KLASIFIKASI ANALISIS SENTIMEN *MOVIE REVIEW* DENGAN METODE *SUPPORT VECTOR MACHINE* MENGGUNAKAN KERNEL *RADIAL BASIS FUNCTION* DAN *INFORMATION GAIN*

Skripsi

disusun sebagai salah satu syarat

untuk memperoleh gelar Sarjana Komputer

Program Studi Teknik Informatika

Oleh

Wahyu Destian Wicaksono
4611416045

JURUSAN ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS NEGERI SEMARANG

2020

PERNYATAAN

Saya menyatakan bahwa skripsi saya yang berjudul "Klasifikasi Analisis Sentimen *Movie Review* dengan Metode *Support Vector Machine* Menggunakan Kernel *Radial Basis Function* dan *Information Gain*" disusun atas dasar penelitian saya dengan arahan dosen pembimbing. Sumber informasi atau kutipan yang berasal dari karya yang diterbitkan telah disebutkan dalam teks dan dicantumkan dalam daftar pustaka di bagian akhir skripsi ini. Dan saya menyatakan bahwa skripsi ini bebas plagiat dan apabila di kemudian hari terbukti terdapat plagiat dalam skripsi ini, maka saya bersedia menerima sanksi sesuai ketentuan perundang-undangan.

Semarang, 19 Maret 2020

Wahyu Destian Wicaksono
4611416045

ii

PERSETUJUAN PEMBIMBING

Nama	: Wahyu Destian Wicaksono
NIM	: 4611416045
Program Studi	: Teknik Informatika S1
Judul Skripsi	: Klasifikasi Analisis Sentimen *Movie Review* dengan Metode *Support Vector Machine* Menggunakan Kernel *Radial Basis Function* dan *Information Gain*

Skripsi ini telah disetujui oleh pembimbing untuk diajukan ke sidang panitia ujian skripsi Program Studi Teknik Informatika FMIPA UNNES.

Semarang, 19 Maret 2020

Pembimbing

Zaenal Abidin S.Si., M.Cs., Ph.D.
NIP 198205042005011001

iii

PENGESAHAN

Skripsi yang berjudul

Klasifikasi Analisis Sentimen *Movie Review* dengan Metode *Support Vector Machine* Menggunakan Kernel *Radial Basis Function* dan *Information Gain*

disusun oleh

Wahyu Destian Wicaksono

4611416045

Telah dipertahankan di hadapan sidang Panitia Ujian Skripsi FMIPA UNNES pada tanggal 26 Maret 2020.

Panitia:

Ketua

Dr. Sugianto, M.Si.
NIP 196102191993031001

Sekretaris

Dr. Alamsyah, S.Si., M.Kom.
NIP 197405172006041001

Penguji 1

Aji Purwinarko S.Si., M.Cs.
NIP 198509102015041001

Penguji 2

Endang Sugiharti, S.Si., M.Kom.
NIP 197401071999032001

Anggota Penguji/
Pembimbing

Zaenal Abidin S.Si., M.Cs., Ph.D.
NIP 198205042005011001

iv

MOTTO DAN PERSEMBAHAN

MOTTO

- Libatkan Allah dalam segala urusanmu. *Keep the faith, Atsiqoh Billah*.
- Tidak akan kecewa seseorang yang berserah diri kepada Allah. Karena sesungguhnya siapapun yang berserah diri kepada Allah dan dia senantiasa berbuat kebaikan, maka dia telah berpegang pada tali yang paling kokoh, dan kepada Allah segala urusan dikembalikan (QS. Luqman: 22).

PERSEMBAHAN

Skripsi ini saya persembahkan kepada:

- Kedua Orang Tua saya Bapak Slamet dan Ibu Srirustinah yang saya sayangi, kasihi, cintai dan selalu saya doakan, semoga ini bisa menghadirkan senyum di bibir dan hati kalian.
- Kakak saya, Ari Sebtiana Rahmawati yang selalu memberikan dukungan, doa, semangat, serta motivasi.
- Ilham Esa Tiffani yang selalu menjadi teman diskusi dan banyak memberikan dukungan serta motivasi
- Teman-teman saya di jurusan Ilmu Komputer, Fakultas MIPA, serta teman-teman di Universitas Negeri Semarang.
- Semua pihak yang tidak dapat disebutkan satu persatu yang telah membantu hingga terselesaikannya penulisan skripsi ini.
- Almamater, Universitas Negeri Semarang.

PRAKATA

Puji syukur penulis panjatkan kepada *Rabb* seluruh alam semesta Allah *Subhanahu wa ta'ala*, atas berkat rahmat, pertolongan, petunjuk, bimbingan serta kebaikan-Nya, penulis dapat menyelesaikan skripsi yang berjudul "**Klasifikasi Analisis Sentimen *Movie Review* dengan Metode *Support Vector Machine* Menggunakan Kernel *Radial Basis Function* dan *Information Gain***". Shalawat serta salam tak lupa selalu terpanjatkan kepada junjungan kita makhluk paling mulia di langit dan di bumi, suri tauladan bagi kita semua, manusia yang membawa peradaban kepada zaman yang terang-benderang, *Rasulullah* Muhammad *Shallallahu alaihi wassalam*, keluarganya, sahabatnya dan para pengikutnya.

Penulis menyadari bahwa penulisan skripsi ini tidak akan selesai tanpa adanya dukungan serta bantuan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih kepada:

1. Prof. Dr. Fathur Rokhman, M.Hum., Rektor Universitas Negeri Semarang.
2. Dr. Sugianto, M.Si., Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang.
3. Dr. Alamsyah, S.Si., M.Kom., Ketua Jurusan Ilmu Komputer FMIPA Universitas Negeri Semarang yang telah meluangkan waktu, membantu, membimbing, mengarahkan dan memberikan saran sehingga penulis dapat menyelesaikan skripsi ini.
4. Zaenal Abidin S.Si., M.Cs., Ph.D., Dosen Pembimbing saya, guru saya yang dengan penuh kesabaran telah meluangkan waktu, membantu, membimbing,

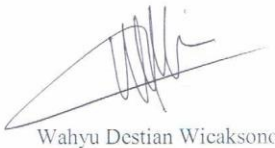
mengarahkan, memberikan saran, dan membagi ilmunya sehingga penulis dapat menyelesaikan skripsi ini.

5. Bapak dan Ibu Dosen Jurusan Ilmu Komputer yang telah memberikan bekal kepada penulis dalam penyusunan skripsi ini.
6. Kedua Orang Tua saya Bapak Slamet dan Ibu Srirustinah yang selalu tanpa lelah mencurahkan kasih sayang, doa, dan dukungannnya.
7. Kakak saya, Ari Sebtiana Rahmawati yang selalu memberikan dukungan, doa, semangat, serta motivasi.
8. Teman-teman saya di jurusan Ilmu Komputer, terutama teman-teman ilkom angkatan 2016, yang telah memberikan semangat dan dukungannya.
9. Teman-teman organisasi Hima Ilkom 2016 yang telah memberikan pengalaman selama kuliah.
10. Semua pihak yang telah membantu terselesaikannya skripsi ini yang tidak dapat penulis sebutkan satu persatu, terimakasih atas bantuannya.

