PENGARUH *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE), REPRESENTASI FITUR, DAN ALGORITMA KLASIFIKASI PADA *SENTIMENT ANALYSIS*

SKRIPSI

Disusun Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Komputer
pada Departemen Ilmu Komputer/Informatika

Disusun Oleh:
Widi Satriaji
24010314130101

DEPARTEMEN ILMU KOMPUTER/INFORMATIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO
2018

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini:

Nama: Widi Satriaji

NIM: 24010314130101

Judul: Pengaruh *Synthetic Minority Oversampling Technique* (SMOTE), Representasi Fitur, dan Algoritma Klasifikasi pada *Sentiment Analysis*

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka.

Semarang, 24 Mei 2018

Widi Satriaji

24010314130101

HALAMAN PENGESAHAN

Judul: Pengaruh *Synthetic Minority Oversampling Technique* (SMOTE), Representasi Fitur, dan Algoritma Klasifikasi pada *Sentiment Analysis*

Nama: Widi Satriaji

NIM: 24010314130101

Telah diujikan pada sidang skripsi pada tanggal 7 Mei 2018 dan dinyatakan lulus pada tanggal 7 Mei 2018.

Semarang, 24 Mei 2018

Mengetahui,
Ketua Departemen Ilmu Komputer/ Informatika
FSM UNDIP

Dr. Retno Kusumaningrum, S.Si, M.Kom
NIP. 198104202005012001

Panitia Penguji Skripsi
Ketua,

Helmie Arif Wibawa, S.Si, M.Cs
NIP. 197805162003121001

HALAMAN PENGESAHAN

Judul: Pengaruh *Synthetic Minority Oversampling Technique* (SMOTE), Representasi Fitur, dan Algoritma Klasifikasi pada *Sentiment Analysis*

Nama: Widi Satriaji

NIM: 24010314130101

Telah diujikan pada sidang skripsi pada tanggal 7 Mei 2018.

Semarang, 24 Mei 2018

Pembimbing

Dr. Retno Kusumaningrum, S.Si, M.Kom.

NIP. 198104202005012001

ABSTRAK

Komentar-komentar pada layanan sewa hotel *online* seperti Traveloka merupakan sumber daya sangat penting yang bisa digunakan bagi pihak penyedia layanan tersebut termasuk pengelola hotel terkait untuk melakukan kontrol kualitas pada layanan sewa hotel mereka, yang berakhir pada meningkatnya kepuasan pelanggan. *Sentiment Analysis* (SA) merupakan *tool* untuk melakukan analisis terhadap komentar-komentar tersebut. Permasalahan-permasalahan yang muncul pada sentiment analysis adalah tidak seimbangnya data komentar (*imbalanced datasets*) dalam hal jumlah dari masing-masing kelas, kemudian algoritma klasifikasi serta representasi fitur yang akan digunakan. Penelitian ini akan mencoba melihat bagaimana SMOTE (*Synthetic Minority Oversampling Technique*) dalam usaha menyeimbangkan jumlah data dari masing-masing kelas, penggunaan algoritma klasifikasi *Naïve Bayes*, *Logistic Regression*, dan *Support Vector Machine*, dan penggunaan representasi fitur *term presence*, *term occurrence*, dan TF-IDF dalam pengaruhnya terhadap hasil kinerja *sentiment analysis*. Penggunaan SMOTE terbilang cukup efektif dalam memperbaiki kinerja model pada kasus klasifikasi dengan data tidak seimbang, yang dibuktikan dengan peningkatan kinerja rata-rata model sebesar kurang lebih 12%. Representasi fitur *term occurrence* menghasilkan nilai *g-mean score* rata-rata sebesar 81,68%, kemudian *term presence* sebesar 79,89%, dan terakhir TF-IDF sebesar 79,31%. Sedangkan untuk algoritma klasifikasi, *Logistic Regression* menghasilkan nilai *g-mean score* rata-rata sebesar 81,65%, kemudian *Support Vector Machine* sebesar 81,55%, dan terakhir *Naïve Bayes* sebesar 77,68%.

