Semoga laporan skripsi ini dapat bermanfaat, baik sebagai sumber informasi maupun sumber inspirasi, bagi para pembaca.

Tangerang, 05 April 2020

Ahsanul Qalbi Fajar Islami

IMPLEMENTASI ALGORITMA RANDOM FOREST MENGGUNAKAN TF-IDF UNTUK ANALISIS SENTIMEN DENGAN PENERAPAN TRANSFER LEARNING

ABSTRAK

Penelitian ini melakukan pembahasan mengenai penerapan salah satu algoritma *machine learning* yaitu Random Forest yang digunakan untuk melakukan klasifikasi teks menjadi dua kategori yaitu positif dan negatif dengan menggunakan TF-IDF sebagai metode untuk mengubah teks berupa bahasa sehari-hari atau *natural language* menjadi vektor representasi. Penelitian ini menggunakan *dataset* berupa ulasan dari pengguna kepada Perusahaan Amazon, Yelp, dan IMDB. Pada penerapannya, penelitian ini menggunakan metode *transfer learning* untuk mengirimkan informasi yang didapatkan oleh model dari *dataset* yang telah di-*training* sebelumnya sebagai *starting point* untuk *dataset* selanjutnya pada proses *training*. *Transfer learning* yang diterapkan adalah *transfer learning feature importances*, yaitu dengan menggunakan informasi yang didapat dari model yang telah di-*training* sebelumnya berupa fitur atau *term* apa saja yang dianggap penting atau memiliki nilai *feature importance* lebih dari nol untuk dijadikan fitur atau *term training* pada *dataset* selanjutnya. Tujuan metode ini adalah untuk mengurangi fitur atau *term* pada *dataset* selanjutnya dan hanya mengambil fitur atau *term* yang berpengaruh saja. Pada penelitian ini juga menerapkan metode *transfer learning* dengan mengirim nilai frekuensi dokumen yang mengandung suatu *term* pada *dataset* selanjutnya yang akan di-*training*. Tujuan dari metode ini adalah untuk mengubah nilai IDF suatu fitur atau *term* dengan informasi dari model dengan *dataset* yang lebih besar dari *dataset* untuk membangun model selanjutnya agar dapat mengubah nilai kepentingan suatu *term* dengan informasi dari *dataset* yang lebih besar. Dengan penelitian ini diharapkan dapat mengetahui apakah penerapan metode *transfer learning* dapat memberikan hasil yang positif untuk melakukan klasifikasi sentimen menggunakan algoritma Random Forest, dan memberikan fasilitas untuk masyarakat untuk mengetahui sentimen dari suatu kalimat.

Kata Kunci: Analisis Sentimen, Random Forest Classifier, Klasifikasi, TF-IDF, *Transfer Learning*, *Feature Importance*, n-grams.

RANDOM FOREST ALGORITHM IMPLEMENTATION USING TF-IDF FOR SENTIMENT ANALYSIS WITH TRANSFERS LEARNING APPLICATION

ABSTRACT

This research discusses about one of machine learning algorithm, Random Forest, which is used to classify text into two categories, positive and negative by using TF-IDF as a method for converting text in the form of everyday language or natural language into a representation vector. This research use dataset in the form of user reviews to the Amazon, Yelp, and IMDB companies. In its application, this research uses the transfer learning method to send information obtained by the model from the dataset that has been previously trained as a starting point for the next dataset in the training process. Transfer learning method that is applied to this research is feature importances transfer learning, by using information obtained from the model that has been previously trained in the form of any feature or term that is considered important or has a feature importance value of more than zero to be used as a feature or term training in the next dataset . The purpose of this method is to reduce the features or terms in the next dataset and only take the features or terms that have effect to *dataset*. In this research also applies the transfer learning method by sending document frequency values that contain a term in the next dataset to be trained. The purpose of this method is to change the IDF value of a feature or term with information from a model with a dataset larger than the dataset to build the next model so that it can change the importance of a term with information from a larger dataset. With this research it is expected to know whether the application of the method of transfer learning can provide positive results for classifying sentiments using the Random Forest algorithm, and providing facilities for the public to find out the sentiments of a sentence.

