ISBN: 978-979-792-691-5

PERBANDINGAN KINERJA ALGORITMA NAIVE BAYES DAN K-NN PENDEKATAN LEXICON PADA ANALISIS SENTIMEN DI MEDIA TWITTER

Azhar[1], Siti Ummi Masruroh[1], Luh Kesuma Wardhani[1], Okfalisa[2]

[1]Prodi Teknik Informatika, Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah
[2]Program Studi Teknik Informatika Fakultas Sains dan Teknologi UIN Sultan Syarif Kasim Riau

*E-mail korespondensi: azhar.roses14@mhs.uinjkt.ac.id

ABSTRACT

Sentiment analysis or opinion mining is a type of natural language, which processed words to find out opinions, attitudes or moods about certain things. Word processing in this study related to the process of classification in textual documents, which was classified into three classes, positive, negative and neutral. Data obtained from social media twitter which were related to the comments of netizens as many as 1000 comments. These data were crawled using keywords of the "Pilpres2019" and "Jokowi". In this study, the performance of the Naive Bayes and K-NN algorithms withLexicon approach were compared in doing classification. The aim of this study wasto compare the level of accuracy, precision and recall of Naive Bayes and K-NN algorithm with the Lexicon approach. From the evaluation, we concluded that the combination of the K-NN algorithm and the Lexicon approach can improve accuracy in this sentiment analysis case. Generally,K-NN algorithm with lexicon approach which k value is k = 5 has better performance with 77% of accuracy level, followed by Naïve Bayes with accuracy of 81% of accuracy level.

Keywords: Sentiment analysis, Classification, Naïve Bayes classifier, Lexicon, K-nearest neighbor

ABSTRAK

Analisis sentiment adalah jenis pemrosesan bahasa alami, yang memproses kata-kata untuk mengetahui pendapat, atau sikap tentang hal-hal tertentu. Analisis sentiment dalam penelitian ini terkait dengan proses klasifikasi pada dokumen tekstual, yang diklasifikasikan menjadi tiga kelas, positif, negatif dan netral. Data diperoleh dari media sosial Twitter, dimana data tersebut terkait dengan komentar netizen sebanyak 1000 komentar. Data ini diperoleh menggunakan teknik crawling dengankata kunci adalah "Pilpres 2019" dan "Jokowi". Dalam penelitian ini, kinerja algoritma Naive Bayes dan K-NN dengan pendekatan Lexicon dibandingkan dalam melakukan klasifikasi. Tujuan dari penelitian ini adalah untuk membandingkan tingkat akurasi, precision dan recall dari algoritma Naive Bayes dan K-NN dengan pendekatan Lexicon. Dari evaluasi, dapat disimpulkan bahwa kombinasi dari algoritma K-NN dan pendekatan Lexicon dapat meningkatkan akurasi dalam kasus analisis sentiment ini. Secara umum, algoritma K-NN yang dikombinasikan dengan pendekatan Lexicon (k = 5) memiliki kinerja yang lebih baik dengan akurasi sebesar77%, diikuti oleh Naïve Bayes denganakurasi 81% .

Kata kunci: Analisis sentimen, Klasifikasi, Naïve Bayes classifier, Lexicon, K-nearest neighbor

PENDAHULUAN

Media sosial seperti facebook, twitter, dan youtubemengalami peningkatan pengguna aktif. Dari data 2015 memaparkan jumlah pengguna twitter lebih dari 285 juta jiwa. Mayoritas penggunanya dari umur 16 tahun hingga 65 tahun keatas.

Media sosial twitter juga belakangan ini mengalami perubahan fungsi yang cukup drastis. Pada awal mula twitter berdiri media

ini digunakan sebagai alat berbalas pesan dan tempat berbagi, namun akibat popularitas yang cukup tinggi penggunaan fungsinya bertambah banyak, seperti pencarian trend,promosi produk, dan pemberian rating. Twitter pada waktu belakangan ini juga kerap kali digunakan dalam ranah sosial–politik misalnya seperti pergerakan sosial bersama, melaporkan informasi tentang kemacetan, politik, kondisi cuaca, bencana alam dan informasi tertentu seperti peringatan terhadap suatu kejadian. Akibatnyafungsi pada twitter yang awal mula sebagai tempat tanya jawab dan berbagi informasi menjadi sarana yang digunakan berbagai aspek salah satunya sebagai media penelitian berupa text mining, sentimen analisis, dan *artificial intelligence* [1].

