REPOSITOR, Vol. 3, No. 2, Februari 2021, Pp. 191-198
ISSN : 2714-7975
E-ISSN : 2716-1382

Sentimen Analisis Tweet Berbahasa Indonesia Pada Pilkada Serentak 2020 Menggunakan Metode Naïve Bayes Berbasis Particle Swarm Optimization

Adam Novrisal[*1], Gita Indah Marthasari[2], Christian Sri Kusuma Aditya[3]
[1,2,3]Universitas Muhammadiyah Malang
damsnov20@gmail.com[*1], gita@umm.ac.id[2], christianskaditya@umm.ac.id[3]

Abstrak

Analisis sentiment merupakan cabang dari text mining, fokus utamanya merupakan menganalisa dokumen berupa teks. Tweet yang berupa teks tersebut dibagi menjadi dua class yaitu class positif dan negatif. Algoritma yang digunakan pada penelitian ini yaitu Naïve Bayes berbasis Particle Swarm Optimization yang digunakan untuk mengetahui apakah terdapat peningkatan akurasi pada hasil klasifikasi. Dataset yang digunakan sebanyak 1000 dan diujikan menggunakan 10 Fold Cross Validation. Hasil klasifikasi yang didapatkan dari penelitian ini menghasilkan akurasi sebesar 81%. Hasil tersebut lebih baik dibandingkan dengan hasil klasifikasi menggunakan Naïve Bayes tanpa ada proses seleksi fitur dengan Particle Swarm Optimization dengan hasil akurasi sebesar 74.14%.

Kata Kunci: *Analisis sentimen, Naïve Bayes, Seleksi Fitur, Python*

Abstract

Sentiment analysis is a form of text mining, the main focus is analyzing text documents. Tweets in the form of text are divided into two classes, namely positive and negative classes. The algorithm used in this study is naïve bayes based on particle swarm optimization which is used to determine whether there is an increase in the accuracy of the classification results. The dataset used is 1000 and tested using 10 fold cross validation. The classification results obtained from this study produce an accuracy of 81% these results are better than the classification results using naïve bayes without any features selection process with particle swarm optimization an accuracy of 74.14%

Keywords: *Sentiment analysis, Naïve Bayes, Feature Selection, Python*

1. Pendahuluan

Indonesia merupakan sebuah negara dengan sistem politik demokrasi. Hal tersebut ditandai dengan diselenggarakannya pemilihan kepala daerah (pilkada) terhadap calon kepala daerah dan wakil kepala daerah. Pemilihan kepala daerah yang dilakukan secara serentak mulai tahun 2015 untuk mengganti kepala daerah dan wakil kepala daerah yang telah menjabat selama 5 tahun. Kebijakan pemilihan kepala daerah secara serentak ini mulai diresmikan pada tahun 2015. Dengan segala pro dan kontra dengan diadakan pemilihan kepala daerah secara serentak di tahun 2020 ini tidak hanya nampak di dunia nyata, namun juga di dunia jejaring sosial seperti twitter. Berdasarkan portal web *tagar.id* tidak semua daerah yang menyelenggarakan pilkada serentak di tahun 2020 setuju dengan kebijakan yang ditetapkan oleh pemerintah bahkan pemerintah itu sendiri menginginkan jika evaluasi mengenai pilkada serentak namun, hal tersebut dibantah oleh kemendagri mengatakan jika evaluasi tersebut bukan mengembalikan pilkada melalui DPRD. Menurut catatan jika pilkada langsung melalui DPRD memiliki dampak yang tidak baik dalam sistem sosial masyarakat indonesia yaitu potensi konflik yang tinggi di beberapa daerah seperti Papua dan Aceh[1]. Hal tersebut yang memunculkan opini pro dan kontra pada pilkada serentak tersebut. Berikut contoh tweet berupa opini pro (positif) ”Mari kita wujudkan Pilkada Serentak 2020 yang sejuk, aman, dan damai” dan kontra (negatif) ”Boikot PON PAPUA 2020, Boikot Pilkada Serentak 2020, Boikot OTSUS, Boikot Pemekaran Provinsi maupun Kabupaten/Kota di TANAH PAPUA. No Dialog, Referendum Yes”.