Semoga skripsi ini dapat memberikan manfaat bagi pembaca di masa yang akan datang.

Semarang, 19 Maret 2020

Penulis



Wahyu Destian Wicaksono

NIM 4611416045

vii

ABSTRAK

Wahyu Destian Wicaksono. 2020. Klasifikasi Analisis Sentimen *Movie Review* dengan Metode *Support Vector Machine* Menggunakan Kernel *Radial Basis Function* dan *Information Gain*. Skripsi, Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang. Pembimbing Zaenal Abidin S.Si., M.Cs., Ph.D.

Kata kunci: Analisis sentimen, klasifikasi, *support vector machine*, kernel RBF, *information gain*

Persoalan klasifikasi analisis sentimen masih menjadi pembahasan hangat dan perhatian bagi para peneliti. Analisis sentimen adalah proses yang bertujuan untuk menentukan isi dari *dataset* yang berbentuk teks apakah mengandung sentimen positif atau negatif. Saat ini *review* suatu produk menjadi sumber data yang penting bagi produsen, maupun bagi calon konsumen. Pada penelitian ini akan dilakukan klasifikasi analisis sentimen pada *Data Movie Review Polarity Dataset V2.0* yang berasal dari situs IMDb. *Support Vector Machine* (SVM) telah diusulkan oleh beberapa peneliti untuk digunakan pada klasifikasi analisis sentimen *movie review*, karena memiliki kelebihan dalam mengolah *dataset* yang berjumlah besar dan tetap bekerja dengan baik pada banyak fitur. Namun, kinerja SVM sangat tergantung pada pemilihan kernel dan parameternya yang dapat mempengaruhi hasil akurasi pada klasifikasi. kernel *Radial Basis Function* (RBF) digunakan agar lebih baik dalam memporses data *nonlinear* dan dilakukan *tuning* pada parameternya untuk mendapat hasil akurasi yang optimal. Dalam klasifikasi analisis sentimen terdapat satu kendala lain di mana jumlah fitur yang digunakan sangat banyak, ini dapat mengurangi performa dari klasifikasi. *Feature selection* dapat digunakan untuk mengurangi jumlah fitur yang terlalu besar dengan membuang fitur yang kurang relevan dan memilih fitur dengan korelasi yang kuat terhadap klasifikasi. *feature selection* yang digunakan adalah *information gain* (IG). Pengujian dilakukan dengan implementasi menggunakan bahasa pemrograman *Python.* Hasil akurasi yang diperoleh dalam klasifikasi analisis sentimen *movie review* menggunakan algoritma SVM ialah sebesar 81,50%. Sedangkan, hasil akurasi algoritma SVM dengan menerapkan kernel RBF ialah sebesar 82,25%. Hasil akhir akurasi algoritma SVM dengan menerapkan kernel RBF dan *feature selection* IG sebesar 87,25%. Dengan demikian, penerapan kernel RBF dan *feature selection* IG pada algoritma SVM dapat meningkatkan hasil akurasi dalam klasifikasi analisis sentimen *movie review* sebesar 5,75%. Hasil ini membuktikan bahwa penerapan kernel RBF dan *feature selection* IG berhasil memperoleh tingkat akurasi lebih baik pada klasifikasi analisis sentimen *movie review*.

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR LAMPIRAN

Lampiran Halaman

BAB 1

PENDAHULUAN

1.1. Latar Belakang

Dalam beberapa tahun terakhir, masalah klasifikasi analisis sentimen telah menarik perhatian bagi para peneliti. Persoalan analisis sentimen ini juga menjadi perhatian bagi para pelaku usaha dan *entrepreneur* dalam memasarkan produk ataupun jasa mereka. Melalui *review*, *rating*, iklan, berita, dan komentar yang ada di berbagai media elektronik ini, para pelaku usaha dapat mengetahui seberapa besar respon positif yang diberikan publik terhadap layanan atau jasa yang mereka tawarkan (Kalaivani & Shunmuganathan, 2013).

Analisis sentimen juga disebut sebagai *opinion mining* yang di mana memiliki arti membangun sistem secara komputasional dengan mengumpulkan pendapat terhadap suatu objek dari berbagai sumber dan menganalisisnya (Medhat, Hassan, & Korashy, 2014). Analisis sentimen merupakan bidang penelitian yang cukup populer, karena dapat memberikan keuntungan untuk berbagai aspek, mulai dari prediksi penjualan (Y. Liu, Huang, An, & Yu, 2007), politik (Park, Lim, Sams, Nam, & Park, 2011), dan pengambilan keputusan para investor (Dergiades, 2012).

Menurut Tripathy, Agrawal, dan Rath (2015), analisis sentimen adalah proses yang bertujuan untuk menentukan isi dari *dataset* berisi opini yang berbentuk teks apakah termasuk dalam kategori opini positif, negatif, atau netral baik itu pada *document level*, *sentence level*, ataupun *aspect level*. Pada *document*

level, tujuan utamanya adalah untuk mengklasifikasikan pendapat ke dalam kelas dokumen positif dan negatif. *sentence level* adalah mengkategorikan emosi yang diungkapkan dalam suatu kalimat. Pada *sentence level*, langkah dasarnya adalah mengenali kalimat sebagai obyek atau subyek. Misalkan kalimatnya adalah subyek, ia akan memutuskan apakah kalimat itu menyatakan pendapat negatif atau positif. Dalam *aspect level* bertujuan untuk mengkategorikan sentimen sehubungan dengan entitas tertentu (Bhavitha, Rodrigues, & Chiplunkar, 2017).

Saat ini, opini konsumen telah menjadi salah satu sumber informasi yang begitu penting terhadap berbagai produk, termasuk juga produk film (Zhu, Wang, Zhu, Tsou, & Ma, 2011). Opini konsumen dapat diperoleh melalui media sosial, *e-commerce*, blog, dan YouTube (C.-L. Liu, Hsaio, Lee, Lu, & Jou, 2011). Industri film sendiri secara global terus mengalami perkembangan, baik dari jumlah film yang dibuat, jumlah penonton, maupun jumlah perputaran uangnya. Data yang didapatkan dari *National Association of Theatre Owners* menunjukkan bahwa ada perkembangan dari tahun 1987 di mana tiket bioskop terjual sebanyak 1,09 miliar tiket, menjadi 1,314 miliar tiket pada tahun 2016 untuk penjualan di wilayah Amerika Serikat dan Kanada saja. Sedangkan untuk pendapatan *box office* di wilayah Amerika Serikat dan Kanada pada tahun 1987 sebesar USD 4,25 miliar menjadi USD 11,372 miliar pada tahun 2016, angka ini tentu saja mengatakan bahwa peningkatan perkembangan dunia film sangat pesat.

