Sentimen Twitter terhadap PILKADA kota Medan menggunakan metode Naïve Bayes

Prasetyo Mimboro*[1]

1 **Program Magister Teknik Informatika Program Pascasarjana Universitas Amikom Yogyakarta**
Jl. Padjajaran, Ring Road Utara, Kel. Condongcatur, Kec. Depok, Kab.Sleman, Prop. Daerah Istimewa Yogyakarta
prasetyo.mimboro@students.amikom.ac.id

Abstrak

Indonesia menjadi negara kelima terbesar pengguna Twitter sebanyak 19,5 juta pengguna. Seiring berkembangnya teknologi informasi, Twitter menjadi salah satu sumber informasi berdasarkan dari sentiment Twitter dan trending serta penggunaan hastag yang menjadi tranding. Belakangan ini vaksin nusantara menuai pro dan kontra, untuk dapat mengkalsifikasikan kalimat positif dan negative dalam sentiment Twitter terhadap vaksin nusantara maka membutuhkan data dari para pengguna Twitter dengan mengambil data berdasarkan kalsidikasi kalimat yang selanjutnya di proses data awal sebelum dimasukan ke dalam model indoBERT yang nantinya akan mengahasilkan tingkat akurasi sentiment Twitter terhadap vaksin nusantara. Indonesia memiliki 19,5 juta pengguna Twitter dari total 500 juta pengguna global dan terus berkembang dari waktu ke waktu. Pengguna Twitter memanfaatkannya sebagai forum terbuka kampanye oleh calon walikota Medan dan relawan mereka diminta Netizen menanggapi. Tanggapan warganet terhadap setiap tweet adalah Positif dan Negatif. Oleh karena itu, penelitian ini mencoba menganalisis tweet tentang sentimen netizen terhadap Pilkada Kota Medan 2020. Opini atau sentimen dari pengguna Twitter bisa tentunya dapat dijadikan sebagai kritik dan saran yang dapat ditampung oleh calon walikota dan wakil walikota Medan. Netizen Twitter sering pendapat tentang Calon Kepala Daerah melalui Unggahannya. Pendapat dari Netizen Twitter masih acak-acakan atau tidak terklasifikasi. Untuk memudahkan proses mengklasifikasikan data opini netizen membutuhkan Analisis Sentimen. Analisis Sentimen dilakukan dengan klasifikasi tweet yang mengandung sentimen Netizen terhadap Penyelenggaraan Pilkada Kota Medan 2020. Metode klasifikasi yang digunakan dalam penelitian ini adalah metode Naïve Bayes yang dikombinasikan dengan ekstraksi fitur TF-IDF. Uji validitas yang diterapkan pada penelitian ini menggunakan matriks konfusi. Dengan fitur tf-idf ekstraksi dan metode Naïve Bayes akan dapat secara otomatis mengklasifikasikan analisis sentimen dengan hasilr akurasi 76,00%.

Kata Kunci analisis sentimen, naive bayes, tf-idf, twitter

Digital Object Identifier 10.36802/jnanaloka.v3-no1-27-32

1 Pendahuluan

Semakin berkembangnya teknologi informasi, Twitter kini menjadi salah satu sumber informasi [1; 2; 3]. Pemanfaatan Twitter dapat dikembangkan menjadi data berdasarkan dari *tweet* yang di publikasi oleh penggunanya. Trending di Twitter dapat dijadikan bahan untuk mencari informasi tentang opini dari pengguna Twitter itu sendiri. Indonesia menjadi negara terbesar kelima pengguna Twitter [4]. Dalam media social Twitter pengguna dapat

* Corresponding author.

mengekspresikan dengan memperperbaharui status yang disebut dengan tweet kepada para pengikutnya atau saling membalas dan mention teman maupun orang lain. Analisis sentiment Twitter menjadi topik penelitian terhangat untuk mendapatkan informasi dan memahami opini publik [5]. Beberapa tahun terkahir, banyak artikel yang memebahas sentiment Twitter sebagai analisis dalam bidang saham, penggunaan bahasa slang, dan masih banyak lagi [5; 6; 7].

Opini dan opini netizen Twitter tentunya dapat dijadikan sebagai kritik dan saran yang dapat diterima oleh calon Walikota Medan dan Wakil Walikota. netizen Twitter sering memiliki pendapat tentang kandidat teratas di wilayah tersebut melalui unggahan. pendapat netizen Twitter belum acak atau dikategorikan. Analisis sentimen diperlukan untuk memfasilitasi metode dalam mengklasifikasikan data opini netizen. Analisis sentimen atau *opinion mining* merupakan bidang ilmu yang menganalisa pendapat, sentimen, evaluasi, penilaian, sikap dan emosi publik terhadap entitas seperti produk, jasa, organisasi, individu, masalah, peristiwa, topik, dan atribut mereka [8]. Keyakinan Analisis ini dapat digunakan sebagai acuan untuk mengklasifikasikan opini publik tentang politisi atau calon kepala daerah tertentu. Opini publik yang dimaksud adalah opini yang disebarkan di Internet atau di media sosial seperti Facebook dan Twitter. Dan lain-lain.