Perkembangan penggunaan twitter sangat begitu cepat dari awal kemunculannya. Berdasarkan portal website *techno.id* pengguna aktif media sosial twitter sebanyak 302 juta pengguna aktif yang 80 persennya berasal dari perangkat mobile dengan rentan usia

penggunanya berusia 18 – 29 tahun sebanyak 37 persen dan 25 persen berusia 30 – 49 tahun[2]. Twitter telah menjadi sebuah media sosial yang banyak digunakan oleh para pengguna media sosial untuk berkomunikasi. Pengguna twitter dapat mem-posting tweet dengan batas hanya 140 karakter tweet. Tweet merupakan sebuah pesan yang berisi informasi mengenai keluh kesah para pengguna media sosial twitter. Isi tweet dapat digunakan untuk mengungkapkan sudut pandang terhadap pemilihan kepala daerah secara serentak di tahun 2020 yang bisa bersifat opini atau penilaian secara subjektif. Opini di dalam setiap tweet tersebut yang nantinya akan dimanfaatkan untuk melihat bagaimana sentimen masyarakat khususnya pengguna media sosial twitter terhadap pemilihan kepala daerah secara serentak di tahun 2020.

Terdapat banyak teknik klasifikasi untuk bisa mengetahui seperti apa sentimen masyarakat mengenai pemilihan kepala daerah secara serentak ditahun 2020 ini diantaranya *Naive bayes classifier, Support Vector Machine, dan Decision Trees.* Dalam penelitian lainnya yang melakukan analisis sentimen pada twitter mengenai Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes dan Pembobotan *Emoji* dari hasil pengujian tersebut didapatkan hasil akurasi sebesar 68,52% untuk pembobotan tekstual, 75,93% untuk pembobotan non-tekstual, dan 74,81% untuk penggabungan kedua pembobotan[3]. Pada penelitian lainnya tentang Penerapan Particle Swarm Optimization (PSO) Untuk Klasifikasi Dan Analsisi Kredit Menggunakan Algoritma C4.5 hasil akurasi yang didapat dari penelitian tersebut sebesar 70% dengan mengkombinasikan algoritma C4.5 dengan *Particle Swarm Optimization* [4]. Pada penelitian lainnya mengenai Analisis Sentimen Opini Publik Bahas Indonesia Terhadap Wisata TMII Menggunakan *Naive Bayes* Dan *PSO* mendapatkan hasil akurasi sebesar 94% setelah menggunakan algoritma *Particle Swarm Optimization* [5].

Dalam penelitian kali ini akan mengintegrasikan metode *Naive Bayes* dan *Feature Selection Particle Swarm Optimization* untuk menganalisis opini masyarakat terhadap pelaksanaan pilkada serentak ditahun 2020 melalui tweet berbahasa indonesia di twitter yang diharapkan dapat meningkatkan tingkat akurasi tersebut.

2. Metode Penelitian

2.1 Pengumpulan Data

Pengumpulan data dilakukan dengan menggunakan *python* dengan menggunakan *library* yang telah tersedia untuk siap digunakan. *Library* tersebut yang akan digunakan adalah twitterscraper yang berbasis bahasa *python*. Untuk bisa menggunakan *library* tersebut harus *install* dengan *syntax* “pip install twitterscraper” pada direktori *file python* tersebut.

2.2 Preprocessing

Pada tahap preprocessing ini akan dilakukan sebuah proses untuk membuat data hasil *crawling* layak untuk digunakan pada proses selanjutnya [6]. Terdapat beberapa tahap dalam proses ini yaitu, *case folding, punctuational removal, stopword removal, stemming,* dan *tokenizing*.

2.2.1 CaseFolding

Pada tahap *case folding* terhadap sebuah proses untuk merubah huruf kapital menjadi huruf kecil serta menghilangkan tanda baca dan angka. Cara kerja dari *case folding* adalah memproses huruf alfabhet dari “a” hingga “z” saja sehingga karakter selain huruf tersebut akan dihapus [7].

2.2.2 Filtering

1. Stopwords

Stopwords adalah kosakata yang bukan termasuk kata ciri pada suatu dokumen dan tidak menyampaikan pesan apapun secara signifikan pada teks atau kalimat[8]. Kosakata yang dimaksud seperti kata penghubung dan kata keterangan yang bukan merupakan kata unik, seperti “dari”, “akan”, dan sebagainya.

2. Removal URL

Removal URL adalah proses untuk menghapuskan URL atau alamat *website* yang terdapat didalam sebuah tweet [9].

2.2.4 Stemming

Stemming merupakan proses untuk mendapatkan kata dasar dengan cara menghilangkan awalan, akhiran, sisipan, dan *confixes* (kombinasi dari awalan dan akhiran). Pada tahap ini akan menggunakan algoritma *Nazief and Adriani's Stemmer* [10].

2.2.5 Tokenizing

Pada tahapan ini dilakukan sebuah proses untuk memilah isi teks sehingga menjadi satuan kata. Dalam proses ini prinsipnya untuk memisahkan setiap kata yang menyusun suatu dokumen.