Keywords: Sentiment Analysis, Random Forest Classifier, Classification, TF-IDF, Transfer Learning, Feature Importance, n-grams.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Analisis sentimen adalah salah satu bidang studi yang menganalisis pendapat, sentimen, evaluasi, penilaian, sikap, dan emosi seseorang terhadap suatu entitas tertentu seperti, jasa, organisasi, individu, masalah, peristiwa, topik dan atribut lainnya (Liu, 2012). Analisis sentimen dapat dimanfaatkan untuk melakukan ekstraksi opini-opini dari dokumen, komentar, sosial media, *review blog*, dan data-data lainnya. Analisis sentimen memliki beberapa metode, salah satu diantaranya adalah metode pendekatan *machine learning*. Analisis sentimen dengan pendekatan *machine learning* dapat menggunakan teknik klasifikasi.

Teknik klasifikasi adalah teknik dalam *data mining* untuk mengelompokkan data berdasarkan keterikatan data terhadap data sampel (Irvi Oktanisa, dan Afif Supianto 2018). Salah satu teknik klasifikasi adalah Random Forest. Random Forest adalah salah satu metode berbasis klasifikasi dan regresi dimana terdapat proses agregasi pohon keputusan (Dhawangkhara, dan Riksakomara 2017). Kelebihan dari algoritma Random Forest diantaranya adalah dapat menghindari *overfitting*, meminimalisir waktu *training data*, berjalan secara efisien pada data yang banyak, dan dapat mempertahankan akurasi walaupun sebagian data hilang.

Penelitian terkait mengenai klasifikasi teks dengan algoritma Random Forest dan TF-IDF telah banyak dilakukan salah satunya adalah penelitian oleh M AliFauzi. (2018) yang mengklasifikasikan teks Bahasa Indonesia dengan algoritma

Random Forest menggunakan fitur Bag-of-word dan metode pembobotan Term Frequency Inverse Document Fequency (TF.IDF). Disimpulkan bahwa penelitian tersebut memberikan hasil akurasi yang baik yaitu kinerja dengan skor out-of-bag (OOB) 0,829. Penelitian terkait dengan penggunaan TF-IDF selanjutnya adalah penelitian yang dilakukan oleh Yulius Denny dkk (2019) yang menerapkan algoritma TF-IDF untuk *text mining*. Penelitian tersebut menyimpulkan bahwa penggunaan TF-IDF merupakan metode yang tepat untuk digunakan dalam pencarian kata ditiap dokemen dan dapat membantu pengguna mendapatkan dokumen terkait sesuai dengan *query* yang telah di inputkan.

Penelitian terkait mengenai analisis sentimen dengan menggunakan teknik N-gram telah banyak dilakukan, salah satunya adalah penelitian oleh Wahyu Candra Inddhiarta (2017) yang melakukan analisis sentimen pemilihan kepala daerah Jakarta dengan menggunakan N-gram dan algoritma Naïve Bayes. Dari penelitian tersebut, disimpulkan bahwa berdasarkan ketiga jenis token N-gram yaitu unigram, bigram, dan trigram dengan metode Naïve Bayes, nilai akurasi tertinggi terdapat pada penggunaan bigram yaitu 0,823, menunjukan bahwa dengan menggunakan bigram ketepatan akurasi dari sistem lebih baik dari pada unigram dan trigram. Nilai presisi tertinggi juga terdapat pada penggunaan bigram yaitu 0,76. Namun pada nilai *recall* nilai tertinggi terdapat pada penggunaan trigram yaitu 0,898, sehingga disimpulkan penggunaan bigram dalam pengklasifikasian data lebih baik daripada menggunakan unigram atau trigram. Analisis sentimen dengan menggunakan *machine learning*, diperlukan *datasets* sebagai data *training* dan data *testing*. Metode pendekatan *machine learning* menghasilkan akurasi yang