Kata kunci: *Sentiment analysis*, hotel, Traveloka, *imbalanced datasets*, SMOTE, *g-mean score*

ABSTRACT

The comments on online hotel reservation services such as Traveloka is a very important resource that can be used by the service provider including hotel manager to quality control their hotel reservation service, which ends in increasing customer satisfaction. Sentiment Analysis (SA) is a tool for analyzing these comments. The problems that arise in sentiment analysis are the unequal number of each class of the data (imbalanced datasets), and the classification algorithm as well as the feature representation. This research will try to look at how SMOTE (Synthetic Minority Oversampling Technique) attempts to balance the amount of data from each class, the use of the Naïve Bayes, Logistic Regression, and Support Vector Machine classification algorithm, and the use of term presence, term occurrence, and TF-IDF feature representations in effect on the performance of sentiment analysis. The use of SMOTE is quite effective in improving model's classification performance when data is unbalanced, as evidenced by average model performance improvement of approximately 12%. Feature representation of term occurrence resulted in average 81.68% of g-mean score, then term presence 79.89%, and TF-IDF 79.31%. As for the classification algorithm, Logistic Regression resulted in average score 81.65% of g-mean score, then Support Vector Machine 81.55%, and Naïve Bayes 77.68%.

Keyword: Sentiment analysis, hotel, Traveloka, imbalanced datasets, SMOTE, g-mean score

KATA PENGANTAR

Segala puji syukur penulis ucapkan kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Pengaruh *Synthetic Minority Oversampling Technique* (SMOTE), Representasi Fitur, dan Algoritma Klasifikasi pada *Sentiment Analysis*" sehingga memperoleh gelar sarjana strata satu pada Departemen Ilmu Komputer/ Informatika pada Fakultas Sains dan Matematika Universitas Diponegoro.

Dalam penyusunan skripsi ini, penulis mendapat bantuan dan dukungan dari banyak pihak. Atas peran sertanya dalam membantu dalam penyelesaian skripsi ini, penulis ingin mengucapkan terimakasih kepada :

1. Prof. Dr. Widowati, S.Si, M.Si, selaku Dekan FSM UNDIP.
2. Dr. Retno Kusumaningrum, S.Si, M.Kom, selaku Ketua Departemen Ilmu Komputer/Informatika dan dosen pembimbing.
3. Helmie Arif Wibawa, S.Si, M.Cs, selaku Koordinator Skripsi.
4. Semua pihak yang telah membantu kelancaaran dalam penyusunan skripsi, yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa masih banyak kekurangan dalam penyusunan laporan skripsi ini, untuk itu penulis mengharapkan saran dan kritik yang bersifat membangun demi kesempurnaan Skripsi ini. Semoga laporan Skripsi ini dapat bermanfaat bagi pembaca pada umumnya dan penulis pada khususnya.

Semarang, 24 Mei 2018

Penulis,

Widi Satriaji

24010314130101

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

Bab pendahuluan membahas mengenai latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, dan sistematika penulisan Skripsi mengenai Pengaruh *Synthetic Minority Oversampling Technique* (SMOTE), Representasi Fitur, dan Algoritma Klasifikasi pada *Sentiment Analysis*.

1.1 Latar Belakang

Berkat menjamurnya perangkat berbasis internet, industri pesan kamar hotel secara *online* menjadi industri yang cukup populer di dunia belakangan ini, tidak terkecuali di Indonesia (Idris, 2017). Banyak layanan-layanan tersebut yang bermunculan baik berbasis *web* maupun aplikasi *smartphone*. Kebanyakan dari penyedia layanan tersebut memberikan kolom komentar bagi para penggunanya untuk mengutarakan kesan dan kritik terhadap kualitas yang mereka dapatkan dari hotel yang mereka pesan. Komentar-komentar tersebut merupakan sumber daya sangat penting yang bisa digunakan bagi pihak penyedia layanan sewa hotel termasuk pengelola hotel untuk melakukan kontrol kualitas pada layanan sewa hotel mereka, yang berakhir pada meningkatnya kepuasan pelanggan.

Sentiment Analysis (SA) merupakan *tool* untuk melakukan analisis terhadap komentar-komentar tersebut. SA, yang juga sering disebut sebagai *opinion mining*, adalah area riset dalam hal menganalisa ekspresi dari pendapat misalnya dari *web* (Mountassir, et al., 2012). Salah satu tugas dari SA, yaitu *sentiment classification*, memungkinkan kita untuk mendapatkan pengetahuan tentang bagaimana tanggapan pelanggan, apakah positif atau negatif, terhadap layanan hotel dari komentar-komentar dengan jumlah besar yang hampir tidak mungkin untuk dilakukan secara manual oleh manusia. Cambria et al. (2013) mengklasifikasikan teknik SA ke dalam empat kategori utama: *keyword-spotting, lexical affinity, concept-based,* dan *statistic and machine learning*. Teknik *machine learning* adalah teknik yang akan digunakan dalam penelitian ini.

