PENERAPAN SVM DAN INFORMATION GAIN PADA ANALISIS SENTIMEN PELAKSANAAN PILKADA SAAT PANDEMI

[*]Aliffia Kulsumarwati[1)], Intan Purnamasari[2)], Budi Arif Dermawan[3)]

[123]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang

***Correspodence author**:* aliffia.kulsumarwati17048@student.unsika.ac.id

DOI: https://doi.org/10.37012/jtik.v7i2.641

Abstrak

Sosial media pada masa kini banyak dimanfaatkan untuk berbagai aktifitas, salah satunya adalah untuk menumpahkan segala tanggapannya terhadap kejadian-kejadian yang tengah terjadi di masyarakat. Seperti banyaknya masyarakat yang memberikan tanggapan terhadap kebijakan pemerintah Indonesia mengenai perlaksanaan Pilkada 2020 yang tetap diselenggarakan meski di tengah pandemi Covid-19 di Twitter. Berbagai tanggapan masyarakat ada yang mendukung maupun tidak setuju dengan diadakannya pilkada 2020 karna dilaksanakan di masa pandemi. Untuk itu maka dilakukan penerapan *data mining* dengan algoritma *Support Vector Machine* dan seleksi fitur *information gain* untuk menganalisis berbagai tanggapan masyarakat mengenai pelaksanaan pilkada 2020. Data yang digunakan merupakan *tweet* dari aplikasi Twitter sebanyak 496 data. Sebelum tahap *data mining*, dilakukan pembagian data menjadi 80% *data traning* dan 20% *data testing*. Hasil klasifikasi data *tweet* dengan *Support Vector Machine* menggunakan kernel linear menghasilkan nilai akurasi yang besar yaitu 92%, *precision* 90%, dan *recall* 92%.

Kata Kunci: *text mining,* pilkada, pandemi, *support vector machine, information gain*

Abstract

Social media is currently used for various activities like expressing responses to events that are happening in society. Like many people who have responded to the Indonesian government's policy regarding the implementation of the 2020 Pilkada is still being held even during the Covid-19 pandemic on Twitter. Various public responses have supported or disagreed with the holding of the 2020 local elections because they were carried out during the pandemic. For this reason, data mining is carried out with the Support Vector Machine algorithm and information gain feature selection to analyze various public responses regarding the implementation of the 2020 elections. The data used are tweets from the Twitter application as many as 496 data. Before the data mining stage, the data is divided into 80% of the training data and 20% of the testing data. The result of data classification with Support Vector Machine using a linear kernel produce accuracy of 92%, 90% precision, and 92% recall.

Keywords: *text mining, pilkada, pandemic, support vector machine, information gain*

PENDAHULUAN

Indonesia termasuk ke dalam negara yang terkontaminasi wabah virus Covid-19 dengan 2 kasus pertama sejak tanggal 2 Maret 2020 (Tosepu et al., 2020). Jumlah pasien yang menderita positif Covid-19 selalu ditemukan bertambah setiap harinya, berdasarkan data yang bersumber dari *SouthCity*, yaitu sebanyak 869.600 pasien dengan 711.205 pasien

dinyatakan sembuh dan 25.246 pasien meninggal dunia per tanggal 1 januari 2021 (SouthCity, 2021).

Pengendalian persebaran Covid-19 telah diupayakan oleh pemerintah dengan mengeluarkan pedoman mengenai anjuran protokol kesehatan kepada masyarakat untuk diterapkan di tempat dan fasilitas umum yang beberapa aturannya yaitu masyarakat diharuskan menggunakan masker dan face shield, mencuci tangan secara teratur, hingga diharuskan untuk menjaga jarak/physical distancing (Kementerian Kesehatan Republik Indonesia, 2020). Hal tersebut bertolak belakang dengan kebijakan pemerintah yang tetap menyelenggarakan Pilkada 2020 di masa pandemi Covid-19.

