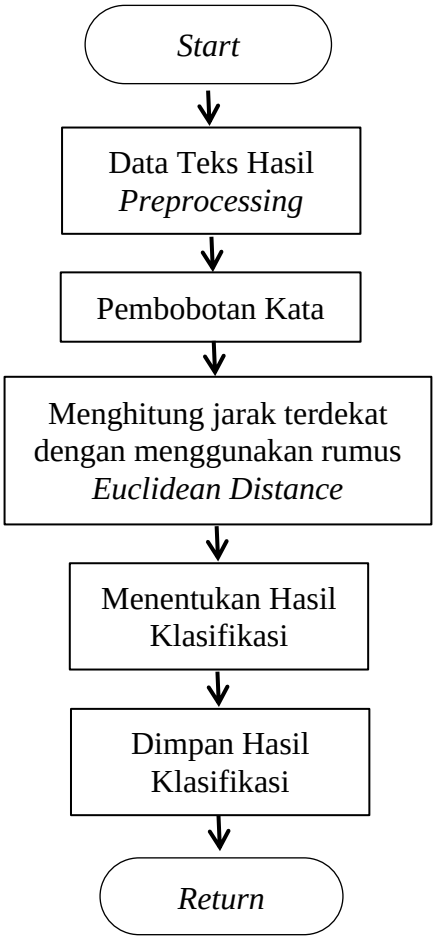
Pattern discovery adalah tahap untuk menemukan ciri-ciri pada setiap dokumen. Tahap ini berusaha menemukan pola atau pengetahuan dari keseluruhan teks, seberapa sering kata itu muncul dari sebuah dokumen.

.

E. Proses Klasifikasi Menggunakan KNN dan SVM

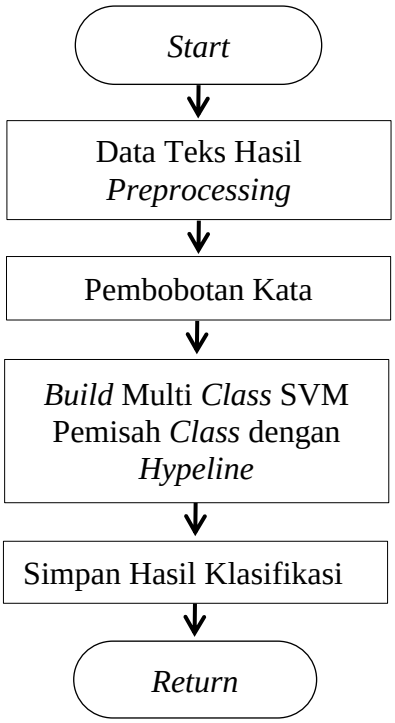
K-Nearest Neigbour (KNN) dan *Support Vector Mechine* (SVM) adalah dua metode yang digunakan pada proses ini. Proses ini menggunakan *library* java yang akan mengubah dokumen dalam bentuk matriks kata dokumen dan telah diberi pembobotan *TF-IDF*. Klasfikasi Menggunakan Metode *K-Nearest Neighbor* (KNN), untuk menghitung nilai kemiripan antara data uji dengan semua data latih pada dokumen artikel dengan melihat jarak terdekat antar tetangganya, Diagram alir *K-Nearest Negbhor* (KNN) dapat dilihat pada Gambar 10 (Lidya, Sitompul dan Efektif, 2015).

\



Gambar 10. Diagram Alir KNN

Proses analisis menggunakan *Support Vector Mechine* (SVM) dimulai dengan mengubah text menjadi data vektor. Vektor dalam penelitian ini memiliki dua komponen yaitu dimensi (*word id*) dan bobot. Bobot ini adalah nilai *tf-idf,* tujuan dari model ruang vektor digunakan untuk memberikan setiap kata dalam dokumen sebuah *id* (dimensi) dan sebuah bobot berdasarkan seberapa penting keberadaannya dalam dokumen (posisi dokumen dalam dimensi tersebut). *Support Vector Mechine* (SVM) mencoba untuk menemukan garis yang terbaik membagi dua kelas, kemudian mengklasifikasi dokumen uji berdasarkan di sisi mana dari garis tersebut mereka muncul. Diagram alir proses klasifikasi dengan SVM ditunjukkan oleh Gambar 11 **(**Seifemichael, Amsalu, Afghah, *et al*., 2015).