Tidak seperti *keyword-spotting* maupun *lexical affinity* yang bersifat *stricly-ruled*/berbasis aturan-aturan kaku yang sudah didefinisikan sebelumnya, *machine*

learning dipilih karena bisa mencari pola aturan/*rule* secara mandiri. Teknik *machine learning* juga lebih dipilih dibanding teknik *concept-based* karena data yang akan diolah memiliki jumlah yang cukup memadai dan biaya usaha yang lebih rendah. Beberapa algoritma *machine learning* yang sering digunakan dan terbukti optimal untuk SA adalah *Naive Bayes, Logistic Regression,* dan *Support Vector Machine.* Penerapan algoritma Naïve Bayes pada SA pernah dilakukan oleh McCallum & Nigam (1998) dan menghasilkan akurasi sebesar 87%, Kaur & Mohana (2016) dengan akurasi sebesar 68,15%, dan Preety & Dahiya (2015) dengan akurasi sebesar 89%. Sedangkan untuk penerapan *Logistic Regression* pada SA pernah dilakukan oleh Al-Tahrawi (2015) dengan nilai *F1-Measure* 86,5%, Naradhipa & Purwarianti (2011) dengan akurasi sebesar 80%, serta Kaur & Mohana (2016) dengan akurasi sebesar 82,15%. Dan untuk *Support Vector Machine* pernah dilakukan oleh Al-Tahrawi (2015) dengan nilai *F1-Measure* sebesar 93,1%, Naradhipa & Purwarianti (2011) dengan akurasi 86,6%, dan Kaur & Mohana (2016) dengan akurasi sebesar 82,2%.

Terlihat bahwa ketiga algoritma klasifikasi tersebut cukup baik ketika digunakan untuk melakukan klasifikasi SA. Akan tetapi belum ada satupun penelitian yang memberikan klaim generalisasi yang luas. Pada umumnya penelitian-penelitian tersebut mengklaim kinerja algoritma *machine learning* yang digunakan adalah yang terbaik, padahal sebenarnya algoritma terpilih tersebut adalah bersifat *domain dependent*, yakni algoritma tersebut hanya baik untuk domain tertentu saja, seperti *review* hotel, *review* produk, *brand perception*, pengamatan *sentiment* pada tokoh atau partai politik, dan sebagainya. Oleh karena itu pada penelitian ini akan dibandingkan kinerja dari ketiga algoritma yang telah disebutkan sebelumnya untuk melakukan satu permasalahan, yaitu *Sentiment Analysis* pada komentar-komentar di suatu *website* pemesanan hotel tertentu.

Permasalahan yang muncul kemudian adalah kebanyakan komentar-komentar tersebut cenderung tidak seimbang (*imbalanced datasets*) dalam hal jumlah dari masing-masing kelas atau condong ke salah satu kutub, misal condong ke kelas negatif atau sebaliknya. Secara umum algoritma *machine learning* untuk klasifikasi akan menghasilkan suatu model dengan tingkat kepekaan yang minim terhadap kelas minoritas ketika menerima *imbalanced datasets* (He & Ma, 2013) dan telah terbukti menjadi permasalahan yang menantang bagi kalangan komunitas riset *machine learning*

(Kubat & Matwin, 1997) karena hal itu tentu saja akan mengakibatkan kinerja yang buruk terhadap SA yang akan dilakukan.

Namun, di kebanyakan penelitian tentang SA sering kali permasalahan mengenai ketidakseimbangan dataset tersebut tidak ditangani. Mereka mengasumsikan keseimbangan antara jumlah sampel data positif dan negatif. Sayangnya, asumsi tersebut tentu saja sering kali tidak tepat dalam keadaan nyata di lapangan. Data dengan distribusi kelas yang seimbang tidak bisa dijamin untuk didapat dari apapun subjeknya, termasuk untuk kasus *sentiment* pada komentar pengguna layanan pesan hotel *online*. Oleh karena itu penting untuk menangani masalah pada *imbalanced datasets*.

