SKRIPSI

PENERAPAN METODE *SUPPORT VECTOR MACHINE* UNTUK MENGANALISIS SENTIMEN MASYARAKAT TERHADAP PENGGUNAAN IDOLA KOREA SELATAN SEBAGAI DUTA MEREK PRODUK LOKAL

VERRA ANDRIANI

NPM: 6161901065

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2023

FINAL PROJECT

APPLICATION OF THE SUPPORT VECTOR MACHINE METHOD FOR ANALYSIS OF PUBLIC SENTIMENT ON THE USE OF SOUTH KOREAN IDOLS AS BRAND AMBASSADORS FOR LOCAL PRODUCTS

VERRA ANDRIANI

NPM: 6161901065

DEPARTMENT OF MATHEMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2023

LEMBAR PENGESAHAN

PENERAPAN METODE *SUPPORT VECTOR MACHINE* UNTUK MENGANALISIS SENTIMEN MASYARAKAT TERHADAP PENGGUNAAN IDOLA KOREA SELATAN SEBAGAI DUTA MEREK PRODUK LOKAL

Verra Andriani

NPM: 6161901065

Bandung, 18 Agustus 2023

Menyetujui,

Pembimbing 1	Pembimbing 2
Liem Chin, M.Si.	Dr. Andreas Parama Wijaya

Ketua Penguji	Anggota Penguji
Maria Anestasia, M.Si., M.Act.Sc.	Dr. Daniel Salim

Mengetahui,

Ketua Program Studi

Dr. Livia Owén

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

PENERAPAN METODE *SUPPORT VECTOR MACHINE* UNTUK MENGANALISIS SENTIMEN MASYARAKAT TERHADAP PENGGUNAAN IDOLA KOREA SELATAN SEBAGAI DUTA MEREK PRODUK LOKAL

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
18 Agustus 2023

Verra Andriani
NPM: 6161901065

ABSTRAK

Fenomena budaya Korea Selatan *(hallyu wave)* mengakibatkan banyak bermunculan artis maupun kelompok *boyband* dan *girlband* yang menarik perhatian masyarakat serta perhatian merek produk khususnya di Indonesia. Banyak merek yang berlomba-lomba memilih artis atau anggota grup dari negara tersebut sebagai duta merek untuk mempromosikan produk-produknya. Pandangan masyarakat terhadap hal ini tentunya mengundang pendapat pro dan kontra atas keputusan merek dalam memilih artis dan anggota grup sebagai duta merek produknya. Pendapat mengenai pandangan masyarakat yang disampaikan pada media sosial dapat dianalisis menggunakan analisis sentimen. Namun, ekstraksi fitur adalah salah satu langkah penting dalam menganalisis sentimen, di mana TF-IDF dan *count vectorizer* merupakan metode ekstraksi fitur yang dapat digunakan. Ekstraksi fitur menghasilkan vektor yang dapat digunakan sebagai input untuk model analisis sentimen menggunakan metode klasifikasi. Pada skripsi ini, metode klasifikasi yang digunakan adalah *support vector machine* (SVM). Adapun data yang digunakan untuk klasifikasi ada sebanyak dua set, yaitu data sintesis dan data dari Twitter. Hasil klasifikasi dari masing-masing data set ini akan dibandingkan dengan menggunakan dua metode ekstraksi fitur. Hasil penelitian menunjukkan bahwa performa metode SVM dengan ekstraksi fitur TF-IDF mampu memberikan performa yang lebih baik pada kedua set data. Data sintesis dengan fungsi kernel linear metode SVM menghasilkan performa F1-*score* terbaik sebesar 80% dan data Twitter dengan fungsi kernel RBF metode SVM menghasilkan performa F1-*score* terbaik sebesar 73,47%.

Kata-kata kunci: analisis sentimen; ekstraksi fitur; *support vector machine*.