Twitter memiliki peran penting sebagai penunjuk *trend* yang terjadi misalnya seperti pada masapemilihan calon presiden tahun 2019. Sejak tahun 2018 sudah banyak *trending topic* di twitter mengenai calon–calon presiden untuk tahun 2019, meskipun kabar tersebut belum pastikebenarannya.

Dengan tingginya minat warganet untukmenanggapi berita populer di Indonesia mengundang banyak penelitian berbasis *machine learning, text mining, artificial intelligence* untuk mencoba menggunakan data dari media twitter. Salah satu penelitiantentanganalisis sentiment menggunakan machine learning dilakukan tahun 2016 oleh Ghulam Asrofi dengan judul Analisis Sentimen Calon Gubernur DKI Jakarta 2017 dan Devika M. D. untuk penelitian perbandingan analisis kinerja pada analisis sentimen[2,3].

Naïve Bayes dan K-NN merupakan beberapa metode *machine learning* yang sering digunakan dalam klasifikasi teks dan memberikan hasil yang cukupbaik. Pada penelitian ini dilakukan perbandingan kinerja terhadap kedua metode tersebut dalam kasus analisis sentiment. K-NN yang digunakan sebagai salah satu metode klasifikasi dikombinasikan dengan pendekatan lexicon, dimana pada beberapa penelitian kombinasi ini memberikan hasil lebih baik dalam kinerja metode klasifikasi.

Data penelitian diperoleh dari dataset yang berasaldarimedia twitter sebanyak 1000 dokumen. Dokumen ini diperoleh menggunakan teknik crawling dengan kata kunci “pilpres 2019” dan “Jokowi”.

Paper ini dibagi kedalam beberapa pembahasan, yaitu pendahuluan yang menjelaskan motivasi penelitian ini. Bagian 2 mengulas tentang tinjauan pustaka yang mendasari penelitian. Selanjutny abagian 3 menjelaskan tentang metode penelitian yang digunakan dalam penelitian ini. Hasil dan pembahasan mengenai penelitian ini dapat dilihat pada bagian 4. Bagian akhir dari penelitian ini adalah kesimpulan dari penelitian.

TINJAUAN PUSTAKA

1. Pre-Processing

Pre-processing dalam proses klasifikasi dokumendigunakanuntukmembangunsebuah index dari koleksi dokumen. Index adalah himpunan *term* yang menunjukkan isi atau topik yang dikandung oleh dokumen [4]. Ekstraksi *term* biasanya melibatkan beberapa operasi utama, antara lain:

1. *Case Folding* merupakan tahap mengubah semua huruf menjadi bentuk huruf kecil (l*ower case*).
2. *Filtering* merupakan tahap menghapus tanda baca seperti “@,.!” Untuk menpermudah tahap berikutnya.
3. Pemisahan rangkaian *term* (*tokenization). Tokenization* adalah tugas memisahkan deretan kata di dalam kalimat, paragraph atau halaman menjadi token atau potongan kata tunggal atau termed word.
4. Normalisasi merupakan tahap pemeriksaan ulang apakah setiap token sudah sesuai ejaan yang ada di KBBI atau belum dan merubah term kata tidak baku menjadi kata sesuai di KBBI.
5. Penghapusan *stop words*. Stop word didefinisikan sebagai term yang tidak berhubungan (irrelevant) dengan subyek utama dari *database* meskipun kata

tersebut sering kali hadir di dalam dokumen. Contoh *stop words* adalah ada, adalah, adanya, adapun, agak.

6. *Stemming*. Kata-kata yang muncul di dalam dokumen sering mempunyai banyak varian. Karena itu, setiap kata yang bukan *stopwords* direduksi ke *stemmed word* (*term*) yang cocok yaitu kata tersebut distem untuk mendapatkan bentuk akarnya dengan menghilangkan awalan atau akhiran.