Gambar 11. Diagram Alir SVM

F. Hasil klasifikasi artikel menggunakan metode SVM dan KNN selanjutnya dihitung akurasinya dari kedua metode dan membandingkan antara metode SVM dengan metode KNN berdasarkan tingkat akurasi ketepatan klasifikasi.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Simpulan yang diperoleh berdasarkan analisis dan pembahasan yang telah dilakukan adalah sebagai berikut:

1. Metode *Support Vector Machine* dengan menggunakan *kernel linier, polynomial* dan *RBF* didapat hasil kernel *polynomial* lebih baik dari kernel *linier* dan *RBF* pada 250 dokumen, untuk dibandingkan dengan hasil KNN digunakan *kernel polynomial* dengan hasil yang didapatkan pada data *trainning* untuk masing-masing pengukuran performa dari nilai rata-rata 10 *fold* didapatkan nilai rata-rata, *recall, precision,* dan *F-Measure* sebesar 80%, 76%, dan 77%.
2. Metode *K-Nearest Neighbor* dengan menggunakan *k* =1, 3, 5, 7, dan 11 pada data *training* dengan 250 dokumen didapatkan hasil tiap performa dari nilai rata-rata 10 fold didapatkan nilai rata-rata, *recall, precision,* dan *F-Measure* adalah 58,3%, 63,2%, dan 55,7%,
3. Perbandingan antara kedua metode SVM dan KNN didapatkan hasil SVM *kernel polynomial* lebih baik dibandingkan dengan KNN untuk kasus penelitian ini.

5.2 Saran

Saran yang akan diberikan berdasarkan penelitian yang telah dilakukan adalah sebagai berikut:

1. Jumlah artikel sebagai data latih (*training*) yang digunakan perlu ditambahkan agar mendapatkan hasil akurasi yang lebih maksimal dalam pengujian klasifikasi.
2. Format *file* untuk *input* artikel dapat dilakukan selain dengan format TXT misalkan, PDF atau DOC/ DOCX.
3. Penelitian ini perlu adanya pengembangan pada proses tokenisasi, agar proses tokenisasi dapat memperhatikan frasa (gabungan dua kata atau lebih yang bersifat nonpredikatif) agar makna kata lebih sesuai.

DAFTAR PUSTAKA

Ariadi, D. dan Fithriasari, K. 2015. Klasifikasi Berita Indonesia Menggunakan Metode *Naive Bayesian Classification* dan *Support Vector Machine* dengan *Confix Stripping Stemme*r. *JURNAL SAINS DAN SENI ITS Vol. 4, No.2*, 4(2):248–253.

Asiyah, S.N. dan Fithriasari, K. 2016. Klasifikasi Berita *Online* Menggunakan Metode Support Vector Machine dan K- Nearest Neighbor. *Jurnal SAINS dan SENI ITS*, 5(2): 317–322.

Bengio, y. 2004. No unbiased estimator of the variance of *k-fold cross-validation. journal of machine learning research* 5 (2004) 1089–1105.

Claudy, Y.I., Perdana, R.S. dan Fauzi, M.A. 2018. Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(8): 2761–2765.

Even, Yair dan Zohar. 2002 *Introduction to Text Mining*.National Center for Supercomputing Applications University of Illinois.

Fildman, R. dan Sanger, J. 2007. *The Text Mining Handbook*. New York : Cambridge University Press.

Hamzah, A. 2012. Klasifikasi Teks dengan *Naive Bayes Classifier* (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III Yogyakarta*, (2011): B-269-B-277.

Han, J. dan Kamber, M., 2006. Data *Mining Concept and Tehniques. Morgan Kauffman, San Fransisco.*

Hotho, A., Andreas, N., Paaß, G dan Augustin, S. 2005. A Brief Survey of Text Mining. 1–37.

Imandoust, S.B. dan Bolandraftar, M. 2013. *Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background*. *International Journal of Engineering Research and Applications*, 3(5): 605–610.