2.3 Term Frequency Inverse Document Frequency

TF – IDF merupakan sebuah metode pembobotan kata yang digunakan untuk mengekstraksi ciri dari suatu teks, terdapat dua hal dalam perhitungan nilai bobot yaitu *term frequency* (TF) dan *inverse document frequency* (IDF). Pada tahap TF digunakan untuk mencari nilai dari kemunculan kata dalam suatu dokumen. Lalu tahap IDF digunakan untuk mencari nilai kemunculan dari kata pada keseluruhan dokumen, nilai IDF berbanding terbalik dengan TF, semakin banyak kata yang muncul maka nilai IDF akan semakin kecil.bUntuk menghitung TF – IDF dari kata, digunakan rumus Persamaan 1.

$$TF - IDF(t, d, D) = tf(t, d)x\ idf(t, D) \tag{1}$$

2.4 Particle Swarm Optimization

Particle Swarm Optimization merupakan sebuah konsep berbasis intelejen segerombolan yang dikemukankan pada tahun 1995 oleh Eberhart dan Kennedy[11]. Metode tersebut digunakan untuk meningkatkan akurasi terhadap atribut yang terdapat pada metode *naïve bayes classifier* dengan menggunakan Persamaan 2 sebagai berikut [12].

$$V_i(t+1) = \omega V_{i(t)} + c_1 r_1\left(P_{i(t)} - X_{i(t)}\right) + c_2 r_2(P_g - X_{i(t)}) \tag{2}$$

2.5 Klasifikasi Naïve Bayes

Naïve Bayes Classifier adalah sebuah algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk klasifikasi data uji pada kategori yang paling tepat. Metode *naïve bayes* merupakan salah satu metode untuk proses klasifikasi teks. Kesederhanaan proses klasifikasi dalam metode ini menjadi salah satu keunggulan dari metode klasifikasi yang lainnya. Dalam metode ini dilakukan dua proses yaitu proses pelatihan data dengan membuat model klasifikasi, yang kedua proses pengujian dengan menggunakan data uji yang dimasukkan kedalam model klasifikasi yang telah dibuat menggunakan proses pelatihan data.

Pada proses klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan *(Vmap)*. Berikut Persamaan 3 yang digunakan dalam proses klasifikasi.

$$V_{map} = \underset{v_j=V}{\arg max} P(v_j) \prod_i P(a_i|V_j) \tag{3}$$

2.6 Pengujian

Validasi dilakukan menggunakan 10 *fold cross validation*. Dimana dengan menggunakan Teknik tersebut akan membagi secara acak ke dalam tiap bagian dimana terdiri dari 10 bagian yang akan dilakukan proses klasifikasi. Untuk pengukuran akurasi diukur dengan *confusion matrix* seperti pada Tabel 1 dengan menggunakan Persamaan 4, Persamaan 5, Persamaan 6, dan Persamaan 7.

Tabel 1. Confussion Matrix

	Positif	Negatif
Positif	True Positif	False Positif
Negatif	False Negatif	True Negatif

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Presisi = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F - measure = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (7)$$

3. Hasil Penelitian dan Pembahasan

Pembahasan pada penelitian kali ini terdiri dari 6 skenario. Skenario pertama dan kedua adalah klasifikasi menggunakan *naïve bayes classifier* dengan membedakan jumlah data uji dan data latih pada masing – masing skenario. Skenario ketiga klasifikasi menggunakan data yang telah melalui proses fitur seleksi menggunakan algoritma *particle swarm optimization* dengan iterasi sebanyak 50 kali. Pada skenario keempat masih menggunakan *particle swarm optimization* sebagai fitur seleksi dengan iterasi sebanyak 100 kali. Lalu pada skenario kelima dan keenam juga proses klasifikasi *naïve bayes classifier* masih menggunakan data yang telah melalui proses fitur seleksi menggunakan *particle swarm optimization* dengan iterasi sebanyak 150 kali dan iterasi sebanyak 200 kali pada masing – masing skenario. Semua skenario pengujian pada penelitian ini menggunakan 10 *cross validation*.