2. Naive Bayes

Metode Naïve bayes memanfaatkan probabilitas atau nilai kemungkinan. Konsepdasar yang digunakan oleh naïve bayes adalah teorema bayes, yaitu melakukan klasifikasi dengan melakukan perhitungan nilai probabilitas p(c), yaitu probabilitas kelas jika diketahui dokumen. Naive bayes menganggap sebuah dokumen sebagai kumpulan dari kata-kata yang menyusun dokumen tersebut, dan tidak memperhatikan urutan kemunculuan kata pada dokumen perhitungan probabilitasnya dapat dianggap sebagai hasil perkalian dari probabilitas kemunculan kata-kata pada dokumen[5].

Probabilitas sebuah dokumen berada di kelas c dapat dihitung menggunakan rumus posterior probability sebagai berikut:

$$P(c_j|w_i) = P(c_j)xP(w_1|c_j)x \dots xP(w_k|c_j) \quad (1)$$

Keterangan:

1. $P(C_j|W_i)$ = *Posterior Probability* adalah peluang kemunculan hasil dari prior dan likelihood untuk menentukan kategori sebuah kelas (dokumen).
2. P(Cj) = *Prior Probability* adalah peluang kemunculan tiap kelas
3. $P(W_i|C_j)$ = *Conditional Probability (Likelihood)* adalah peluang kata-kata pada kelas (dokumen) tertentu.

3. K-Nearest Neighbor

K-NN adalah algoritma untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples* (data latih). Dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas darikategori pada *K-NN*. Algoritma *K-NN* menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru[6]. Data latih akan dibangun denga nmemperhatikan keseimbangan dokumen satu sama lain. Adapun algoritma *K-NN* dapat dijelaskan dengan keterangan berikut[7]:

a. Hitung jarak antara data sampel (data uji) dengan data latih yang telahdibangun. Salah satu persamaan dalam menghitung jarak kedekatan dapat menggunakan persamaan *Cosine Similarity*.
b. Menentukan parameter nilai k = jumlah tetanggaan terdekat.
c. Mengurutkan jarak terkecil dari data sample
d. Pasangkan kategori sesuai dengan kesesuaian
e. Cari jumlah terbanyak daritetanggaan terdekat. Kemudian tetapkan kategori.

Jarak yang digunakan dalam penelitian ini adalah *Cosine Similarity.*

$$Cos(i,k) = \frac{\sum k\,(di.dk)}{\sqrt{\sum k\, d_{ik}^2}\sqrt{\sum k\, d_{jk}^2}} \quad (2)$$

Keterangan :

$\sum k\,(d1\,dk)$= vector produk dari i dan k

$\sqrt{\sum k\, d_{ik}^2}$ = panjang vektor i

$\sqrt{\sum k\, d_{jk}^2}$ = panjang vektor k

4. Metode Simulasi

Simulasi adalah suatu metodologi untuk melaksanakan percobaan dengan menggunakan model dari suatu sistem nyata. Ada berbagai jenis *lifecycle* menurut [8] yang dapat digunakan untuk studi pada pemodelan dan simulasi. Terdapat langkah-langkah dasar yang harus dipertimbangkan dalam melakukan studi simulasi. *Lifecycle* tidak harus diartikan sebagai sebuah sekuensial yang *strict*, bersifat

iteratif, dan kadang juga bertransisi dalam arah yang berlawanan [8].

METODE PENELITIAN

Pada penelitian ini penulis mengumpulkan data dan informasi yang dapat menunjang proses dalam penelitian dimana proses pengumpulan data sebagai berikut :

Studi Pustaka

Studi pustaka dilakukan dengan pengumpulan teori yang berkaitan dengan penulisan artikel ini sebagai bahan untuk melengkapi penelitian ini. Sumber teori berasal dari buku referensi,hasil penelitian (jurnal dan skripsi) dan artikel-artikel terkait. Selain itu peneliti juga mengunjungi situs-situs yang terkait aplikasi *natural language processing, text mining,* pendekatan *lexicon* dan algoritma klasifikasi tentang *Naive Bayes*dan *K-Nearest Neighbor* (*K-NN*).