Beberapa pendekatan yang telah diajukan untuk menangani permasalahan ketidakseimbangan distribusi pada dataset di antaranya adalah, *re-sampling* (Chawla, et al., 2002), *one-class classification* (Juszczak & Duin, 2003), dan *cost-sensitive learning* (Zhou & Liu, 2006). Pendekatan pertama merupakan penanganan yang dilakukan dengan melakukan modifikasi pada dataset, yaitu dengan membentuk data sintetis untuk menyeimbangkan jumlah data dari masing-masing kelas, sedangkan pendekatan kedua dan ketiga berupaya menangani masalah tersebut dengan melakukan modifikasi pada algoritma *classifier*-nya. Pendekatan yang akan dilakukan pada penelitian ini adalah teknik *re-sampling*. *Re-sampling* itu sendiri secara umum dikelompokkan menjadi dua, *under-sampling* kelas mayor dan *over-sampling* kelas minor. Karena tidak ingin kehilangan data yang penting yang kemungkinan ditimbulkan oleh teknik *under-sampling* (Ah-Pine & Morales, 2016) dan karena jumlah dataset yang relatif kecil, penelitian ini akan menggunakan teknik *over-sampling* untuk penanganan *imbalanced datasets*.

SMOTE (*Synthetic Minority Over-sampling TEchnique*) merupakan salah satu teknik *over-sampling* yang sering digunakan untuk menangani masalah *imbalanced datasets* dengan membuat data sintetis pada kelas data minor sehingga data menjadi seimbang (Chawla, et al., 2002) dan diharapkan berimbas pada kinerja klasifikasi yang lebih baik (Ah-Pine & Morales, 2016). Bahkan di suatu kasus penelitian, terjadi peningkatan kinerja (dalam hal ini akurasi) dari sekitar 65% untuk distribusi data awal yang kemudian secara bertahap dilakukan SMOTE yang pada akhirnya data menjadi benar-benar seimbang dengan akurasi sekitar 80%, dengan rata-rata peningkatan akurasi

dari kondisi awal tanpa SMOTE ke kondisi data menjadi benar-benar imbang berada di 70% (Pears, et al., 2014).

Permasalahan ketiga yang sering muncul pada SA adalah mengenai pemilihan jenis representasi *feature vector*. Penelitian ini menggunakan pendekatan *Bag of Words* (BoW) dalam mengolah dan menganalisis dokumen teks. BoW dibentuk dengan melakukan tokenesisasi per kata (*unigram*) kemudian dilakukan penghapusan *stopwords* (seperti kata depan, kata sambung, dan sebagainya) yang sebelumnya sudah diberlakukan proses *stemming* (mengubah kata menjadi kata dasarnya). Kemudian representrasi ke bentuk vektor (*vectorization*) akan dibandingkan 3 jenis fitur, yaitu bentuk *term presence*, *term occurrence*, dan *term frequency-inverse document frequency* (TF-IDF). *Term presence* menunjukkan muncul atau tidaknya suatu kata pada suatu dokumen tertentu, yang dilambangkan dengan nilai *boolean* dalam bentuk skalar 1 untuk menandakan *presence* dan 0 untuk sebaliknya. Kemudian untuk *term occurrence* adalah vektorisasi yang menunjukkan frekuensi kemunculan suatu kata pada suatu dokumen. Kemudian yang terakhir adalah TF-IDF, yaitu *term occurrence* atau *term frequency* pada masing-masing kata pada dokumen setelah bobotnya dipengaruhi pula oleh frekuensi kemunculan suatu kata tersebut dari seluruh dokumen yang ada.

Penggunaan metrik *accuracy* (akurasi) bukan merupakan metrik yang cukup baik untuk melihat kinerja suatu model pada data yang tidak seimbang, karena akurasi tidak menunjukkan kepekaan klasifikasi untuk masing-masing kelas dari data yang akan diteliti (Bekkar, et al., 2013). Misalnya saja suatu hasil penelitian didapatkan akurasi sebesar 90% untuk data tidak seimbang, dengan distribusi 10% untuk kelas minor, dan 90% untuk kelas mayor, dan akurasi 90% tersebut didapat dengan mengklasifikasikan seluruh data tersebut ke dalam kelas mayor, itu artinya tidak ada data kelas minor yang terklasifikasi ke kelas yang seharusnya, dan itu tidak bagus.

Oleh karena itu, metrik kinerja yang akan diajukan untuk digunakan pada penelitian ini adalah *G-mean score*/*geometric mean score* (Kubat & Matwin, 1997; Bekkar, et al., 2013). Metrik ini menunjukkan keseimbangan kinerja baik untuk kelas mayor maupun minor, yang artinya adalah kinerja yang buruk pada kelas minor akan menghasilkan nilai *G-mean* yang buruk walaupun kinerja dari kelas mayor sangat baik (Shohei, et al., 2009). *G-mean* telah digunakan oleh beberapa peneliti untuk menilai kinerja klasifikasi pada *imbalanced datasets* (Ah-Pine & Morales, 2016; Mountassir, et

al., 2012; Li, et al., 2011; Xu, et al., 2015; Ertekin, et al., 2007; Karagiannopoulos M., et al., 2007; Su & Hasio, 2007; Zhang & Wang, 2013; Barua, et al., 2012).