Pada skenario pertama menggunakan data uji sebanyak 40% dan data latih sebanyak 60%, skenario kedua menggunakan data uji sebanyak 70% dan data latih sebanyak 30%, pada skenario ketiga menggunakan data uji sebanyak 40% dan data latih sebanyak 60%, serta menggunakan iterasi sebanyak 50 kali. Skenario keempat menggunakan data uji sebanyak 70% dan data latih sebanyak 30%, serta menggunakan iterasi sebanyak 100 kali. Skenario pengujian kelima menggunakan data uji sebanyak 30% dan data latih sebanyak 70%, serta menggunakan iterasi sebanyak 150 kali. Skenario keenam menggunakan data uji sebanyak 90% dan data latih sebanyak 10%, serta menggunakan iterasi sebanyak 200 kali. Berikut hasil dari setiap skenario pengujian.

3.1 Klasifikasi menggunakan metode *naïve bayes classifier*

Pada Tabel 2, skenario pertama klasifikasi ini didapatkan nilai *true positif 258, false positif* 46*, true negatif* 62*, false negatif* 34. Hasil tersebut didapatkan dari total data uji sebanyak 400 data.

Tabel 2. Confussion Matrix

	Predicted Label	
	Positif	Negatif
Positif	258	46
Negatif	34	62

Berdasarkan hasil *confussion matrix* yang telah didapatkan pada proses klasifikasi skenario pertama, didapatkan hasil evaluasi data berdasarkan tabel *confussion matrix* tersebut. Tabel 3 berikut hasil akurasi, recall, presisi, dan f – measure.

Tabel 3. Evaluasi Klasifikasi

	Hasil
Accuracy	73.75%
Recall	75%
Precision	73%
F Measure	74%

Pada Tabel 4, skenario kedua klasifikasi ini didapatkan nilai *true positif 489, false positif* 10*, true negatif* 40*, false negatif* 70. Hasil tersebut didapatkan dari total data uji sebanyak 700 data.

Tabel 4. Confussion Matrix

	Predicted Label	
	Positif	Negatif
Positif	489	10
Negatif	70	40

Berdasarkan hasil *confussion matrix* yang telah didapatkan pada proses klasifikasi skenario pertama, didapatkan hasil evaluasi data berdasarkan tabel *confussion matrix* tersebut. Tabel 5 berikut hasil akurasi, recall, presisi, dan f – measure.

Tabel 5. Evaluasi Klasifikasi

	Hasil
Accuracy	74.14%
Recall	74%
Precision	76%
F Measure	74%

3.2 Klasifikasi metode naïve bayes classifier berbasis particle swarm optimization

Pada skenario ketiga klasifikasi ini didapatkan nilai *true positif 320, false positif* 80*, true negatif* 0*, false negatif* 0. Hasil Tabel 6 berikut didapatkan dari total data uji sebanyak 400 data.

Tabel 6. Confussion Matrix

	Predicted Label	
	Positif	Negatif
Positif	320	80
Negatif	0	0

Berdasarkan hasil *confussion matrix* yang telah didapatkan pada proses klasifikasi skenario pertama, didapatkan hasil evaluasi data berdasarkan tabel *confussion matrix* tersebut. Tabel 7 berikut hasil akurasi, recall, presisi, dan f – measure.

Tabel 7. Evaluasi Klasifikasi

	Hasil
Accuracy	80%
Recall	100%
Precision	80%
F Measure	89%

Pada skenario keempat klasifikasi ini didapatkan nilai *true positif 559, false positif* 141*, true negatif* 0*, false negatif* 0. Hasil Tabel 8 berikut didapatkan dari total data uji sebanyak 700 data.

Tabel 8. Confussion Matrix

	Predicted Label	
	Positif	Negatif
Positif	559	141
Negatif	0	0

Berdasarkan hasil *confussion matrix* yang telah didapatkan pada proses klasifikasi skenario pertama, didapatkan hasil evaluasi data berdasarkan tabel *confussion matrix* tersebut. Tabel 9 berikut hasil akurasi, recall, presisi, dan f – measure.

Tabel 9. Evaluasi Klasifikasi

	Hasil
Accuracy	79.85%
Recall	100%
Precision	80%
F Measure	89%

Pada Tabel 10, skenario kelima klasifikasi ini didapatkan nilai *true positif 234, false positif* 66*, true negatif* 0*, false negatif* 0. Hasil tersebut didapatkan dari total data uji sebanyak 300 data.

Tabel 10. Confussion Matrix

	Predicted Label	
	Positif	Negatif
Positif	234	66
Negatif	0	0

Berdasarkan hasil *confussion matrix* yang telah didapatkan pada proses klasifikasi skenario pertama, didapatkan hasil evaluasi data berdasarkan tabel *confussion matrix* tersebut. Tabel 11 berikut hasil akurasi, recall, presisi, dan f – measure.