Crawling Data

Data twitter diperoleh dengan metode crawling pada media twitter menggunakan Twitter API. Data yang diambil adalah tentang komentar netizen terhadapsalah satu calon pilpres 2019 Joko Widodo pada pemilihan calon pilpres di tahun 2019. *Crawling* data dimulai dari tanggal 08 Februari 2018 hingga 24 Februari2018. *Crawling* kedua dimulai tanggal 12 Juni sampai 08 Agustus 2018. Pengambilan data memanfaatkan fitur dari twitter yang diperuntukkan untuk *developer* yang terdapat pada *website* https://developers.twitter.com/.

HASIL DAN PEMBAHASAN

Beberapa penelitian sebelumnya [2-6] diperoleh informasi bahwa analisis sentimen dapat dilakukan dengan membandingkan algoritma *Naive Bayes* dan *K-NN* sebagai klasifikasi. Metode Lexicon digunakan untuk mempermudah pemberian label pada data latih sejumlah 900 data dan data uji sebesar 100 data.

1. *Simulation Phase*

Pada tahapan simulasi ini dilakukan beberapa percobaan simulasi yang berkaitan dengan pengujian tingkat akurasi algoritma *Naive Bayes* dan *K-NN*. Adapun faktor-faktor dalam proses simulasi dapat dilihat pada tabel 1 berikut.

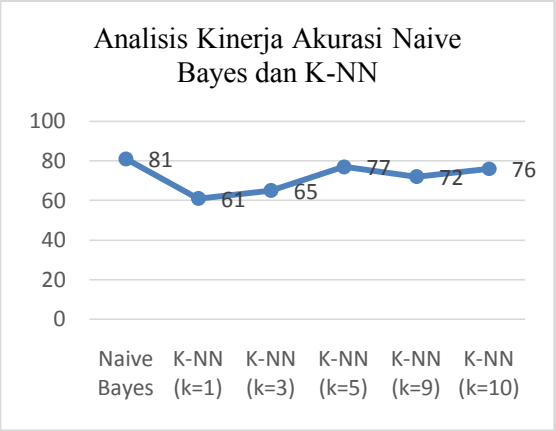
Tabel1. Tahapansimulasi

Variabel/Parameter Simulasi	Tahap Simulasi
Faktor 1	Tahap klasifikasi sentimen data latih dengan pendekatan *Lexicon*
Faktor 2	Tahap pelatihan data pada data latih
Faktor 3	Tahap pengujian data uji dengan algoritma *Naive Bayes* dan algoritma *K-NN* berdasarkan nilai k pada algoritma *K-NN* adalah 1,3,5,9 dan 10
Faktor 4	Tahap pengujian akurasi menggunakan model *confusion matrix*

2. *Verification, Validation and Experimentation*

Dalam proses validasi dilakukan pengujian kebenaran sistem yaitu dengan melakukan perbandingan hasil klasifikasi algoritma *naive bayes* dan *K-NN* dengan pendekatan *lexicon*yang dihitung secara manual dengan hasil pada aplikasi analisis orientasi sentimen sehingga menghasilkan keakuratan sistem.

Setelah percobaan dilakukan dalam enam kali hasil analisis kinerja akurasi kedua algoritma dapat dilihat pada gambar 1 berikut.



Gambar 1. Hasil analisis kinerja akurasi

Kinerja *naive bayes* dan *K-NN* kemudian dibandingkan dengan nilai k terbaik, untuk diketahui perbedaannya yaitu,

a. Hasil tingkat akurasialgoritma *naive bayes*

Tabel2. Tabel hasil klasifikasi *naive bayes*

Sentimen		Kelas hasil prediksi		
		Positif	Negatif	Netral
Kelas Sebenarnya	Positif	a=31	b=1	c=4
	Negatif	d= 5	e=18	f =3
	Netral	g=2	h=4	i=32

Berdasarkan hasil pengujian dari Table 2 dapat diambil nilai akurasi sebagai berikut :

$$= \frac{31 + 18 + 32}{31 + 1 + 4 + 5 + 18 + 3 + 2 + 4 + 32} x100$$

$$akurasi = \frac{81}{100} \times 100 = 81\%$$

b. Hasil Tingkat Akurasi Algoritma *K-NN* dengan nilai *k*= 5.