Popularitas internet mendorong orang untuk mencari pendapat pengguna dari internet sebelum membeli sebuah produk atau jasa termasuk di dalamnya yaitu film (C.-L. Liu *et al*., 2011). Menurut Koh, Hu, dan Clemons (2010), pendapat

orang-orang dapat mengurangi keraguan calon konsumen terhadap suatu produk tertentu dan membantu konsumen menyimpulkan kualitas dari produk tersebut. Seiring berjalannya waktu, hari ini banyak sekali situs yang menyediakan ulasan tentang suatu produk yang dapat mencerminkan pendapat pengguna, salah satunya adalah situs *Internet Movie Database* (IMDb) (C.-L. Liu *et al*., 2011). IMDb adalah situs yang berhubungan dengan film dan produksi film. Informasi yang diberikan IMDb sangat lengkap, mulai dari aktor/aktris yang bermain di film itu, sinopsis singkat film, *link* untuk *trailer* film, tanggal rilis untuk beberapa negara, dan ulasan dari *user-user* yang lain (Chandani & Wahono, 2015). Ketika seseorang ingin membeli atau menonton suatu film, komentar-komentar orang lain dan *rating* film biasanya mempengaruhi minat beli atau menonton mereka.

Konsep dasar dalam analisis sentimen yaitu *text mining*. *Text mining* adalah proses mengubah data teks menjadi data semi terstruktur (Wang, Li, Song, Wei, & Li, 2011). Data teks yang tidak terstruktur harus diubah menjadi data semi terstruktur sehingga dapat ditemukan pola pada data teks baru. Proses pengubahan data teks menjadi data semi terstruktur merupakan tahap *text pre-procesing*. Dalam *text mining* proses *text pre-processing* melalui beberapa tahapan, seperti *tokenization*, *stopwords filtering*, dan *stemming*. Setelah itu akan dilakukan *indexing*, yang di mana hasilnya berupa *bag-of-word*. Setelah terbentuk *bag-of-word*, langkah berikutnya adalah mentransformasikan *token*/*term* menjadi fitur vektor numerik menggunakan metode *word vectorization*. Dengan begitu proses *text mining* dengan algoritma klasifikasi dapat dilakukan.

Teknik untuk melakukan analisis sentimen pada sebuah *review* produk sendiri telah diusulkan oleh para peneliti, yaitu bisa dengan menggunakan algoritma klasifikasi yang diantaranya adalah *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *K-Nearest Neighbor* (KNN) (Kalaivani & Shunmuganathan, 2013), *Maximum Entropy* (ME) (Shivaprasad & Shetty, 2017), *Artificial Neural Network* (ANN) (Chandani & Wahono, 2015), dan *Multilayer Layer Perceptron* (MLP) (Ahmed *et al*., 2017). Dari semua algoritma klasifikasi yang telah diusulkan, SVM dipilih untuk digunakan pada penelitian kali ini karena memiliki kelebihan dibanding algoritma yang lainnya, yaitu meskipun *dataset* yang akan diproses berjumlah besar hasil akurasinya akan tetap tinggi, tetap bekerja dengan baik pada banyak dimensi fitur (*linear* atau *nonlinear*), dan memiliki perlindungan *overfitting* yang berarti tidak selalu tergantung pada jumlah fitur *dataset* (Bhavitha *et al*., 2017). Ini artinya SVM sangat cocok jika digunakan untuk melakukan klasifikasi analisis sentimen pada data *review* sebuah produk yang biasanya memiliki jumlah *dataset* yang besar dan juga memiliki jumlah fitur dalam bentuk *token*/*term* yang sangat banyak di dalamnya.

Beberapa peneliti juga telah melakukan pengujian pada algoritma SVM melalui komparasi menggunakan beberapa *dataset review* melalui cara membandingkan hasil akurasi yang didapatkan dengan algoritma lainnya. Rana dan Singh (2016) pada penelitiannya melakukan perbandingan algoritma SVM dan NB untuk melakukan analisis sentimen data dokumen *movie review* dengan jumlah 2000 data, kemudian mengklasifikasikannya ke dalam dua kelas yaitu kelas dengan label positif dan negatif, hasilnya SVM berhasil memperoleh hasil akurasi terbaik

dengan 78,75%, dan NB dengan 73,75%. Pada penelitian yang lain Shivaprasad dan Shetty (2017) juga mencoba membandingan algoritma SVM, NB, dan ME untuk melakukan analisis sentimen pada data *review* di situs toko online, dengan cara mengekstraksi data *review*, melakukan pemrosesan *Natural Languange Processing* (NLP), komputasi linguistik, dan klasifikasi. Hasilnya pada kategori produk olahraga, gawai, dan komputer rata-rata algoritma SVM berhasil memperoleh hasil akurasi tertinggi dengan 99,10%.

Meskipun SVM memiliki beberapa kelebihan namun menurut Gomes, Prudêncio, Soares, Rossi, dan Carvalho (2012) kinerja SVM sangat tergantung pada pemilihan kernelnya. Pemilihan fungsi kernel yang tepat dalam SVM adalah hal yang sangat penting, karena fungsi kernel ini akan menentukan *feature space* di mana fungsi *classifier* akan dicari, maka dari itu pada SVM dikenal dengan *kernel trick* (Naufal, Wahono, & Syukur, 2015), ada beberapa kernel yang dapat digunakan untuk proses penyeleksian parameter diantaranya *Radial Basis Function* (RBF) *Kernel*, *Linear Kernel*, dan *Polynomial Kernel*. Pada penelitian kali ini akan digunakan kernel RBF, karena kernel RBF dapat menangani pemisahan *linear* pada data *input nonlinear* berdimensi tinggi seperti dalam klasifikasi teks (Ahmed *et al*., 2017), juga bermanfaat mengurangi sulitnya membaca data numerik, karena nilai kernel berada diantara nol dan satu. selain itu kernel RBF juga menunjukkan *tradeoff* parameter c dalam algoritma SVM yang sangat mempengaruhi hasil dari klasifikasinya (Zhou, Liu, & Ye, 2009). Kernel RBF dalam SVM akan mendapatkan performa klasifikasi dan akurasi yang lebih baik ketika melakukan pemilihan nilai parameter $gamma$ dan konstanta *soft margin* C yang tepat.

Untuk melihat seberapa efektif penerapan kernel RBF pada algoritma SVM Ahmed *et al*. (2017) dalam penelitiannya di bidang *text mining* telah mencoba melakukan identifikasi dan analisis terhadap perbandingan akurasi algoritma SVM dengan *poly kernel*, SVM dengan kernel RBF, NB, dan MLP pada 3001 data *movie review* dalam klasifikasi analisis sentimen. Hasilnya menunjukkan bahwa SVM dengan kernel RBF memperoleh hasil akurasi terbaik dengan 97,20% unggul dari SVM *poly kernel* 94,80% dan NB serta MLP yang mendapatkan akurasi berturut-turut 91,16% dan 91,60%. Pada penelitian lain, Jadav dan Vaghela (2016) mencoba melakukan optimasi algoritma SVM dengan menggunakanan kernel RBF dan memodifikasi parameter C serta *gamma*-nya dengan tujuan melihat seberapa baik hasil akurasi klasifikasi analisis sentimennya jika dibanding dengan SVM saja dan NB dengan menggunakan *dataset polarity movie review dataset*, *twitter dataset*, dan *gold dataset* dari *amazon.com*, hasil yang diperoleh SVM dengan kernel RBF memperoleh rata-rata akurasi terbaik pada percobaan di semua *dataset* yaitu 75,5%, SVM 74,80%, dan NB 73,40%.