Proses penggalian opini dapat dilakukan dengan menggunakan metode *text mining* dan *machine learning*. Salah satu algoritma *text mining* adalah ekstraksi fitur Tfidf dengan metode Naïve Bayes. Analisis Sentimen Ulasan Film melakukan survei pada perbandingan algoritma klasifikasi pembelajaran mesin dan pemilihan fitur. Hasil perbandingan berasal dari penelitian ini. Algoritma terbaik adalah algoritma support vector machine dengan akurasi 81,10% dan area di bawah kurva 0.904 [9]. Survei tentang analisis sentimen data tweet menggunakan Delta Weighting TfIdf menggunakan model jaringan saraf tiruan dengan hasil bahwa pembobotan *delta* TFIDF lebih unggul dari pada TFIDF biasa, dibuktikan dengan hasil akurasi semua skenario. Delta TFIDF dan TFIDF secara verturut-turut memperoleh hasil akurasi tertinggi yaitu 70,6% dan 68,5% [10]. Analisis sentimen Twitter menggunakan metode support vector machine, dengan nilai akurasi mencapai 78,12% didapatkan dalam penelitian [7]. Metode K-Means dengan pengujian berdasarkan kata tweet diperoleh tingkat akurasi sebesar 92.80 % sedangkan pengujian berdasarkan tweet harian diperoleh tingkat akurasi sebesar 89.80 % di dalam penelitian [11].

Tujuan dari penelitian ini adalah mengidentifikasi kalimat yang mengandung kata negatif dan positif terhadap hasil analisis sentiment dengan menggunakan model kata negative dan positif berdasarkan opini masyarakat melalui Twitter dan secara otomatis mengklasifikasikan opini positif dan negatif pengguna internet Twitter terkait Pilkada Medan 2020.

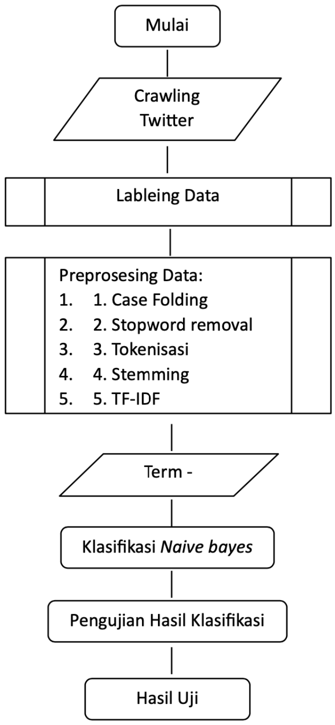
2 Metodologi

Proses pengambilan data penelitian ini dari Twitter dengan kata kunci "pilkada medan". Data yang berhasil diambil, diberi label negatif dan positif dengan peringkat yang ditentukan pengguna. Berdasarkan label, jumlah data yang diambil adalah tertampil pada Tabel 1.

Tabel 1 Sebaran dataset

sentimen	Jumlah data
negatif	210
positif	90
total	300

Data berdasarkan hasil *crawling* didapatkan sebanyak 300 baris record data untuk setiap periode, dan pembagian data uji adalah 20% dan 80% data latih. Tahapan proses pada sistem analisis sentiment terhadap opini masyarakat terhadap pilkada kota medan tahun 2020 pada jejaring sosial Twitter guna mengetahui pro dan kontranya terhadap masing-masing pasangan calon. Tahapan fase penelitian dapat dilihat pada Gambar 1.