baik, akan tetapi performa dari klasifikasinya bergantung pada dataset yang digunakan untuk data *training* masalah ini berkaitan dengan *transfer learning. Transfer learning* adalah metode Deep Learning yang menerapkan pengetahuan atau *knowledge* dari domain yang berbeda namun terkait ke domain tujuan (Di Zhang dkk, 2019). Berdasarkan penelitian yang telah ada sebelumnya, penelitian ini mencoba untuk menganalisis seberapa besar pengaruh *transfer learning* terhadap akurasi klasifikasi analisis sentimen dengan menggunakan algoritma Random Forest dengan bantuan N-gram dan TF-IDF.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijabatkan sebelumnya, rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana mengimplementasikan algoritma Random Forest dengan N-gram dan TF-IDF pada analisis sentimen?
2. Seberapa besar tingkat performa model yang dirancang dengan menggunakan metode *transfer learning* ?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini dapat dijabarkan menjadi beberapa poin sebagai berikut.

1. *Datasets* terdiri dari kumpulan *review* dari pengguna pada produk yang dimiliki oleh perusahaan Amazon, dan Yelp.
2. *Datasets review* yang digunakan berupa teks Bahasa Inggris.
3. *Transfer learning* akan dilakukan dengan mengirimkan pengetahuan dari

satu model Random Forest Classifier berupa fitur-fitur dan nilai kepentingan setiap fitur-fitur tersebut kepada model baru yang akan dibangun, sehingga model tersebut tidak menggunakan fitur yang dianggap tidak penting.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang dijelaskan sebelumnya, penelitian ini memiliki tujuan sebagai berikut.

1. Menerapkan algoritma Random Forest dengan menggunakan metode Bigrams dan TF-IDF untuk klasifikasi analisis sentimen *review* oleh *user*.
2. Mengukur dan mengetahui seberapa besar hasil performa model jika diterapkan metode *transfer learning* pada analisis sentimen dengan algoritma Random Forest.

1.5 Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat menerapkan metode *transfer learning* dengan algortima Random Forest pada analisis sentimen sehingga jika memberi pengaruh yang positif untuk performa model diharapkan dapat menanggulangi keterbatasan set data untuk membangun model yang serupa.

BAB II

LANDASAN TEORI

2.1 Analisis Sentimen

Sentimen analisis adalah riset komputasional dari opini sentimen dan emosi yang diekspresikan secara tekstual (Ira Zulfa, dan Edi Winarko 2017). Dalam analisis sentimen, teks data yang didapatkan akan diklasifikasikan menjadi beberapa jenis seperti teks sentimen "positif", "negatif", dan "netral". Pada penerapannya, analisis sentimen dimanfaatkan untuk memberikan nilai reputasi pada pelayanan pelanggan, produk perusahaan, dan reputasi seorang tokoh publik.

2.2 Word Embedding

Word Embedding adalah istilah yang digunakan untuk teknik mengubah sebuah kata menjadi sebuah *vector* atau *array* yang terdiri dari kumpulan angka. *Word Embedding* adalah sebuah pendekatan yang digunakan untuk. merepresentasikan *vector* kata. *Word Embedding* merupakan pengembangan komputasi permodelan kata-kata yang sederhana seperti perhitungan menggunakan jumlah dan frekuensi kemunculan kata dalam sebuah dokumen (Yulius Denny dkk, 2019).