Masalah lain yang ada pada klasifikasi analisis sentimen adalah banyaknya fitur yang digunakan pada sebuah *dataset* (Wang *et al*., 2011). Pada umumnya klasifikasi analisis sentimen merupakan proses mengolah data berupa teks dan ini menghasilkan fitur dengan jumlah yang sangat banyak, jika semua fitur tersebut digunakan dalam proses klasifikasi maka akan mengurangi kinerja komputasi yang bisa mengakibatkan menurunnya akurasi klasifikasi (Wang, Li, Zhao, & Zhang, 2013). Menurut Wang *et al*. (2011) *feature selection* memiliki konsep mengurangi ruang fitur yang besar, seperti membuang fitur yang kurang relevan dalam

klasifikasi untuk mereduksi jumlah fitur dari *dataset*. *Feature selection* dalam penelitian ini menggunakan *information gain* (IG), karena dengan menggunakan IG dapat diketahui bobot dari suatu fitur dan bisa dipilih fitur terbaik berdasarkan rangking data (*top k*) (Aggarwal & Philip, 2008).

Pada penelitian yang dilakukan oleh Chandani dan Wahono (2015), melakukan perbandingan pada algoritma SVM, NB, ANN untuk melakukan klasifikasi analisis sentimen pada *dataset movie review* yang dikombinasikan dengan *feature selection* untuk mengatasi masalah jumlah fitur yang terlalu besar, metode *feature selection* yang diterapkan yaitu IG, *chi-square, forward selection*, dan *backward selection*. Hasilnya menunjukkan bahwa kombinasi menggunakan *feature selection* IG mendapatkan hasil terbaik dengan rata-rata peningkatan akurasinya 15,46%. Kemudian Kalaivani dan Shunmuganathan (2013) membandingkan tiga algoritma klasifikasi SVM, NB, dan KNN untuk mencari algoritma klasifikasi dengan hasil akurasi terbaik melalui penerapan *feature selection* IG pada *dataset movie review* dalam klasifikasi analisis sentimen, hasil penelitian ini menunjukkan bahwa kombinasi dari SVM dan IG sukses memperoleh akurasi terbaik dengan 81.45%. dari hasil penelitian tersebut terbukti bahwa *feature selection* IG efektif mengatasi jumlah fitur yang terlalu besar dalam sebuah *dataset*.

Berdasarkan uraian permasalahan di atas, maka diusulkan penelitian untuk mengukur tingkat akurasi algoritma SVM dengan kernel RBF dan menerapkan IG sebagai *feature selection* untuk klasifikasi analisis sentimen *dataset Data Movie Review Polarity Dataset V2.0*. Hal inilah yang menjadi latar belakang penulis dalam melakukan penelitian yang berjudul “Klasifikasi Analisis Sentimen *Movie Review*

dengan Metode *Support Vector Machine* Menggunakan Kernel *Radial Basis Function* dan *Information Gain*".

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas, maka rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana penerapan kernel RBF dan *feature selection* IG pada algoritma SVM dalam klasifikasi analisis sentimen *movie review*?
2. Bagaimana akurasi dari algoritma SVM dengan menerapkan kernel RBF dan *feature selection* IG dalam klasifikasi analisis sentimen *movie review*?

1.3. Batasan Masalah

Pada penelitian ini diperlukan batasan-batasan agar tujuan penelitian dapat tercapai. Adapun batasan masalah yang dibahas dalam penelitian ini adalah sebagai berikut.

1. Algoritma klasifikasi yang digunakan dalam penelitian ini adalah algoritma *Support Vector Machine* (SVM).
2. Kernel yang digunakan pada algoritma SVM adalah kernel *Radial Basis Function* (RBF).
3. *Feature selection* yang digunakan adalah *Information Gain* (IG).
4. Data yang digunakan pada penelitian ini adalah *Data Movie Review Polarity Dataset V2.0* (Pang & Lee, 2004).

5. Bahasa pemrograman yang digunakan dalam pembuatan sistem pada penelitian ini adalah *Python*.

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut.

1. Menerapkan kernel RBF dan *feature selection* IG pada algoritma SVM dalam klasifikasi analisis sentimen *movie review*.
2. Mengetahui akurasi dari algoritma SVM sebelum dan sesudah menerapkan kernel RBF dan IG dalam klasifikasi analisis sentimen *movie review*.

1.5. Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut:

1. Memahami penerapan kernel RBF dan *feature selection* IG pada algoritma SVM dalam klasifikasi analisis sentimen *movie review*.
2. Mengetahui hasil peningkatan akurasi antara algoritma SVM dan algoritma SVM yang menerapkan kernel RBF dan *feature selection* IG dalam klasifikasi analisis sentimen movie *review*.

1.6. Sistematika Penulisan

Sistematika penulisan berguna untuk memudahkan dalam memahami jalan pemikiran secara keseluruhan skripsi. Penulisan skripsi ini secara garis besar dibagi menjadi tiga bagian, yaitu sebagai berikut.

1.6.1 Bagian Awal Skripsi

Bagian awal skripsi terdiri dari halaman judul, halaman pengesahan, halaman pernyataan, halaman motto dan persembahan, abstrak, kata pengantar, daftar isi, daftar gambar, daftar tabel dan daftar lampiran.

1.6.2 Bagian Isi Skripsi

Bagian isi skripsi terdiri dari lima bab, yaitu sebagai berikut.

1. BAB 1: PENDAHULUAN

   Bab ini berisi tentang latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian serta sistematika penulisan skripsi.

2. BAB 2: TINJAUAN PUSTAKA

   Bab ini berisi penjelasan mengenai definisi maupun pemikiran-pemikiran yang dijadikan kerangka teoritis yang menyangkut dan mendasari pemecahan masalah dalam skripsi ini.

3. BAB 3: METODE PENELITIAN

   Bab ini berisi penjelasan mengenai studi pendahuluan, tahap pengumpulan data, tahap analisis data, model yang digunakan dalam penelitian dan analisis kebutuhan serta perancangan sistem.

4. BAB 4: HASIL DAN PEMBAHASAN

   Bab ini berisi hasil penelitian berserta pembahasannya.

5. BAB 5: PENUTUP

   Bab ini berisi simpulan dari penulisan skripsi dan saran yang diberikan penulis untuk mengembangkan skripsi ini.

1.6.3 Bagian Akhir Skripsi

Bagian akhir skripsi ini berisi daftar pustaka yang merupakan informasi mengenai buku-buku, sumber-sumber dan referensi yang digunakan penulis serta lampiran-lampiran yang mendukung dalam penulisan skripsi ini.