Gambar 1 Tahapan analisis sentimen

Data set hasil *crawling* Twitter di ekspor ke dalam format MS Excel (*.xlsx) lalu dilakukan tahapan *preprosesing* data dengan menggunakan Google Colab. *Preprocessing* merupakan proses perubahan data tidak terstruktur menjadi data terstruktur untuk digunakan pada proses selanjutnya. Proses ini dilakukan untuk menyeleksi data yang akan diolah pada sistem untuk menghasilkan data yang sesuai dan terstruktur dengan baik. *Preprocessing* pada umumnya terdiri dari, *cleansing*, tokenisasi, *casefolding*, *stopword removal* dan *stemming.*

Pembersihan adalah langkah untuk menghapus tweet dari kata-kata yang tidak diperlukan seperti tagar, nama pengguna atau url. Tokenisasi digunakan untuk membagi kalimat, paragraf, atau frasa menjadi sebuah kata. Dalam proses tokenisasi juga memungkinkan Anda untuk menghapus karakter tanda baca atau pungutuasi. Perubahan huruf kapital menjadi non kapital merupakan tahapan *casefolding* sedangkan penghapusan kata yang dianggap stopword dilakukan seperti pada kata, jika, maka, di, di, juga hanya, tetapi hanya dan lain sebagainya. Proses penting dalam prepocessing adalah menerjemahkan kata dalam dokumen menjadi akar kata (*root words*) menurut aturan tertentu. Akar kata atau kata dasar ini diturunkan dengan menghilangkan awalan, sisipan atau akhiran.

Salah metode ekstraksi fitur dapat digunakan untuk memberikan nilai pada setiap kata dalam data latih adalah dengan TF-IDF (*term frequency and inverse document frequency*). Sistem yang melakukan klasifikasi diharapkan dapat mengklasifikasikan semua data dengan benar, namun tidak dapat dihindari bahwa kinerja sistem tidak dapat 100% benar, sehingga sistem klasifikasi juga harus diukur kinerjanya. Pengukuran kinerja untuk klasifikasi biasanya dilakukan dengan menggunakan matriks konfusi. Matriks konfusi adalah tabel yang mencatat hasil kinerja klasifikasi.

Klasifikasi Naïve Bayes adalah metode klasifikasi yang menggunakan metode statistik dan probabilistik yang diusulkan oleh ilmuwan Thomas Bayes. Perbandingan penggunaan Naïves Bayes dengan SVM dilakukan dalam [12] klasifikasi emosi restoran online, naive bay mengungguli SVM (*support vector machine*), yang mencapai akurasi 95,33% di Naïves Bayes melampaui dan SVM dengan akurasi 90%.

3 Hasil dan pembahasan

Diskriminasi negatif dan positif dilakukan pada data yang berhasil diambil. Pemberian bobot dengan TFIDF ditunjukkan pada Gambar 2.

	Doc 1	Doc 2	Doc 3	Doc 4	...	Doc 17	Doc 18	Doc 19	Doc 20
akhyarsalman	1	1	1	1	...	0	1	1	0
medan	2	0	1	0	...	1	0	1	0
di	2	0	1	0	...	2	0	0	0
tim	0	0	0	0	...	1	0	1	0
suara	0	0	2	1	...	0	0	0	0
akhyar	0	0	0	0	...	1	0	0	0
tak	0	1	0	1	...	0	0	0	0
berita	0	0	0	0	...	0	0	0	0
kpu	0	0	0	0	...	0	1	0	0
penggelembungan	0	0	1	1	...	0	0	0	0
dugaan	0	0	1	0	...	0	0	0	0
rekapitulasi	0	0	0	0	...	0	0	0	0
hasil	0	0	1	0	...	0	0	0	0
pilkada	0	0	1	0	...	0	0	1	0
pleno	0	0	0	0	...	0	1	0	0
acara	0	0	0	0	...	0	0	0	0
bobbyaulia	1	1	1	0	...	0	0	0	0
dari	0	1	0	1	...	0	0	0	0
kejanggalan	0	0	0	0	...	0	0	1	0
rapat	0	0	0	0	...	0	1	0	0
kalah	0	0	0	1	...	0	0	0	0
unggul	0	1	0	0	...	0	0	0	0
ada	0	0	0	0	...	0	0	0	0
tims	0	0	0	0	...	0	0	0	0
saksi	0	0	0	0	...	0	1	0	0
golput	3	0	0	0	...	0	0	0	0
salman	0	0	0	0	...	1	0	0	0
ogah	0	0	0	0	...	0	0	0	0
teken	0	0	0	0	...	0	0	0	0
soal	0	0	1	1	...	0	0	0	0

[30 rows x 20 columns]

Gambar 2 Matrix term 20 kata teratas

Berdasarkan pelabelan dataset diperoleh hasil sentimen dengan total jumlah data sebanyak 300 baris tweet dengan komposisi negatif dan positif adalah 70,00% dibanding 30,00%. Untuk mengetahui tingkat akurasi pengujian dengan metode Naïve Bayes dilakukan pengujian salah satunya dengan menggunakan *confusion matrix.* Hasil pengujian dengan metode Naïve Bayes adalah seperti tertampil dalam Gambar 3.