Contoh cara tradisional untuk membaca teks dan mengubah menjadi vektor angka, misalnya terdapat sebuah kalimat yakni "sore ini merupakan sore yang indah". Langkah pertama adalah membuat sebuah *dictionary* yang berisi *list* dari seluruh kata yang *unique* atau tidak berulang, sehingga *dictionary* yang terbentuk adalah ["Sore", "ini", "merupakan", "yang", "indah"]. Langkah selanjutnya adalah

menggunakan metode *one-shot encoding* yang akan mengeluarkan *output* vektor berupa vektor ‘1’ merepresentasikan tempat kata tersebut pada *list*, dan vektor ‘0’ untuk merepresentasikan tempat kata lainnya. Contoh vektor representasi pada kata ‘merupakan’ mengacu pada metode *one-shot encoding* adalah [0, 0, 1, 0 ,0].

2.3 Teknik N-gram

Menurut Wahyu Candra Indhiarta (2017) N-gram merupakan penggabungan kata sifat yang sering muncul untuk menunjukan suatu sentimen. Teknik N-gram memiliki jenis-jenisnya berupa Unigram (n = 1), Bigram (n = 2), Trigram (n = 3), dan seterusnya. Berikut merupakan contoh penerapan N-gram pada kalimat “Pembelajaran mesin merupakan salah satu mata kuliah jurusan informatika”.

Tabel 2.1 Jenis-jenis N-Gram

N-Gram	Hasil Penerapan
Unigram	Pembelajaran, mesin, merupakan, salah, satu, mata, kuliah, jurusan, informatika
Bigram	Pembelajaran mesin, mesin merupakan, merupakan salah, salah satu, satu mata, mata kuliah, kuliah jurusan, jurusan informatika
Trigram	Pembelajaran mesin merupakan, mesin merupakan salah, merupakan salah satu, salah satu mata, satu mata kuliah, mata kuliah jurusan, kuliah jurusan informatika

2.4 Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency atau TF-IDF merupakan algoritma yang berguna untuk mengetahui bobot setiap kata atau seberapa sering kata tersebut. Musfiroh Nurjannah, dkk. (2013) menyatakan bahwa metode ini

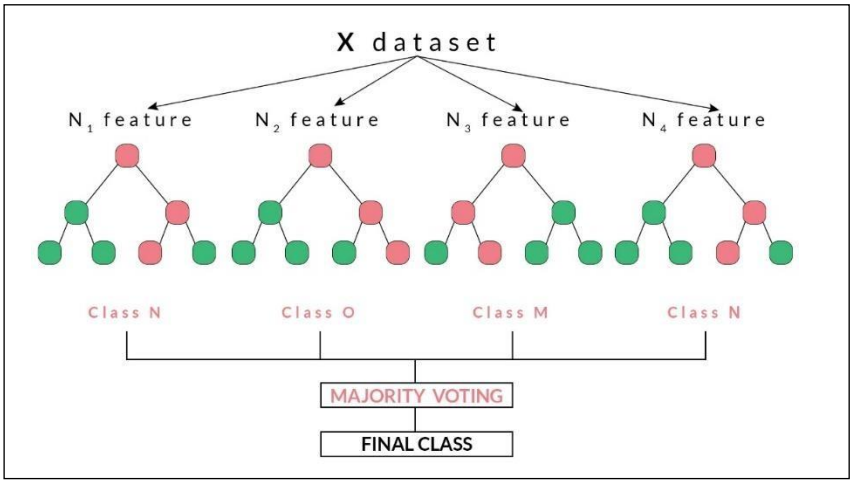
menggabungkan perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen (Term Frequency) tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut (Inverse Document Frequency).

Frekuensi kemunculan kata (Term Frequency) di dalam dokumen menunjukkan seberapa penting kata tersebut di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut (Inverse Document Frequency) menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen Musfiroh Nurjannah, dkk. (2013).