BAB 2

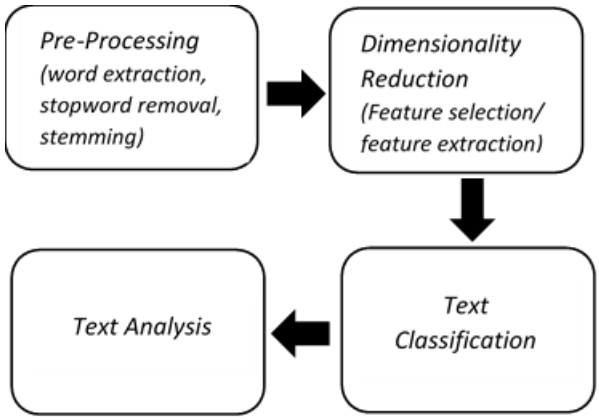
TINJAUAN PUSTAKA

2.1. *Text Mining*

Text mining adalah proses menambang data yang berupa teks di mana data tersebut bersumber salah satunya dari dokumen. Tujuan dari *text mining* adalah mencari kata tertentu yang memiliki makna khusus guna mewakili isi dari sebuah dokumen, sehingga dapat dilakukan analisis keterhubungan antardokumen. *Text mining* bisa juga diartikan sebagai suatu proses ekstraksi pola unik berupa informasi dan pengetahuan yang berguna dari sejumlah sumber data yang berupa teks (Kotu & Deshpande, 2014).

Teknik *text mining* memungkinkan pengguna untuk mengekstrak informasi khusus dari sejumlah besar informasi dan untuk mengidentifikasi hubungan dengan informasi lainnya, di mana ini juga melibatkan kategorisasi teks (Hong, Lee, & Han, 2015). Untuk membuat komputer melakakukan analisis mendalam terhadap informasi yang ditulis dalam bahasa manusia dan guna untuk menemukan informasi khusus dari informasi yang diberikan, maka digunakan sumber daya linguistik dan algoritma pembelajaran pola statistik (Hong *et al*., 2015). Ketika data menjadi sebuah *big data*, dan sumber data seperti media sosial, YouTube, serta blog semakin besar, *text mining* banyak digunakan untuk iklan, pemasaran, analisis kasus hukum, pencarian informasi, dan analisis tren.

Klasifikasi teks dilakukan menggunakan algoritma klasifikasi data yang biasanya digunakan dalam *data mining*. Klasifikasi data menggunakan kriteria klasifikasi yang telah ditentukan untuk mempelajari data melalui *data training* dan kemudian menggunakan hasil tersebut yang biasa disebut sebagai model untuk mengklasifikasikan *data testing* ke dalam kelas yang telah ditentukan. Dengan demikian, dimungkinkan untuk menunjukkan karakteristik data yang diberikan dan untuk mengekstrak data guna mencari informasi khusus di dalamnya. Artinya, kita dimungkinkan untuk membuat prediksi pada data baru berdasarkan pada *data training*, proses tahapan *text mining* ditunjukkan pada Gambar 2.1 (Shamsinejadbabki & Saraee, 2012).



Gambar 2.1 Kerangka proses *text mining*

Proses dasar *text mining* adalah sebagai berikut, tahap yang paling dasar yaitu sekumpulan dokumen teks yang dilakukan *text pre-processing*, adapun teknik yang digunakan antara lain *tokenize, lower case, stopword removal* dan *stemming*. Selanjutnya untuk mereduksi jumlah fitur dalam *dataset*, digunakan teknik *feature extraction*/*feature selection* (Wang *et al*., 2011). Kemudian tahap klasifikasi teks

berdasarkan kelas menggunakan algoritma klasifikasi, terakhir dilakukan analisis dari hasil klasifikasi teks atau dokumen.

Menurut Miner *et al*. (2012), pekerjaan yang dapat dilakukan dengan memanfaatkan *text mining* dapat dikelompokkan menjadi 7 macam sebagai berikut.

1. Pencarian dan perolehan informasi (*search and information retrieval*), yaitu penyimpanan dan pencarian dokumen teks, misalnya dalam mesin pencari (*search engine*) dan pencarian kata kunci (*keywords*).
2. Pengelompokan dokumen, dilakukan pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunakan metode *clustering* pada *data mining*.
3. Klasifikasi dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunakan metode *classification* pada *text mining* berdasarkan model terlatih yang telah memiliki label.
4. *Web Mining*, merupakan penggalian informasi dari internet dengan skala fokus tertentu.
5. *Information extraction*, yaitu mengidentifikasi dan mengekstraksi informasi dari data yang sifatnya semi terstruktur atau tidak terstruktur dan mengubahnya menjadi data yang terstruktur.
6. *Natural language processing* (NLP), yaitu pembuatan program yang memiliki kemampuan untuk memahami bahasi manusia.
7. Ekstraksi konsep, merupakan pengelompokan kata atau frase ke dalam kelompok yang mirip secara konsep semantik.

2.2. Analisis Sentimen

Analisis sentimen atau *opinion mining* merupakan proses memahami, mengekstrak, dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Medhat *et al*. (2014) menjelaskan bahwa analisis sentimen atau *opinion mining* adalah proses secara komputasional terhadap sebuah masalah atau objek. Entitas tersebut direpresentasikan secara tunggal, majemuk, maupun berupa topik. Analisis sentimen mengidentifikasi sentimen yang terkandung dalam teks kemudian menganalisisnya apakah cenderung beropini positif atau negatif.

Analisis sentimen merupakan salah satu cabang penelitian dalam *text mining* yaitu *text classification* yang bertujuan untuk menentukan isi dari *dataset* yang berbentuk teks baik itu berupa dokumen, kalimat, maupun paragraf yang bersifat positif, negatif, atau netral. C.-L. Liu *et al*. (2011) menjelaskan bahwa analisis sentimen atau *opinion mining* berpedoman pada pembahasan yang lebih dalam dari pengolahan bahasa manusia, komputasi linguistik, dan *text mining* yang bertujuan menganalisis pendapat, sentimen, sikap, penilaian, dan emosi seseorang baik itu terhadap produk berupa barang atau jasa.

Tujuan utama dalam analisis sentimen adalah mengumpulkan opini dalam bentuk teks untuk diidentifikasi sentimen yang terkandung di dalamnya, kemudian mengklasifikasi polaritasnya. Salah satu contoh penggunaan analisis sentimen dalam dunia nyata adalah identifikasi kecenderungan pasar dan opini pasar terhadap suatu objek barang atau jasa. Besarnya pengaruh dan manfaat dari analisis sentimen menyebabkan penelitian dan aplikasi berbasis analisis sentimen berkembang pesat.

Bahkan di Amerika terdapat sekitar 20-30 perusahaan yang memfokuskan pada layanan analisis sentimen (B. Liu, 2010).