Dari data di Gambar 3, *confusion matrix* tersebut dapat dilihat bahwa nilai TP (*true positive*) adalah 34 dan nilai TN (*true negative*) adalah 5. Adapun nilai akurasi, presisi, recall dan f1-score yang didapatkan dari penelitian ini adalah 65%, 78%, 65% dan 57%.

```
['Negative' 'Negative' 'Negative' 'Negative' 'Negative' 'Negative'
 'Negative' 'Negative' 'Negative' 'Negative' 'Negative' 'Negative'
 'Negative' 'Negative' 'Negative' 'Negative' 'Negative' 'Positive'
 'Negative' 'Negative' 'Negative' 'Negative' 'Negative' 'Negative'
 'Positive' 'Negative' 'Negative' 'Negative' 'Negative' 'Negative'
 'Negative' 'Negative' 'Negative' 'Negative' 'Positive' 'Positive'
 'Negative' 'Negative' 'Negative' 'Negative' 'Positive' 'Negative'
 'Negative' 'Negative' 'Negative' 'Negative' 'Negative' 'Negative'
 'Negative' 'Negative' 'Negative' 'Negative' 'Negative' 'Negative'
 'Negative' 'Negative' 'Negative' 'Negative' 'Negative' 'Negative']
MultinominalNB Accuracy: 0.65
MultinominalNB Precision: 0.7836363636363636
MultinominalNB Recall: 0.65
MultinonimalNB F1_score: 0.5727437477346864
confusion matrix:
 [[34  0]
 [21  5]]
```

Gambar 3 Pengujian metode Naïves Bayes

4 Kesimpulan dan saran

Penentuan klasifikasi sentimen dapat dilakukan menggunakan metode Naïve Bayes dengan kecenderungan tweet ditentukan oleh teks *processing*. Nilai akurasi sebesar 65% dan masih adanya nilai *false negatif* sebanyak 21 perlu menjadi perhatian tersendiri. Upaya untuk meningkatkan akurasi dan mengurangi kesalahan tipe II dalam matrik konfusi bisa dilakukan dalam penelitian selanjutnya.

Pustaka

1 H. Rosenberg, S. Syed, and S. Rezaie, "The twitter pandemic: The critical role of twitter in the dissemination of medical information and misinformation during the covid-19 pandemic," *Canadian journal of emergency medicine*, vol. 22, no. 4, pp. 418–421, 2020.

2 B. Mønsted, P. Sapieżyński, E. Ferrara, and S. Lehmann, "Evidence of complex contagion of information in social media: An experiment using twitter bots," *PloS one*, vol. 12, no. 9, p. e0184148, 2017.

3 K. Rudra, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting and summarizing situational information from the twitter social media during disasters," *ACM Transactions on the Web (TWEB)*, vol. 12, no. 3, pp. 1–35, 2018.

4 "Pengguna twitter indonesia masuk daftar terbanyak di dunia, urutan berapa?" 2022, diakses 2 Januari 2022. [Online]. Available: https://databoks.katadata.co.id/datapublish/2022/03/23/pengguna-twitter-indonesia-masuk-daftar-terbanyak-di-dunia-urutan-berapa

5 A. Reyes-Menendez, J. R. Saura, and C. Alvarez-Alonso, "Understanding# worldenvironmentday user opinions in twitter: A topic-based sentiment analysis approach," *International journal of environmental research and public health*, vol. 15, no. 11, p. 2537, 2018.

6 C. Gu and A. Kurov, "Informational role of social media: Evidence from twitter sentiment," *Journal of Banking & Finance*, vol. 121, p. 105969, 2020.

7 A. Novantirani, M. K. Sabariah, and V. Effendy, "Analisis sentimen pada twitter untuk mengenai penggunaan transportasi umum darat dalam kota dengan metode support vector machine," *eProceedings of Engineering*, vol. 2, no. 1, 2015.

8 B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

9 V. Chandani, R. S. Wahono *et al.*, "Komparasi algoritma klasifikasi machine learning dan feature selection pada analisis sentimen review film," *Journal of Intelligent Systems*, vol. 1, no. 1, pp. 56–60, 2015.

10 C. Amalia and Y. Sibaroni, "Analisis sentimen data tweet menggunakan model jaringan saraf tiruan dengan pembobotan delta tf-idf," *eProceedings of Engineering*, vol. 7, no. 2, 2020.

11 A. Faesal, A. Muslim, A. H. Ruger, and K. Kusrini, "Sentimen analisis pada data tweet pengguna twitter terhadap produk penjualan toko online menggunakan metode k-means," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 19, no. 2, pp. 207–213, 2020.

12 Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of internet restaurant reviews written in cantonese," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7674–7682, 2011.