ABSTRACT

The phenomenon of South Korean culture, known as the hallyu wave, has led to the emergence of many artists as well as boybands and girlbands that capture the attention of the public, as well as the focus of brands, especially in Indonesia. Many brands compete to select artists or group members from that country as brand ambassadors to promote their products. The public's perception of this naturally invites both positive and negative opinions about brands' decisions in choosing artists and group members as brand ambassadors for their products. Opinions expressed by the public on social media can be analyzed using sentiment analysis. However, feature extraction is a crucial step in sentiment analysis, where TF-IDF and count vectorization are methods of feature extraction that can be employed. Feature extraction generates vectors that can serve as input for sentiment analysis models using classification methods. In this thesis, the classification method employed is the Support Vector Machine (SVM). Two sets of data are used for classification: synthetic data and data from Twitter. The classification results from each of these data sets will be compared using two feature extraction methods. The research findings demonstrate that the performance of the SVM method with TF-IDF feature extraction is superior in both data sets. The synthetic data, when combined with the linear kernel function of the SVM method, yields the best F1-score performance of 80%, while the Twitter data, when combined with the SVM method using the RBF kernel function, yields the best F1-score performance of 73,47%.

Keywords: sentiment analysis; feature extraction; support vector machine.

It's okay to make mistakes sometimes
Because anyone can do so
-Breathe by Lee Hi

KATA PENGANTAR

Puji dan syukur kepada Tuhan Yesus Kristus karena atas berkat, kasih karunia, dan pernyertaan-Nya, penulis dapat menyelasaikan skripsi ini sebagai salah satu syarat untuk mendapatkan gelar Sarjana Sains dari Program Studi Matematika Fakultas Teknologi Informasi dan Sains, Universitas Katolik Parahyangan. Dalam kesempatan ini, penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada pihak-pihak yang telah memberikan dukungan serta bantuan, yaitu kepada:

1. Papa, Mama, dan Koko yang telah mendukung penulis selama menjalankan studi hingga proses penulisan skripsi selesai.
2. Bapak Liem Chin, M.Si. selaku dosen pembimbing 1 dan Bapak Dr. Andreas Parama Wijaya selaku dosen pembimbing 2 yang sudah membimbing penulis selama proses penulisan skripsi, memberikan saran, nasihat, serta memberikan semangat kepada penulis sehingga skripsi ini dapat selesai.
3. Ibu Maria Anestasia, M.Si., M.Act.Sc. dan Bapak Dr. Daniel Salim selaku dosen penguji yang memberikan komentar serta saran agar skripsi ini dapat menjadi lebih baik lagi.
4. Seluruh dosen, staf tata usaha, serta karyawan UNPAR yang telah memberikan ilmu, bantuan, dan kenyamanan selama perkuliahan.
5. Natasya, Ezraes, Febry, dan Jessyca yang telah menjadi tempat berkeluh kesah dan mendukung penulis sejak dahulu.
6. Stephanie, Aspira, Abel, Aldi, Anton, teman seperjuangan yang saling mendukung untuk menyelesaikan skripsi serta Kirana, Sherina, Annisa dan teman-teman Matematika angkatan 2019 lainnya yang tidak dapat disebutkan satu-satu.
7. Seluruh pihak lain yang tidak dapat penulis sebutkan satu per satu. Semoga Tuhan membalas semua kebaikan yang telah diberikan.

Penulis menyadari bahwa penulisan skripsi ini masih banyak kekurangan. Oleh karena itu, penulis menerima kritik dan saran yang membangun untuk skripsi ini sehingga dapat menjadi lebih baik lagi. Penulis berharap skripsi ini dapat memberikan manfaat bagi siapa saja yang membacanya.

Bandung, 18 Agustus 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Fenomena budaya Korea Selatan yang mendunia, atau dikenal dengan *hallyu wave*, menawarkan hiburan seperti drama, film, dan musik yang digemari oleh berbagai kalangan masyarakat termasuk di Indonesia. Hal ini mengakibatkan banyak bermunculan artis maupun kelompok *boyband* dan *girlband* (atau lebih sering disebut dengan *idol group*) baru yang menarik banyak perhatian orang sehingga terbentuk komunitas penggemar idola tersebut. Fenomena ini tentunya menarik perhatian merek produk yang ada, khususnya di Indonesia. Banyak merek yang akhirnya berlomba-lomba memilih artis atau anggota grup dari negara tersebut sebagai duta (*brand ambassador*) untuk mempromosikan produk-produknya. Beberapa produk di Indonesia yang menggunakan idola Korea Selatan sebagai duta merek, antara lain adalah NCT Dream pada produk Lemonilo dan Somethinc, NCT 127 pada produk pasta gigi Click, Twice pada produk Scarlet, Sehun EXO pada produk Whitelab, dan produk-produk lokal lainnya.