Analisis sentimen biasanya diaplikasikan dalam tiga level berbeda yaitu, *sentence level*, *document level*, dan *aspect level*. Tujuan dari *document level* yaitu mengklasifikasikan seluruh dokumen kedalam kelas positif atau negatif. *Sentence level* berdasarkan pada poalritas masing-masing kalimat secara individu (Tripathy *et al*., 2015). Medhat *et al*. (2014) menjelaskan bahwa tujuan utama dari analisis sentimen *document level* adalah mengklasifikasikan opini sebuah dokumen sebagai opini yang bersentimen positif atau negatif berdasarkan beberapa dokumen berukuran besar dengan satu topik yang sama. *Sentence level* pada analisis sentimen, mengklasifikasikan sentimen pada masing-masing kalimat dengan mengindentifikasi kalimat tersebut apakah termasuk opini positif atau negatif.

Penelitian di bidang analisis sentimen atau *opinion mining* mulai marak pada tahun 2002. Turney (2002) melakukan penelitian bertema *opinion mining* dengan menggunakan data berupa data ulasan konsumen suatu produk. Metode yang digunakan adalah *semantic orientation* menggunakan *pointwise mutual information* (SO-PMI). Hasil akurasi terbaik yang dicapai adalah 84% terhadap data ulasan kendaraan bermotor dan 66% untuk data *movie review*. Pang dan Lee (2004) mengklasifikasikan ulasan dari film pada level dokumen yang memiliki pendapat positif atau negatif dengan menggunakan teknik *supervised learning*. Sekumpulan ulasan dari film yang sebelumnya telah diberikan label positif atau negatif digunakan sebagai *data training* untuk beberapa algoritma *machine learning* yang sudah ada, hasilnya akurasi yang didapatkan berkisar antara 72% sampai 83%.

2.3. *Dataset Movie Review*

Dataset merupakan kumpulan dari objek dan fiturnya. Fitur merupakan sifat atau karakteristik dari suatu objek, contohnya bisa warna mata seseorang, suhu, atau kata (dalam *text minig*). Fitur juga dikenal sebagai variabel, *field*, karakteristik atau atribut. Kumpulan dari fitur menggambarkan sebuah objek bisa disebut *record*, titik, kasus, sampel, entitas atau *instance* (Hermawati, 2013). Sedangkan menurut Jena dan Kamila (2015), *dataset* merupakan suatu kumpulan data atau satu data statistik di mana setiap fitur data dapat mewakili suatu objek dan setiap *instance* memiliki deskripsi sendiri.

Dataset yang akan digunakan pada penelitian bersumber dari *review* suatu film. Perfilman merupakan salah satu industri yang terus berkembang hingga saat ini. Seiring dengan hal tersebut muncul sebuah situs tentang *review* berbagai film. Salah satu contoh situs yang menyediakan *review* tentang produk film adalah *Internet Movie Database* (IMDb). Chandani dan Wahono (2015) menambahkan bahwa agar industri film terus berkembang ke arah yang lebih baik maka dibutuhkan penilaian-penilaian dari para penikmat film. Banyak masyarakat memanfaatkan IMDb untuk mengetahui kulitas sebuah film sebelum membeli atau menonton suatu film, dengan melalui komentar-komentar orang lain dan peringkat film tersebut biasanya mempengaruhi tingkat ketertarikan seseorang untuk membeli ataupun menonton film tersebut.

Pada penelitian kali ini *dataset movie review* yang digunakan adalah *Data Movie Review Polarity Dataset V2.0* (Pang & Lee, 2004). Sumber data awal adalah arsip IMDb kemudian dipilih hanya *review* di mana peringkat penulis dinyatakan

dengan bintang atau nilai numerik (konvensi lain terlalu beragam untuk memungkinkan pemrosesan otomatis). Untuk metode yang dijelaskan dalam penelitian ini hanya berkonsentrasi pada membedakan antara sentimen positif dan negatif. *Dataset* ini tersedia secara *online* di http://www.cs.cornell.edu/people/pabo/- movie-review-data, yang terdiri dari 2000 data *review* dengan 1000 *review* positif dan 1000 negatif.

2.4. *Term Frequency – Invers Document Frequency (TF-IDF)*

Metode TF-IDF adalah metode yang digunakan untuk merubah kata dalam *dataset* menjadi vektor numerik melalui teknik pembobotan yang digunakan pada *information retrieval.* Metode ini juga terkenal efisien dan mudah digunakan. Metode ini akan menghitung nilai TF dan IDF pada setiap *token*/*term* pada setiap dokumen di dalam kelas label (Jindal, Malhotra, & Jain, 2015).

TF adalah jumlah kemunculan kata pada suatu dokumen. Semakin banyak suatu kata muncul pada dokumen, maka semakin besar kata tersebut berpengaruh pada dokumen tersebut. Sebaliknya, semakin sedikit suatu kata muncul pada dokumen, maka semakin kecit kata tersebut berpengaruh pada dokumen tersebut. IDF adalah pembobotan kata yang didasarkan pada banyaknya dokumen yang mengandung kata tertentu. Semakin banyak dokumen yang mengandung suatu kata tertentu, semakin kecil pengaruh kata tersebut pada dokumen. Sebaliknya, semakin sedikit dokumen yang mengandung suatu kata tertentu, semakin besar pengaruh kata tersebut pada dokumen (Feldman & Sanger, 2007).

2.5. *Feature Selection*

Menurut Chandani dan Wahono (2015) *feature selection* merupakan bagian penting untuk mengoptimalkan kinerja dari *classifier*. menurut Tang, Alelyani, dan Liu (2014), *feature selection* merupakan teknik yang secara umum digunakan untuk mengurangi dimensi dikalangan praktisi. Lebih jelasnya *feature selection* dapat didasarkan pada pengurangan fitur yang besar, misalnya dengan mereduksi fitur yang kurang relevan dan bisa juga memberi nilai bobot pada setiap fitur.

Hal ini bertujuan untuk memilih *subset* kecil dari fitur yang relevan dari yang asli berdasarkan kriteria evaluasi relevansi tertentu, yang biasanya menyebabkan kinerja *learning* menjadai lebih baik, seperti proses komputasi yang ringan dan model *interpretability* yang lebih baik. Jindal *et al*. (2015) menjelaskan bahwa penerapan metode *feature selection* digunakan untuk mengurangi dimensi dari set fitur dengan menghapus fitur yang tidak relevan, karena jumlah fitur hasil *text pre-processing* dalam *bag-of-word* masih sangat banyak dengan tujuan mengurangi dimensi data untuk meningkatkan performa akurasi klasifikasi, dan mengurangi *overfitting*.

Menghilangkan fitur yang relevan atau membiarkan fitur yang kurang relevan dapat menyebabkan kerugian dan menyebabkan bias pada algoritma untuk mendapatkan hasil yang baik. *Feature selection* memiliki sejumlah keunggulan seperti ukuran *dataset* yang lebih kecil, menyusutnya ruang pencarian, dan komputasi yang lebih ringan. Tujuannya adalah pengurangan ukuran dimensi untuk menghasilkan peningkatan akurasi klasifikasi, metode pada *feature selection* dalam klasifikasi dokumen teks menggunakan fungsi evaluasi yang diterapkan per

token/*term*. Pembobotan dari fitur berbentuk *token*/*term* (fitur individu terbaik) dapat dilakukan menggunakan beberapa teknik, salah satunya yaitu IG. Metode pembobotan ini memberikan peringkat dan bobot pada fitur (Azhagusundari & Thanamani, 2013).