Tabel 11. Evaluasi Klasifikasi

	Hasil
Accuracy	78%
Recall	100%
Precision	78%
F Measure	88%

Pada Tabel 12, skenario keenam klasifikasi ini didapatkan nilai *true positif 729, false positif* 171*, true negatif* 0*, false negatif* 0. Hasil tersebut didapatkan dari total data uji sebanyak 900 data.

Tabel 12. Confussion Matrix

	Predicted Label	
	Positif	Negatif
Positif	729	171
Negatif	0	0

Berdasarkan hasil *confussion matrix* yang telah didapatkan pada proses klasifikasi skenario pertama, didapatkan hasil evaluasi data berdasarkan tabel *confussion matrix* tersebut. Tabel 13 berikut hasil akurasi, recall, presisi, dan f – measure.

Tabel 13. Evaluasi Klasifikasi

	Hasil
Accuracy	81%
Recall	100%
Precision	81%
F Measure	90%

Dari hasil perhitungan diatas diketahui jika nilai rata – rata akurasi, presisi, recall, dan f – measure proses penelitian ini menggunakan metode *naïve bayes classifier* dibandingkan dengan klasifikasi *naïve bayes* berbasis *particle swarm optimization* didapatkan hasil paling baik dari klasifikasi ini adalah yang menggunakan *particle swarm optimization* sebagai fitur seleksi dibandingkan klasifikasi tanpa menggunakan fitur seleksi.

4. Kesimpulan

Hasil yang didapatkan dari penelitian sentiment analisis tweet berbahasa Indonesia pada pilkada serentak 2020 menggunakan *naïve bayes* berbasis *particle swarm optimization* dihasilkan nilai akurasi paling baik pada proses klasifikasi *naïve bayes* yang menggunakan *particle swarm optimization* sebagai fitur seleksi pada proses klasifikasi nya. Didapatkan hasil akurasi sebesar 81%, recall 100%, presisi 81%, f – measure 90%.

Dari proses tersebut bisa disimpulkan jika penggunaan *particle swarm optimization* sebagai fitur seleksi dapat meningkatkan nilai akurasi dari proses klasifikasi dengan metode *naïve bayes*.

Daftar Notasi

Keterangan Confussion Matrix

TP : True Positif
FP : False Positif
TN : True Negatif
FN : False Negatif

Referensi

[1] S. Harahap W, "Pro dan Kontra Pilkada Langsung," *Tagar.id*, 2019. [Online]. Available: https://www.tagar.id/pro-dan-kontra-pilkada-langsung.
[2] Grafelly Delvit, "Bagaimana perkembangan Twitter saat ini?," *Techno.id*, 2015. [Online]. Available: https://www.techno.id/social/bagaimana-perkembangan-twitter-saat-ini-1509122.html. [Accessed: 13-Sep-2015].
[3] A. Rossi, T. Lestari, R. Setya Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017.
[4] S. Saprudin, "Penerapan Particle Swarm Optimization (PSO) untuk Klasifikasi dan Analisis Kredit dengan Menggunakan Algoritma C4.5," *J. Inform. Univ. Pamulang*, vol. 2, no. 4, p. 214, 2017.
[5] R. Y. Hayuningtyas and R. Sari, "Analisis Sentimen Opini Publik Bahasa Indonesia Terhadap Wisata Tmii Menggunakan Naïve Bayes Dan Pso," *J. Techno Nusa Mandiri*, vol. 16, no. 1, pp. 37–42, 2019.
[6] J. Weng, J. Weng, E. Lim, and J. Jiang, "Institutional Knowledge at Singapore Management University Twitterrank : Finding topic-sensitive influential Twitterers TwitterRank : Finding Topic-sensitive Influential Twitterers," 2010.
[7] T. Kurniawan, "Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier Dan Support Vector Machine Media Mainstream Menggunakan Naïve Machine," p. 1, 2017.
[8] E. Dragut, F. Fang, P. Sistla, C. Yu, and W. Meng, "Stop word and related problems in web interface integration," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 349–360, 2009.
[9] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.
[10] D. Wahyudi, T. Susyanto, and D. Nugroho, "Implementasi Dan Analisis Algoritma Stemming Nazief & Adriani Dan Porter Pada Dokumen Berbahasa Indonesia," *J. Ilm. SINUS*, vol. 15, no. 2, 2017.
[11] Y. Shi, "Particle Swarm Optimization: Developments, Applications and Resources Russell," *Adv. Neural Inf. Process. Syst.*, pp. 329–336, 2007.
[12] K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," *Proc. 2nd Int. Conf. Intell. Comput. Control Syst. ICICCS 2018*, no. May, pp. 685–690, 2019.