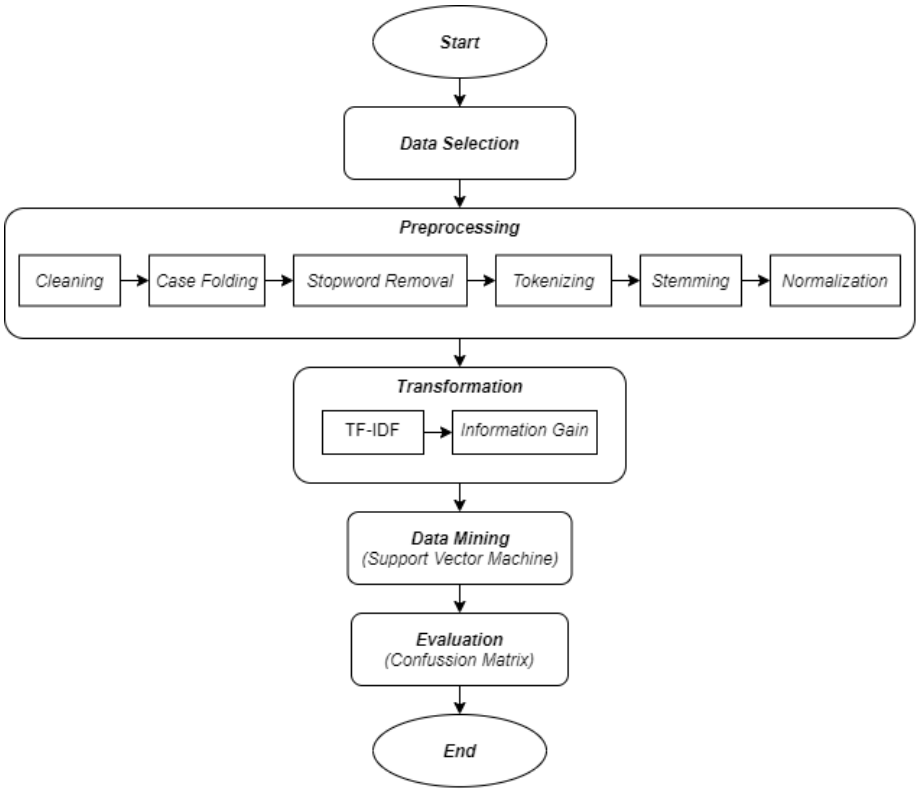
Rencana pengadaan Pilkada 2020 sempat ditunda dikarenakan pandemi, namun Keputusan Presiden (Keppres) Nomor 22 Tahun 2020 pemerintah menetapkan hari libur pada tanggal 9 Desember 2020 sebagai pelaksanaan Pilkada sesuai protokol kesehatan (Sekretariat Kabinet Republik Indonesia, 2020). Mengenai hal ini, seharusnya fokus pemerintah lebih tertuju pada hal-hal dasar dan fundamental seperti upaya penanganan virus Covid-19 serta memperhatikan masyarakat yang terdampak secara merata (Hasibuan, 2020).

Pelaksanaan Pilkada yang dilakukan di masa pandemi dapat mengundang masyarakat untuk beropini mengenai situasi tersebut. Sebagian besar masyarakat menyampaikan opininya pada aplikasi media sosial Twitter. Opini atau sentimen masyarakat yang dalam istilah Twitter biasa disebut sebagai “cuitan” ini dapat berupa opini negatif maupun positif. Dengan adanya perbedaan opini tersebut diperlukan sebuah analisis untuk mengklasifikasikan berbagai sentimen masyarakat mengenai pelaksanaan Pilkada yang berlangsung di masa pandemi. Analisis sentimen merupakan teknik untuk menganalisis berbagai pendapat, sentimen, sikap maupun emosi dari sebuah teks terhadap produk, jasa, individu, organisasi hingga topik dan sebuah peristiwa (Alvianda & Adikara, 2019).