2.5 Random Forest Classifier

Random Forest adalah salah satu teknik *machine learning* yang dapat digunakan untuk melakukan klasifikasi. Random Forest merupakan salah satu metode dalam *decision tree*. Menurut (Aditya Yanuar, 2018) *decision tree* atau pohon pengambil keputusan adalah sebuah diagram alir yang berbentuk seperti pohon yang memiliki sebuah *root node* yang digunakan untuk mengumpulkan data, Sebuah *inner node* yang berada pada *root node* yang berisi tentang pertanyaan tentang data dan sebuah *leaf node* yang digunakan untuk memecahkan masalah serta membuat keputusan. Random Forest memiliki beberapa *decision tree*, kemudian algoritma Random Forest mengambil keputusan berdasarkan hasil *voting* terbanyak dari semua *decision tree*. Kelebihan dari Random Forest adalah jika terdapat data yang hilang dengan jumlah tertentu, Random Forest masih dapat melakukan klasifikasi dengan akurasi yang stabil karena tidak bergantung dengan

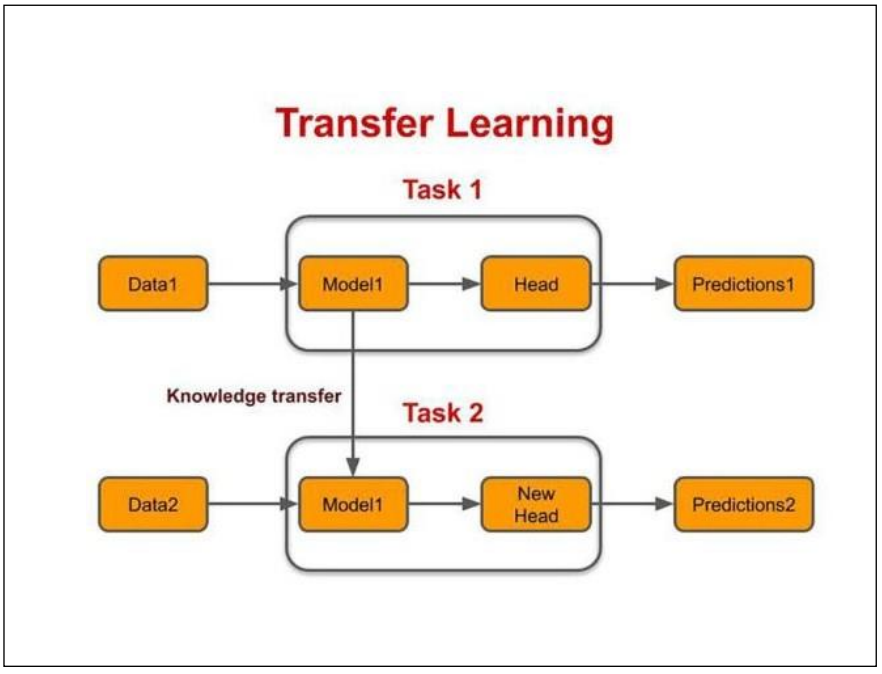
satu *decision tree* saja melainkan membandingkan data *voting decision tree* lainnya.



dGambar 1

Gambar 2.1 Struktur Algortima Random Forest
(Shagufta, 2019)

2.6 Transfer Learning



Gambar 2.2 Visualisasi gambaran *transfer learning*

(Pratik, 2019)

Transfer learning adalah metode Deep Learning yang menerapkan pengetahuan atau *knowledge* dari domain yang berbeda namun terkait ke domain tujuan (Di Zhang dkk, 2019). Menurut (Reza Fuad, 2018) *transfer learning* bertujuan untuk mengurangi penggunaan set data dengan skala besar. Pada penelitian ini *transfer learning* akan dilakukan dengan membangun satu model dasar Random Forest Classifier dan mengirimkan pengetahuan dari model tersebut berupa *feature importance* kepada model baru yang akan dibangun, sehingga tidak menggunakan fitur yang dianggap tidak penting dan akan mencoba mengirimkan nilai frekuensi dokumen yang memiliki suatu *term* pada model yang telah dilatih.