Kumar, L., dan Bhatia, P.K. 2013. TEXT *MINING* : CONCEPTS , PROCESS AND APPLICATIONS. *Journal of global Research in Computer Science*, 4(3):36–39.

Kurniawan. B., dan Effendi, O.S.S. 2012. Klasifikasi Konten Berita Dengan Metode Text *Mining*. *Jurnal Dunia Teknologi Informasi*, 1(2): 14–19.

Khamis, H.S., Cheruiyot, K.W, dan Kimani, S. 2014. *International Journal of Information and Communication Technology Research Application of K-Nearest Neighbour Classification in Medica*l Data *Mining*. *International Journal of Information and Communication Technology Research*, 4(4): 8. Tersedia di http://www.esjournals.org.

Latif, S. 2018. *Text Mining Untuk Klasifikasi Konten Abstrak Jurnal Bahasa Inggris Menggunakan metode Reduksi Dimensi dan Naive Bayes*. Universitas Hasannuddin.

Lidya, S. K., Sitompul, O. S., dan Efektif, S. 2015. Sentiment Analysis pada Text Bahasa Indonesia Menggunakan (SVM) dan (KNN). Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015). ISSN, 2089-9815.

Nasyuha, W.F., Husaini., dan Mursyidah 2016. Klasifikasi Dokumen Tugas Akhir Menggunakan Algoritma *K-Means.* Jurnal Infomedia. ISSN 2527-9858

Nugroho, A.S., Witarto, A. B., dan Handoko, D. 2003. Support Vector Macgine: Teori dan Aplikasinya dalam Bio Informatika. Kuliah Umum Ilmu Komputer . Com. Indonesia.

Pramesti, R.P.A. 2013. *Identifikasi Karakter Plat Nomor Kendaraan Menggunakan Ekstraksi Fitur ICZ dan ZCZ Dengan Metodeklasifikasi K-NN*. *Institut Pertanian Bogor*. Institut Pertanian Bogor.

Prasetyo, H. 2014. Data *Mining* Mengolah Data Menjadi Informasi. Andi Offset. Yogyakarta.

Seifemichael B. Amsalu, Afghah, F., Ramyar, S., dan Kur,. A. 2015 , Driver Bihavior Modeling near Interdrction Using Support Vector Machine based on Statistical Feature Extranction , in Proc. IEEE (IV) Symp.

Somantri, O., Wiyono, S., dan Apriliani, D 2018, *Support Vector Machine* Berbasis Feature Selection untuk Sentiment Analysus Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal, Jurusan Teknik Informatika. Politeknik Harapan Bersama. Tegal.

Suguna, N., dan Thanushkodi, K. 2010. An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. *IJCSI International Journal of Computer Science*, 7(4): 7–10.

Sukhjit, P., Sehra, S., dan Nayyar, P.A. 2013. A Review Paper on Algorithms Used For Text Classification. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 2(3): 90–99.

Tala, F. Z. n. d. A. 2001. Study of *Steamming Effect*s on I*nformation Retrieval in* Bahasa Indonesia. Tesis. U*niversiteit van Amsterdam*. *Netherland.*

Tan, A. 1997. Cascade ARTMAP: integrating neural computation and symbolic knowledge processing. *IEEE Transactions on Neural Networks*, 8(2): 237–250.

Ting, S.L., Ip, W.H., dan Tsang, A.H.C. 2011. Is Naï ve Bayes a Good Classifier for Document Classification ? *International Journal of Software Engineering and Its Applica*, 5(3): 37–46.

Walega, P.A. 2014. Overfitting Problem In A Virtual Sensor Obtained With W-M Method. Warsawa. : Institue of philosophy.

Widaningsi, A.S. 2018. Klasifikasi Jurnal Ilmu Komputer Berdasarkan Pembagian. *Seminar Nasional Teknologi Informasi dan Komunikasi 2018 (SENTIKA 2018)*, 2018(Sentika): 23–24.

Yunliang, J., Qing, S., Jing, F., dan Xiongtao, Z. 2010,. The Classification for E-government Document Based on SVM. *In Web Information Systems and Mining (WISM), 2010 International Conference on* (Vol. 2, pp. 257-260).

Zhang, G. P. 2008. Data *Mining* dan Knowledge Discovery Handbook. Springer. Israel.