Penelitian mengenai analisis sentimen Pilkada serentak pada tahun 2017 sebelumnya telah dilakukan menggunakan algoritma SVM dengan hasil akurasi yang tinggi yaitu sebesar 82% (Rahmawati et al., 2017). Penelitian selanjutnya mengenai analisis Twitter dengan metode NBC dan SVM yang dilakukan oleh Buntoro pada tahun 2016 menghasilkan akurasi pada metode SVM lebih tinggi dari metode NBC. Berdasarkan penelitian sebelumnya, maka penelitian ini akan menggunakan algoritma SVM karena memiliki akurasi yang baik. Perbedaan pada penelitian sebelumnya yaitu pada penelitian sebelumnya belum diterapkan teknik seleksi fitur. Oleh karena itu pada penelitian ini diterapkan seleksi fitur menggunakan *Information Gain.*

METODE PENELITIAN

Knowledge Discovery in Database (KDD) akan diterapkan sebagai metodologi pada penelitian ini. Menurut Kurgan dan Musilek dalam Siregar & Puspabhuana (2017) mengemukakan bahwa *data mining* dapat didefinisikan sebagai tahapan yang merupakan bagian dari proses KDD itu sendiri karena *data mining* pada umumnya digunakan oleh banyak peneliti sebagai penerapan dari proses KDD. Alur penelitian yang menerapkan metode KDD dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

Berikut ini merupakain uraian alur penelitian yang menerapkan metode KDD:

1. *Data Selection*

   Data selection merupakan tahapan pertama dalam KDD sebelum penggalian informasi untuk mengumpulkan data yang kemudian akan digunakan untuk melakukan proses analisisis. Tahap ini dimulai dengan melakukan pengambilan data berupa *tweet* pada aplikasi Twitter menggunakan teknik *crawling* dengan bahasa pemrograman *python*. "pilkada covid" dan "pilkada pandemi" merupakan kata kunci yang dipakai dalam *crawling data* yang di posting sejak tanggal 1 November 2020 hingga 30 Januari 2021.

2. *Preprocessing*

   Tahap preprocessing merupakan proses penyesuaian data dengan kriteria data yang dibutuhkan untuk penelitian. Tahapan ini terdiri dari *Cleaning* yaitu pembersihan kata dari karakter yang tidak diperlukan, *case folding, stemming, stopword removal dan normalization.*

3. *Transformation*

   Tahapan ini akan dilakukan pembobotan kata menggunakan TF-IDF. Setelah data telah memiliki bobot maka dilakukan seleksi fitur menggunakan information gain yang bertujuan untuk mengurangi jumlah fitur yang tidak terlalu berpengaruh. *Information Gain* adalah teknik pemilihan fitur yang dapat mengukur seberapa banyak informasi dalam keputusan klasifikasi yang benar dalam kategori apapun yang memengaruhi ada atau tidaknya kata. (Negara et al., 2020). Persamaan Information gain dapat dilihat dibawah ini:

$$Info(D) = -\sum_{i=1}^{m}(P_i)log_2(P_i)$$

   Dimana m adalah banyaknya kelas dan Pi adalah probabilitas bahwa sampel acak di partisi D *term*asuk dalam kelas Ci.

4. *Data Mining*

   Pada tahapan ini algortima SVM diimplementasikan untuk mengklasifikasi data menggunakan kernel Linear. SVM merupakan salah satu metode untuk melakukan klasifikasi dan bertujuan untuk membuat batas antara dua kelas atau dikenal dengan *Hyperplane* (Huang et al., 2018). Penggunaan kernel linear pada penelitian ini merujuk kepada hasil penelitian sebelumnya yang dilakukan oleh Irfani (2020) yaitu kernel linear menghasilkan akurasi terbaik dibandingkan kernel RBF dan *polynomial*. Persamaan SVM dapat dilihat dibawah ini:

$$k(x_i, x_j) = x_i^T x_j$$

5. *Evaluation*

   Pada tahapan ini akan dilakukan pengujian menggunakan *Confusion Matrix*. *Confusion Matrix* dilakukan untuk melihat kinerja dari *Classifier/Supervised Learning* dimana setiap kolom dari matriks mewakili sebuah kelas yang akan dilakukan proses prediksi dan baris matriks mewakili kelas yang sebenarnya (Nasir & Budiman, 2017). Menurut Andono et.al dalam (Amalina, 2019) klasifikasi memerlukan perhitungan beberapa data dalam tabel matriks menggunakan 4 (empat) istilah yaitu *True Positive* (TP), *False*