Pandangan masyarakat terhadap hal ini mengundang pendapat pro dan kontra atas keputusan merek yang memilih menggunakan idola dan artis Korea sebagai duta merek produknya. Banyaknya penggemar suatu grup atau artis juga dapat menjadi peluang bisnis yang baik dan menjanjikan karena loyalitas yang diberikan penggemar untuk mendukung idolanya [1]. Pendapat pro dan kontra yang disampaikan masyarakat pada media sosial, seperti *Twitter*, *Instagram*, dan *Youtube*, dapat dianalisis untuk melihat pandangan masyarakat secara umum dalam keputusan suatu produk lokal yang menjadikan idola Korea Selatan sebagai duta merek produknya.

Pendapat mengenai pandangan masyarakat yang disampaikan pada media sosial tersebut dapat dianalisis menggunakan analisis sentimen. Analisis sentimen merupakan bidang dari *Natural Languange Processing* (NLP) untuk mengenali dan mengekstrak data sentimen yang diklasifikasi ke dalam sentimen yang positif dan negatif [2]. Tujuannya adalah memberikan informasi yang tersirat dari kumpulan data yang besar serta tidak terstruktur. Dari hasil tersebut, informasi yang relevan dapat berguna bagi pemilik merek dan masyarakat untuk memahami keseluruhan opini yang diberikan.

Salah satu langkah penting dalam menganalisis sentimen sebelum melakukan klasifikasi adalah proses ekstraksi fitur. *Term frequency - inverse document frequency* (TF-IDF) dan *count vectorizer* merupakan metode yang dapat digunakan dalam ekstraksi fitur untuk memberi bobot pada kata-kata yang terdapat dalam data. Bobot kata yang dihasilkan melalui ekstraksi fitur nantinya digunakan sebagai input untuk model analisis sentimen menggunakan metode klasifikasi.

Terdapat beberapa metode klasifikasi yang bisa digunakan, seperti pohon keputusan, *Naïve*

Bayes, *logistic regression*, *random forest*, K-*nearest neighbor* (K-NN), dan *support vector machine* (SVM). Di antara semua metode tersebut, klasifikasi dengan menggunakan SVM memiliki keunggulan dalam memberikan hasil akurasi yang tinggi dan bekerja dengan baik pada data yang memiliki ruang dimensi tinggi [3]. Penelitian [1] membahas mengenai analisis sentimen dari komentar di *Twitter* terhadap penggunaan artis Korea Selatan sebagai duta produk kecantikan lokal dengan menggunakan metode SVM dan *naïve Bayes*. Hasil terbaik diperoleh dari metode SVM dengan melihat performa kinerja model yang didapatkan berupa F1-*score* sebesar 91,06%. Lalu, pada penelitian [4] mengenai analisis sentimen masyarakat terhadap pembelajaran daring menggunakan metode K-NN didapatkan akurasi tertinggi saat $K = 10$ dengan akurasi sebesar 84,65%, presisi mencapai 87%, dan *recall* 86%. Terdapat juga penelitian [5] yang melakukan perbandingan metode klasifikasi teks, yaitu dengan metode regresi logistik, SVM, dan *gradient boosting* untuk menganalisis sentimen terhadap ulasan produk. Hasil penelitiannya menunjukkan bahwa ketiga metode tersebut memiliki performa yang baik, di mana F1-*score* untuk SVM dan regresi logistik sebesar 94% dan *gradient boosting* sebesar 92%.

Pada skripsi ini, digunakan metode SVM untuk mengklasifikasikan sentimen komentar-komentar di *Twitter* terhadap penggunaan idola Korea Selatan sebagai duta merek produk lokal. Fokus lain yang dibahas adalah perbandingan hasil klasifikasi terhadap 2 metode ekstraksi fitur, yaitu TF-IDF dan *count vectorizer*. Untuk lebih memahami keseluruhan proses, metode-metode diuji dengan data sintetis terlebih dahulu. Komentar-komentar buatan dibangkitkan secara manual dan diberi label sesuai dengan rasa (sentimen) penulis.

1.2 Rumusan Masalah

Beberapa masalah yang dapat dirumuskan pada skripsi ini adalah sebagai berikut:

1. Bagaimana proses ekstraksi fitur dengan metode TF-IDF dan *count vectorizer*?
2. Bagaimana implementasi ekstraksi fitur pada data sintesis dan data asli?
3. Bagaimana hasil analisis klasifikasi menggunakan metode SVM?