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan sebelumnya dapat disusun rumusan masalah yaitu bagaimana melakukan *Sentiment Analysis* secara tepat terhadap dataset yang tidak seimbang dengan menerapkan algoritma SMOTE serta kombinasi beberapa algoritma dan beberapa jenis representasi fitur.

1.3 Tujuan dan Manfaat

Tujuan umum yang ingin dicapai dari penelitian Skripsi ini adalah mengetahui bagaimana mendapatkan model klasifikasi terbaik untuk *Sentiment Analysis* dilihat dari tiga aspek yaitu, penerapan SMOTE untuk melakukan *oversampling* dataset yang tidak seimbang, kemudian algoritma apa yang digunakan, dan fitur apa yang dipilih untuk merepresentasikan dokumen teks yang diolah. Secara lebih mendetail, tujuan tersebut dapat dijabarkan ke dalam beberapa tujuan khusus, yaitu:

1. Mengetahui bagaimana kinerja klasifikasi *Sentiment Analysis* sebelum dan sesudah dilakukan *oversampling* SMOTE terhadap *imbalanced datasets*.
2. Mengetahui pengaruh pemilihan representasi fitur yang digunakan terhadap kinerja *Sentiment Analysis* pada *imbalanced datasets*.
3. Mengetahui pengaruh dari masing-masing algoritma klasifikasi yang digunakan terhadap kinerja *Sentiment Analysis* pada *imbalanced datasets*.

Adapun manfaat yang diharapkan dari penelitian Skripsi ini adalah aplikasi yang telah dikembangkan dapat memberikan kontribusi terhadap penetilian terkait penanganan pada permasalahan ketidakseimbangan pada dataset yang dapat menjadi pandangan untuk penelitian-penelitian sejenis selanjutnya.

1.4 Ruang Lingkup

Ruang lingkup dalam melaksanakan penelitian Skripsi tentang Pengaruh *Synthetic Minority Oversampling Technique* (SMOTE), Representasi Fitur, dan Algoritma Klasifikasi pada *Sentiment Analysis* adalah sebagai berikut:

1. *Input* aplikasi ini adalah data komentar berbahasa Indonesia yang diperoleh dari *website* Traveloka sub layanan hotel melalui proses *scraping*.

2. Jumlah komentar berjumlah 1500 yang dipilih secara acak.
3. Beberapa teknologi yang digunakan di antaranya adalah bahasa pemrograman Python dan JavaScript, HTML, CSS, dan beberapa *library* pendukung.

1.5 Sistematika Penulisan

Sistematika yang digunakan dalam laporan Skripsi tentang Pengaruh *Synthetic Minority Oversampling Technique* (SMOTE), Representasi Fitur, dan Algoritma Klasifikasi pada *Sentiment Analysis* terbagi menjadi beberapa pokok bahasan, yaitu:

BAB I PENDAHULUAN

Bab pendahuluan membahas mengenai latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, dan sistematika penulisan Skripsi.

BAB II TINJAUAN PUSTAKA

Bab ini menjelaskan tentang keseluruhan dari teori-teori yang digunakan dalam pengerjaan penelitian, yaitu *Sentiment Analysis, imbalanced datasets* dan SMOTE, *preprocessing, k-fold cross validation*, algoritma klasifikasi yang digunakan, *performance metric*, dan pengembangan perangkat lunak.

BAB III METODOLOGI PENELITIAN

Bab ini menyajikan tahapan penyelesaian skripsi yang diawali dengan garis besar penyelesaian masalah dalam bentuk blok proses. Garis besar penyelesaian masalah diawali dengan *preprocessing* seperti penghilangan *stopwords* dan *stemming* untuk membentuk kamus kata, kemudian proses seleksi fitur dan vektorisasi, dilanjutkan dengan proses *resampling*, dan terakhir adalah evaluasi kinerja dari tiap algoritma yang digunakan.

BAB IV HASIL EKSPERIMEN DAN ANALISIS

Bab ini menjelaskan hasil eksperimen dan analisa yang telah dilakukan dimulai dari bagaimana cara pengumpulan data, penjelasan beberapa skenario eksperimen, dan analisa dari tiap hasil eksperimen.

BAB V PENUTUP

Bab ini berisi tentang kesimpulan dari uraian yang telah dijelaskan pada bab-bab sebelumnya beserta dengan saran yang dapat diajukan sebagai modal pengembangan penelitian lebih lanjut.