Positive (FP), *True* Negatif (TN) dan *False* Negatif (FN) untuk dijadikan sebagai representasi dari hasil klasifikasi. Menurut (Amrizal, 2018) dalam pengujian evaluasi ini, hasilnya akan berupa nilai *recall, precision, dan accuracy.*

HASIL DAN PEMBAHASAN

Data Selection

Kata kunci "pilkada covid" mengasilkan data sebanyak 1067 *tweet* sedangkan "kata kunci pilkada pandemi" menghasilkan sebanyak 2016 *tweet.* Data tersebut kemudian dilakukan penggabungan data dan dihapus data yang mengandung duplikan hingga menyisakan 737 data *tweet.* Selanjutnya dilakukan seleksi data yang mengandung sentimen negatif dan positif serta melakukan pelabelan data menggunakan kelas tersebut. Pelabelan kelas dilakukan secara manual dengan bantuan seorang Pakar Bahasa Indonesia. Sebanyak 498 *tweet* lolos tahap penyeleksian dan telah memiliki kelas sentimen. Data hasil *data selection* dan pelabelan ditunjukan pada Gambar 2.

	username	tweet	value
0	sakami1808	@KangenPiknik @blogdokter Kaya DKI gak pilkada...	negatif
1	dew0w1snu	Dengan partisipasi pemilih mencapai 76% meskip...	positif
2	alipuisimaji	@republikaonline Jangan cuma bacain laporan do...	negatif
3	dew0w1snu	Anggota @DPR_RI Komisi II menilai pelaksanaan ...	positif
4	aryprasetyo85	Perhelatan Pilkada yang digelar di tengah pand...	positif
...	...	...	...
493	safril_lamno	@tvOneNews Data dari satgas covid ribuan org r...	negatif
494	shae_ll	Pilkada Serentak 9 Desember 2020 yg telah berl...	positif
495	henkmizell	@detikcom *PARA PEJABAT DAERAH PARA PANITIA PE...	negatif
496	zakfaruq	@itsme_ini @msaid_didu Alhamdulillah semoga se...	negatif
497	aryprasetyo85	#ProkesPilkadaSukses Bawaslu Catat 2.126 Pelan...	negatif

498 rows × 4 columns

Gambar 2. Hasil Data Selection

Preprocessing

Pada tahapan ini terdapat beberapa proses yang di lakukan untuk membentuk data teks agar sesuai kriteria penelitian. Tabel 2 menunjukan contoh dari tahapan *preprocessing* yang sudah dilakukan.

Tabel 1. Contoh Text Preprocessing

Tahap preprocessing	Hasil
Dataset	Pemerintah: Belum ada bkti Pilkada 2020 jadi klaster penularan Covid-19 https://t.co/01trcD7VgU Mungkin @mohmahfudmd lagi ga ada kuota buat baca berita :)
Cleaning	Pemerintah Belum ada bkti Pilkada jadi klaster penularan Covid Mungkin lagi ga ada kuota buat baca berita
Case Folding	pemerintah belum ada bkti pilkada jadi klaster penularan covid mungkin lagi ga ada kuota buat baca berita
Stemming	pemerintah belum ada bkti pilkada jadi klaster tular covid mungkin lagi ga ada kuota buat baca berita
Stopword Removal	pemerintah belum bkti pilkada jadi klaster tular covid mungkin ga ada kuota buat baca berita
Tokenizing	Pemerintah, belum, bkti, pilkada, jadi, klaster, tular, covid, mungkin, ga, ada, kuota, buat, baca, berita
Normalization	Pemerintah, belum, bukti, pilkada, jadi, klaster, tular, covid, mungkin, tidak, ada, kuota, buat, baca, berita