Tabel 3. Hasil klasifikasi *K-NN* dengan nilai k=5

Sentimen		Kelas hasil prediksi		
		Positif	Negatif	Netral
Kelas Sebenarnya	Positif	a=32	b=3	c=1
	Negatif	d=5	e=14	f=7
	Netral	g=3	h=4	i=31

Berdasarkan hasil pengujian dari Tabel 3 di atas dapat diambil nilai akurasi sebagai berikut :

$$= \frac{32 + 14 + 31}{32 + 3 + 1 + 5 + 14 + 7 + 3 + 4 + 31} x100$$

$$akurasi = \frac{77}{100} \times 100 = 77\%$$

KESIMPULAN

Kesimpulan dari penelitian ini adalah sebagai berikut:

1. Kombinasi algoritma klasifikasi supervised learning dengan pendekatan *lexicon* dapat diterapkan pada analisis sentimen kata.
2. Nilai k optimal dalam melakukan proses klasifikasi algoritma *K-NN* mencapai tingkat akurasi pada k=5 dengan tingkat akurasi mencapai 77% menggunakan dataset hasil *crawling data* twitter dengan kata kunci pilpres 2019 dan Jokowi.
3. Pada perbandingan algoritma*Naive Bayes* dengan *K-NN* selisih tingkat akurasi mencapai 4%. Berdasarkan pengujian yang telah dilakukan oleh peneliti, kombinasi algoritma *naive bayes* dan *K-NN* dengan pendekatan *lexicon* telah terbukti meningkatkan akurasi dalam mengklasifikasi sentimen.
4. Akurasi pada nilai k dalam algoritma *K-NN* mengalami pola naik sebelum mencapai nilai k terbaik lalu mengalami penurunan tingkat akurasi setelah nilai k terbaik.
5. Dengan penggunaan jumlah data latih yang sedikit, *naive bayes* memiliki sedikit kelebihan dibandingkan *K-NN* yang mungkin menyebabkan akurasi *naive bayes* sedikit diatas *K-NN*.

REFERENSI

1. Buntoro, G. A. (2017) “Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter,” 1(1), hal. 32–41.

2. D, D. M., Sunitha C dan Ganesh, A. (2016) "Sentiment Analysis : A Comparative Study On Different Approaches," *Procedia - Procedia Computer Science*. The Author(s), 87, hal. 44–49. doi: 10.1016/j.procs.2016.05.124.
3. Indriani, A. (2014) "Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier," hal. 5–10.
4. Yosyaningsih, Y. V. (2015) *Klasifikasi Dokumen Bahasa Jawa Menggunakan Metode Naïve Bayesian*. Skripsi Program Studi Teknik Informatika, Universitas Sanata Dharma, Yogyakarta.
5. Krisandi, N., Prihandono, B. dan Helmi (2013) "Algoritma K - Nearest Neighbor Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada PT. MINAMAS Kecamatan Parindu," *Buletin Ilmiah Math.Stat.dan Terapannya(Bimaster)*, 02(1), hal. 33–38.
6. Oktinas, W. (2017) "Analisis Sentimen Pada Acara TV Menggunakan Improved K-Nearest Neighbor."
7. Madani, S. A., Kazmi, J. dan Mahlknecht, S. (2010) "Wireless sensor networks: modeling and simulation," *Discrete Event Simulations*, hal. 1–16. Tersedia pada: http://www.intechopen.com/articles/show/title/wireless-sensor-networks-modelling-and-simulation.
8. Zhang H., Gan W, Jiang B. (2014). "Machine Learning and Lexicon for Sentiment Classification : A Survey." 2014 11th Web Information System and Application Conference. 12-14 Sept. 2014, Tianjin, China. DOI: 10.1109/WISA.2014.55
9. Dey L., Chakraborty S., Biswas A., Tiari S. (2016). "Sentiment Analyst of Review Datasets Using Naive Bayes and K-NN Classifier." Inte*rnational Journal of Information Engineering and Electronic Business* 8(4):54-62

Artikel ini menggunakan lisensi Creative Commons Attribution 4.0 International License