1.3 Tujuan

Berdasarkan rumusan masalah tersebut, tujuan dari skripsi ini adalah sebagai berikut:

1. Memahami proses ekstraksi fitur menggunakan TF-IDF dan *count vectorizer*.
2. Mengetahui hasil implementasi metode ekstraksi fitur untuk data sintesis dan data asli.
3. Menganalisis performa klasifikasi dengan metode SVM untuk data sintesis dan data asli.

1.4 *State of the Art*

Beberapa penelitian yang membahas penggunaan metode ekstraksi fitur pada analisis sentimen adalah sebagai berikut. Pada penelitian [6] mengenai analisis sentimen penanganan COVID-19 dengan *support vector machine*, peneliti melakukan uji perbandingan metode ekstraksi fitur dan leksikon

bahasa Indonesia. Tujuan dari penelitian tersebut adalah untuk mengukur kinerja kamus sentimen dan mengetahui pengaruh pemilihan metode ekstraksi fitur *term presence*, (*bag of words*) BoW, dan TF-IDF. Dalam perbandingan metode ekstraksi fitur baik pada tahap klasifikasi sentimen maupun evaluasi model, ditunjukkan bahwa penggunaan metode ekstraksi fitur TF-IDF menghasilkan skor akurasi tertinggi dibandingkan dengan metode ekstraksi fitur lainnya. Penggunaan TF-IDF menghasilkan nilai presisi, sensitivitas, dan F1-*score* rata-rata yang lebih tinggi pada label sebenarnya, yaitu sebesar 79,6%, 74,6%, dan 76,9%.

Pada penelitian [7] analisis eksperimental tentang dampak *count vectorizer* dan TF-IDF pada prediksi sentimen menggunakan model *deep learning*, peneliti melakukan perbandingan akurasi pada kedua teknik ekstraksi fitur. Tujuan penelitian tersebut adalah untuk menganalisis bagaimana masyarakat bereaksi terhadap COVID-19 dan berbagai tahapannya dengan dilakukan klasifikasi *tweet* ke dalam sentimen negatif, positif, atau netral. Penelitian [7] melaporkan bahwa dari kedua metode ekstraksi fitur yang digunakan, didapatkan TF-IDF lebih efisien dibandingkan dengan *count vectorizer* karena data yang digunakan memiliki volume yang besar dan dalam kasus penelitian yang digunakan, yaitu *tweet* mengenai COVID, keduanya memiliki performa yang hampir sama. Hal ini dapat dilihat dari hasil evaluasi performa keempat model klasifikasi yang digunakan, yaitu SVM, Bernoulli Naïve Bayes, *single-multi layer perceptron*, dan regresi logistik. Model SVM memiliki akurasi terbaik untuk kedua metode ekstraksi fitur, dengan akurasi untuk TF-IDF sebesar 93,15% dan *count vectorizer* sebesar 93,07%.

Pada penelitian [1] dilakukan perbandingan metode klasifikasi antara SVM dengan Naïve Bayes untuk menganalisis sentimen masyarakat terhadap penggunaan artis Korea Selatan sebagai duta produk lokal. Namun, dalam skripsi ini, dilakukan perbandingan metode ekstraksi fitur, yaitu antara TF-IDF dengan *count vectorizer* dan kemudian diklasifikasi menggunakan metode SVM. Selain itu, penerapan ekstraksi fitur dan metode SVM dilakukan pada dua set data, yaitu data sintesis dan data asli serta diberi label sesuai dengan sentimen penulis.

1.5 Sistematika Pembahasan

Bab 1 Pendahuluan
Bab ini berisi latar belakang, rumusan masalah, tujuan, dan sistematika pembahasan.
Bab 2 Landasan Teori
Bab ini berisi teori dan penjelasan mengenai pembelajaran mesin, ekstraksi fitur, *support vector machine* (SVM), dan evaluasi.
Bab 3 Proses dan Hasil Pengolahan Data Sintesis
Bab ini berisi penjelasan mengenai data sintesis, prapengolahan data, pelabelan data, ekstraksi fitur, *cross validation*, hasil klasifikasi data sintesis, dan algoritma klasifikasi dengan SVM.
Bab 4 Proses dan Hasil Pengolahan Data *Twitter*
Bab ini berisi penjelasan mengenai pengolahan data asli *Twitter* menggunakan proses yang sama dengan Bab 3 dan pembahasan.
Bab 5 Kesimpulan dan Saran
Bab ini berisi kesimpulan dari hasil analisis dan saran untuk penelitian selanjutnya.