Transformation

Proses pertama dalam tahapan ini dimulai dengan *splitting data* dengan persentase 80% *data training* dan 20% *data testing*. Jumlah *data training* sebanyak 396 sedangkan jumlah data testing sebanyak 100 data yang nantinya akan digunakan untuk dilakukan pengujian klasifikasinya. Setelah membagi data menjadi *data training* dan *data testing*, teknik selanjutnya yang dilakukan adalah menghitung *term* dari setiap dokumen dan dilanjutkan dengan pembobotan kata dengan TF-IDF. Hasil pembobotan dengan TF-TDF dapat dilihat pada Gambar 3.

	abah	abei	abis	abk	absen	absurd	acara
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...
391	0.0	0.0	0.0	0.0	0.0	0.0	0.0
392	0.0	0.0	0.0	0.0	0.0	0.0	0.0
393	0.0	0.0	0.0	0.0	0.0	0.0	0.0
394	0.0	0.0	0.0	0.0	0.0	0.0	0.0
395	0.0	0.0	0.0	0.0	0.0	0.0	0.0

396 rows × 1670 columns

Gambar 3. Hasil TF-IDF

Dari hasil TF-IDF pada Gambar 3 terbentuk matriks berukuran 396 X 1670 dimana terdapat 397 dokumen dan 1670 *term*. Setiap *term* sudah memiliki nilai TF-IDF. Nilai nol yang muncul merupakan *term* yang jarang muncul pada dokumen. Untuk itu pada tahapan

selanjutnya dilakukan proses seleksi fitur menggunakan algoritma *information gain* yang bertujuan untuk menghilangkan beberapa *term* yang paling jarang muncul atau fitur yang tidak terlalu berpengaruh. Proses *information gain* dilakukan dengan menggunakan *threshold* sebesar 0,0005. Gambar 4 menyajikan hasil dari penyeleksian data menggunakan *information gain*.

```
Treshold: 0.0005.
Previous number of features: 1670
Number of features after information gain: 800
```

Gambar 4. Hasil Seleksi Fitur

Data Mining dan *Evaluation*

Pada tahapan *data mining* dilakukan klasifikasi sentimen terhadap pelaksanaan pilkada 2020 menggunakan *Support Vector Machine* dengan kernel linear menghasilkan akurasi sebesar 92%. Kemudian dilakukan evaluasi pengujian menggunakan *confussion matrix.* Hasil dari algoritma *support vector machine* dengan pengujian *confusion matrix* ditunjukan pada Gambar 5.

```
              precision    recall  f1-score   support

           0       0.99      0.93      0.96        97
           1       0.22      0.67      0.33         3

    accuracy                           0.92       100
   macro avg       0.61      0.80      0.65       100
weighted avg       0.97      0.92      0.94       100

[[90  1]
 [ 7  2]]
```

Gambar 5. Hasil pengujian confusion matrix

Pada Gambar 5 menunjukan kernel linear dengan nilai akurasi 92%, *precision* 22%, dan *recall* 67% pada 90 data kelas negatif diprediksi dengan benar dan 2 data kelas positif diprediksi dengan benar.

KESIMPULAN DAN SARAN

Kesimpulan

1. Penelitian mengenai analisis sentimen twitter terhadap kebijakan Pilkada 2020 di masa pandemi menggunakan *Support Vector Machine* dan *Information Gain* menghasilkan bahwa seleksi fitur mampu mengubah fitur yang semula berjumlah 1670 menjadi 800 fitur. Selain itu agoritma SVM dengan kernel linear berhasil mendapatkan akurasi sebesar 92%, dengan nilai *precision* 22%, dan *recall* 67% dengan 90 data kelas negatif berhasil diprediksi dengan benar dan 2 data kelas positif diprediksi dengan benar.

2. Berdasarkan kuantitas kelas negatif yang lebih banyak daripada kelas positif maka diketahui bahwa mayoritas masyarakat tidak menyetujui akan pelaksanaan Pilkada 2020 karena pandemi.