2.6. *Information Gain (IG)*

IG sering digunakan untuk merangking fitur yang paling berpengaruh terhadap kelasnya. Nilai *gain* dari suatu fitur, diperoleh dari nilai *entropy* sebelum pemisahan dikurangi dengan nilai *entropy* setelah pemisahan. Tujuan pengurangan fitur pengukuran nilai informasi diterapkan sebagai tahap sebelum pengolahan awal. Menurut Bramer (2007) hanya fitur yang memenuhi kriteria (*top k*) yang dipertahankan untuk digunakan oleh algoritma klasifikasi. Ada tiga tahapan dalam pemilihan fitur menggunakan IG, yaitu sebagai berikut.

1. Hitung nilai *gain* informasi untuk setiap fitur dalam *dataset* asli.
2. Buang semua fitur yang tidak memenuhi kriteria yang ditentukan.
3. *Dataset* direvisi.

Penghitungan fitur ini dipelopori oleh Claude Shannon pada teori informasi (Gallager, 2001) dan dituliskan oleh Han dan Kamber (2011) pada persamaan 1.

$$info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \quad (1)$$

Keterangan:

D: himpunan Kasus

m: jumlah partisi D

p_i: proporsi dari D_i terhadap D

Dalam hal ini p_i adalah probabilitas sebuah *tuple* pada D masuk ke kelas C_i dan diestimasi dengan $|C_i, D|/|D|$. Fungsi log diambil berbasis dua karena informasi dikodekan berbasis bit. Selanjutnya menurut Han dan Kamber (2011) proses mencari nilai *entropy* setelah pemisahan dijelaskan pada persamaan 2.

$$info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} . info(D_j) \quad (2)$$

Keterangan:

D : himpunan kasus

A : fitur

v : jumlah partisi fitur A

$|D_j|$: jumlah kasus pada partisi ke j

$|D|$: jumlah kasus dalam D

$info(D_j)$: total *entopy* dalam partisi

Untuk mencari nilai IG fitur A diperoleh dengan persamaan 3 (Han & Kamber, 2011).

$$Gain(A) = info(D) - Info_A(D) \quad (3)$$

keterangan:

$Gain(A)$ *: information* fitur A

$Info(D)$: total *entopy*

$Info_A(D)$: *entropy A*

Artinya $Gain(A)$ adalah pengurangan yang diharapkan di dalam *entropy* yang disebabkan oleh pengenalan nilai fitur dari A. Fitur yang memiliki nilai IG terbesar dipilih sebagai uji fitur untuk himpunan S. Selanjutnya suatu simpul dibuat

dan diberi label dengan label fitur tersebut, dan cabang-cabang dibuat untuk masing-masing nilai fitur.

2.7. *Support Vector Machine (SVM)*

Menurut Wu dan Kumar (2009) SVM merupakan salah satu dari sepuluh algoritma terbaik dalam *data mining*. Boser, Guyon, dan Vapnik pada tahun 1992 untuk pertama kali mengembangkan dan mempresentasikan teori dari algoritma SVM di *Annual Workshop on Computational Learning Theory*, meskipun dasar untuk SVM sendiri telah ada sejak 1960-an (Suyanto, 2017). Menurut Han dan Kamber (2011) metode SVM menjadi sebuah metode baru yang menjanjikan untuk mengklasifikasi data, baik data *linear* maupun *nonlinear*.

SVM adalah metode yang cepat dan efektif untuk klasifikasi teks (Feldman & Sanger, 2007). Dalam istilah geometris, SVM *classifier* adalah sebuah *hyperplane* pada ruang *feature* yang memisahkan titik yang merepresentasikan *instance* kelas positif dan negatif. Pang dan Lee (2004), menyatakan bahwa algoritma SVM telah terbukti sangat efektif untuk kategorisasi teks tradisional mengalahkan algoritma NB.

Teknik SVM menarik digunakan oleh para peneliti dalam bidang *data mining/text mining* maupun *machine learning* karena performa yang meyakinkan dalam memprediksi kelas suatu data baru. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas pada *input space*. *Pattern* yang merupakan anggota dari dua buah kelas -1 dan +1 dan berbagai alternatif garis pemisah, *pattern* yang

BAB 5

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, maka dapat ditarik kesimpulan sebagai berikut.

1. Penerapan kernel RBF pada penelitian ini digunakan supaya algoritma SVM lebih baik dalam menangani data *nonlinear* pada *dataset movie review*, dan untuk meningkatkan kinerja klasifikasi analisis sentimen melalui modifikasi nilai parameter C dan $gamma$ dengan memilih nilai parameter optimal yang memberikan akurasi lebih baik dalam klasifikasi. Sedangkan penerapan *feature selection* IG pada penelitian ini digunakan untuk memilih jumlah fitur optimal yang digunakan dalam klasifikasi berdasarkan rangking fitur dengan *gain* tertinggi yang membuat akurasi klasifikasi lebih baik.
2. Pada penelitian ini hasil evaluasi akurasi dari klasifikasi analisis sentimen *movie review* menunjukkan bahwa algoritma SVM dengan menerapkan kernel RBF dan *feature selection* IG menghasilkan akurasi klasifikasi yang lebih tinggi mengungguli pendekatan menggunakan algoritma SVM saja. Berdasarkan hasil tersebut, *review* sentimen negatif dan positif pada *dataset movie review* dapat di klasifikasikan dengan akurasi yang lebih baik.

5.2. Saran

Adapun saran dari penelitian ini adalah sebagai berikut.

1. Melakukan penelitian dengan menerapkan selain daripada kernel RBF pada algoritma SVM. Seperti kernel *linear*, *polynomial* atau *sigmoid* untuk melihat kemungkinan bagaimana tingkat kinerja dari klasifikasi analisis sentimen pada *dataset movie review*.
2. Perlu dilakukan penyelidikan lebih lanjut untuk mengetahui tingkat akurasi klasifikasi analisis sentimen dengan melakukan uji komparasi lanjut tidak hanya berdasarkan jumlah fitur yang digunakan tetapi juga melakukan perbandingan jenis fitur yang lain.

DAFTAR PUSTAKA

Aggarwal, C. C., & Philip, S. Y. (Eds.). (2008). *Privacy-preserving data mining: models and algorithms*. New York, NY: Springer Science & Business Media.

Ahmed, E., Sazzad, M. A. U., Islam, M. T., Azad, M., Islam, S., & Ali, M. H. (2017, March). Challenges, comparative analysis and a proposed methodology to predict sentiment from movie reviews using machine learning. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)* (pp. 86-91). IEEE.

Azhagusundari, B., & Thanamani, A. S. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2*(2), 18-21.

Azmi, Z., & Dahria, M. (2013). Decision tree berbasis algoritma untuk pengambilan keputusan. *Jurnal Ilmiah SAINTIKOM, 12*(3), 157-164.