Saran

Berdasarkan penelitian yang telah dilakukan maka terdapat beberapa saran atau masukan untuk penelitian selanjutnya, yaitu:

1. Membandingkan lebih dari satu algoritma seleksi fitur untuk mengetahui seleksi fitur manakah yang memiliki dampak terbaik untuk hasil akurasi. Selain *information gain* terdapat beberapa seleksi fitur lainnya seperti *chi-square* dan *genetic algorithm.*
2. Dikarenakan kernel RBF tidak dapat memprediksi data kelas positif, yang mungkin terjadi karena data yang tidak seimbang maka pada penelitian selanjutnya dapat menggunakan teknik *oversampling* atau *undersampling* untuk menyeimbangkan data.

REFERENSI

Alvianda, F., & Adikara, P. P. (2019). Analisis Sentimen Konten Radikal Di Media Sosial Twitter Menggunakan Metode Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *3*(1), 241–246.

Amalina, N. (2019). Uji Akurasi Aplikasi Augmented Reality Pembelajaran Huruf Alfabet Bahasa Isyaratindonesia (Bisindo) Pada Vuforia Menggunakan Confusion Matrix. Skripsi Oleh : Niela Amalina. *Universitas Islam Negerimaulana Malik Ibrahim, Malang*.

Amrizal, V. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). *Jurnal Teknik Informatika*, *11*(2), 149–164. https://doi.org/10.15408/jti.v11i2.8623

Buntoro, G. A. (2016). Analisis Sentimen Hatespeech Pada Twitter Dengan Metode Naïve Bayes Classifier Dan Support Vector Machine. *Jurnal Dinamika Informatika*, *5*.

Hasibuan, R. P. P. M. (2020). *Urgensitas Perppu Pilkada*. *4*(1), 121–128.

Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics and Proteomics*, *15*(1), 41–51. https://doi.org/10.21873/cgp.20063

Irfani, F. F. (2020). Analisis Sentimen Review Aplikasi Ruangguru Menggunakan

Algoritma Support Vector Machine. *JBMI (Jurnal Bisnis, Manajemen, Dan Informatika)*, *16*(3), 258–266. https://doi.org/10.26487/jbmi.v16i3.8607

Kementerian Kesehatan Republik Indonesia. (2020). Pedoman Pembatasan Sosial Berskala Besar dalam Rangka Percepatan Penanganan Corona Virus Disease 2019 (COVID-19).

Nasir, M. N., & Budiman, I. (2017). Perbandingan Pengaruh Nilai Centroid Awal Pada Algoritma K-Means Dan K-Means ++ Confusion Matrix. *Seminar Nasional Ilmu Komputer*, *1*, 118–127.

Negara, A. B. P., Muhardi, H., & Putri, I. M. (2020). Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(3), 599. https://doi.org/10.25126/jtiik.2020711947

Rahmawati, A., Marjuni, A., & Zeniarja, J. (2017). Analisis Sentimen Publik Pada Media Sosial Twitter Terhadap Pelaksanaan Pilkada Serentak Menggunakan Algoritma Support Vector Machine. *CCIT Journal*, *10*(2), 197–206. https://doi.org/10.33050/ccit.v10i2.539

Sekretariat Kabinet Republik Indonesia | Pemerintah Tetapkan Hari Pilkada Serentak 9 Desember 2020 Sebagai Libur Nasional- Sekretariat Kabinet Republik Indonesia. (2020). Retrieved from. https://setkab.go.id/pemerintah-tetapkan-hari-pilkada-serentak-9-desember-2020-sebagai-libur-nasional/

SouthCity. (2021). *Update Terkini Mengenai Virus Corona di Indonesia (15 Januari 2021)*. https://southcity.co.id/en/news-updates/update-terkini-mengenai-virus-corona-di-indonesia-15-januari-2021/. Diakses pada tanggal 15 agustus 2021.

Tosepu, R., Gunawan, J., Savitri, D., Ode, L., Imran, A., & Lestari, H. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of the Total Environment*, *725*.