Bhavitha, B. K., Rodrigues, A. P., & Chiplunkar, N. N. (2017, March). Comparative study of machine learning techniques in sentimental analysis. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 216-221). IEEE.

Bramer, M. (2007). *Principles of data mining* (Vol. 180). London: Springer.

Chandani, V., & Wahono, R. S. (2015). Komparasi algoritma klasifikasi Machine Learning dan feature selection pada analisis sentimen review film. *Journal of Intelligent Systems, 1*(1), 56-60.

Dergiades, T. (2012). Do investors' sentiment dynamics affect stock returns? Evidence from the US economy. *Economics Letters, 116*(3), 404-407.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge university press.

Gallager, R. G. (2001). Claude E. Shannon: A retrospective on his life, work, and impact. *IEEE Transactions on Information Theory, 47*(7), 2681-2695.

Gomes, T. A. F., Prudêncio, R. B. C., Soares, C., Rossi, A. L. D., & Carvalho, A. (2012). Combining meta-learning and search techniques to select parameters for support vector machines. *Neurocomputing, 75*(1), 3-13.

Hamel, L. H. (2011). *Knowledge discovery with support vector machines* (Vol. 3). New York, NY: John Wiley & Sons.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. New York, NY: Elsevier.

Hermawati, F. A. (2013). *Data Mining*. Yogyakarta: Andi.

Holovaty, A., & Kaplan-Moss, J. (2009). *The definitive guide to Django: Web development done right*. Berkeley, CA: Apress.

Hong, S.-S., Lee, W., & Han, M.-M. (2015). The feature selection method based on genetic algorithm for efficient of text clustering and text classification. *International Journal of Advances in Soft Computing & Its Applications, 7*(1), 22-40.

Indraswari, R., & Arifin, A. Z. (2017). RBF kernel optimization method with particle swarm optimization on svm using the analysis of input data's movement. *Jurnal Ilmu Komputer dan Informasi, 10*(1), 36-42.

Jadav, B. M., & Vaghela, V. B. (2016). Sentiment analysis using support vector machine based on feature selection and semantic analysis. *International Journal of Computer Applications, 146*(13), 26-30.

Jena, L., & Kamila, N. K. (2015). Distributed data mining classification algorithms for prediction of chronic-kidney-disease. *International Journal of Emerging Research in Management &Technology , 9359*(11), 110-118.

Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for text classification: Literature review and current trends. *Webology, 12*(2), 1-28.

Kalaivani, P., & Shunmuganathan, K. L. (2013). Sentiment classification of movie reviews by supervised machine learning approaches. *Indian Journal of Computer Science and Engineering, 4*(4), 285-292.

Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16.

Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications, 9*(5), 374-385.

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Burlington, MA: Morgan Kaufmann.

Langgeni, D. P., & Baizal, Z. A. (2010, July). Clustering artikel berita berbahasa indonesia menggunakan unsupervised feature selection. In *Seminar Nasional Informatika* (Vol. 1, No. 4). Yogyakarta: "Veteran" University of National Development Yogyakarta.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (pp. 627–666). Boca Raton, FL: CRC press.

Liu, C.-L., Hsaio, W.-H., Lee, C.-H., Lu, G.-C., & Jou, E. (2011). Movie rating and review summarization in mobile environment. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(3), 397-407.

Liu, Y., Huang, X., An, A., & Yu, X. (2007, July). ARSA: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 607-614). New York, NY: ACM Press.

Loper, E., & Bird, S. (2002, July). NLTK: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* (pp. 69–72). Association for Computational Linguistics.

Lutz, M. (2013). *Learning python: Powerful object-oriented programming*. Sebastopol, CA: O' Reilly Media.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing, 14*(9), 1-9.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093-1113.

Miner, G., Elder Iv, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Amsterdam: Academic Press.

Naufal, A. R., Wahono, R. S., & Syukur, A. (2015). Penerapan bootstrapping untuk ketidakseimbangan kelas dan weighted information gain untuk feature selection pada algoritma support vector machine untuk prediksi loyalitas pelanggan. *Journal of Intelligent Systems, 1*(2), 98-108.

Nugroho, A. S. (2008). Support vector machine: paradigma baru dalam softcomputing. *Neural Networks*, 92-99.

Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1, p. 85). USA: Trelgol Publishing.

Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (p. 271). Association for Computational Linguistics.

Park, S. J., Lim, Y. S., Sams, S., Nam, S. M., & Park, H. W. (2011). Networked politics on Cyworld: The text and sentiment of Korean political profiles. *Social Science Computer Review, 29*(3), 288-299.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830.

Pawlik, A., Segal, J., Sharp, H., & Petre, M. (2014). Crowdsourcing scientific software documentation: a case study of the NumPy documentation project. *Computing in Science & Engineering, 17*(1), 28-36.

Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi menggunakan MATLAB.* Yogyakarta: Andi.

Rana, S., & Singh, A. (2016, October). Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)* (pp. 106-111). IEEE.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Amsterdam, Netherlands: Centrum voor Wiskunde en Informatica.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge university press.

Shamsinejadbabki, P., & Saraee, M. (2012). A new unsupervised feature selection method for text clustering based on genetic algorithms. *Journal of Intelligent Information Systems, 38*(3), 669-684.

Shivaprasad, T. K., & Shetty, J. (2017, March). Sentiment analysis of product reviews: a review. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 298-301). IEEE.

Solanki, A. V. (2014). Data mining techniques using weka classification for sickle cell disease. *International Journal of Computer Science and Information Technologies, 5*(4), 5857-5860.

Somantri, O., Wiyono, S., & Dairoh, D. (2016). Metode k-means untuk optimasi klasifikasi tema tugas akhir mahasiswa menggunakan support vector machine (SVM). *Scientific Journal of Informatics, 3*(1), 34-45.

Sugiyono, S. (2012). *Business research methods*. Bandung, Indonesia: Alfabeta.

Suyanto, D. (2017). *Data Mining untuk klasifikasi dan klasterisasi data*. Bandung: Informatika Bandung.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37. Boca Raton, FL: CRC Press.

Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of sentimental reviews using machine learning techniques. *Procedia Computer Science, 57*, 821-829.

Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417-424). Association for Computational Linguistics.

Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications, 38*(7), 8696-8702.

Wang, S., Li, D., Zhao, L., & Zhang, J. (2013). Sample cutting method for imbalanced text sentiment classification based on BRC. *Knowledge-Based Systems, 37*, 451-461.

Wu, X., & Kumar, V. (2009). *The top ten algorithms in data mining*. Boca Raton, FL: CRC press.

Zhou, S.-S., Liu, H.-W., & Ye, F. (2009). Variant of gaussian kernel and parameter setting method for nonlinear SVM. *Neurocomputing, 72*(13-15), 2931-2937.

Zhu, J., Wang, H., Zhu, M., Tsou, B. K., & Ma, M. (2011). Aspect-based opinion polling from customer reviews. *IEEE Transactions on Affective Computing, 2*(1), 37-49.