Perbandingan Algoritma Naïve Bayes, K-Nearest Neighbors dan Random Forest untuk Klasifikasi Sentimen Terhadap BPJS Kesehatan pada Media Twitter

by Tam Rizal

Submission date: 27-Oct-2021 10:12AM (UTC-0400)
Submission ID: 1685601205
File name: Cek_turnitin.docx (276.37K)
Word count: 1518
Character count: 10182

Perbandingan Algoritma Naïve Bayes, K-Nearest Neighbors dan Random Forest untuk Klasifikasi Sentimen Terhadap BPJS Kesehatan pada Media Twitter

Tamrizal A.M[1*], Ainul Yaqin [2]

[1]*Magister Teknik Informatika, Universitas Amikom Yogyakarta,*
Jl. Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta, Indonesia 55283
[2] *Informatics. Faculty of Computer Science, Universitas Amikom Yogyakarta,*
Jl. Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta, Indonesia 55283
tamrizal@gmail.com

***Abstrak* :**
Dari sejak didirikan, BPJS terus berusaha meningkatkan kualitas pelayanan termasuk menyediakan berbagai layanan pengaduan. Selain fasilitas pengaduan yang telah disediakan oleh BPJS, media sosial seperti twitter sebenarnya dapat dijadikan sebagai tempat untuk mengumpulkan informasi yang berkaitan dengan BPJS. Berbagai keluhan maupun apresiasi terhadap pelayanan BPJS sering disuarakan melalui media twitter. Pada penelitian ini, dilakukan pengujian tiga algoritma machine learning yaitu Naïve Bayes, K-Nearest Neighbors dan Random Forest, untuk mengetahui dan membandingkan tingkat akurasi dari masing-masing algoritma tersebut dalam melakukan klasifikasi terhadap sentimen masyarakat terhadap BPJS Kesehatan melalui media twitter. Pada penelitian ini dataset diperoleh dengan melakukan scrapping menggunakan twitter API. Data yang diperoleh kemudian diseleksi dan dilakukan labeling. Dari hasil seleksi dan labeling didapatkan dataset sebanyak 150 tweet yang terdiri atas 50 tweet positif, 50 tweet negative dan 50 tweet netral yang akan digunakan dalam percobaan. Pada percobaan dengan menggunakan 90% data untuk training dan 10% data untuk testing, didapatkan tingkat akurasi sebesar 80% Naive Bayes, 67% K-Nearest Neighbors dan 87% Random Forest.

***Keywords*:**
Sentimen Analisis;
Naïve Bayes;
KNN;
Random Forest l;
Twitter;
BPJS

Article history:
Received ……., 20xx
Revised ……, 20xx
Accepted ……., 20xx

DOI:
10.22441/incomtech.v10i3.7777

1. PENDAHULUAN

Badan Penyelenggara Jaminan Sosial (BPJS) telah resmi beroperasi sejak 1 Januari 2014 sebagai tindak lanjut dari ditetapkannya undang-undang nomor 24 tahun 2011 tentang Badan Penyelenggara Jaminan Sosial (BPJS)[1][2][3][5]. Kehadiran BPJS dimaksudkan untuk memastikan seluruh penduduk Indonesia terlindungi oleh jaminan kesehatan yang komprehensif, adil, dan merata.

Sejak berdirinya sampai dengan sekarang BPJS terus berusaha meningkatkan kualitas pelayanan termasuk menyediakan berbagai layanan pengaduan melalui unit penangan pengaduan peserta yang ada di kantor BPJS, petugas informasi dan pengaduan peserta (PIPP) yang ada di rumah sakit, BPJS Kesehatan Care Center di nomor 1500400 serta menu lapor pada aplikasi mobile JKN. Melalui berbagai fasilitas layanan pengaduan tersebut diharapkan dapat memberikan masukan untuk perbaikan layanan BPJS Kesehatan dan stakeholder.

Selain fasilitas pengaduan yang telah disediakan oleh BPJS tersebut, media sosial seperti twitter sebenarnya dapat dijadikan sebagai tempat untuk mengumpulkan informasi yang berkaitan dengan BPJS. Berbagai keluhan maupun apresiasi terhadap pelayanan BPJS sering disuarakan melalui media twitter. Hal tersebut sesungguhkan dapat dijadikan peluang untuk mengumpulkan informasi sebanyak-banyaknya untuk menjadi bahan evaluasi bagi BPJS Kesehatan.

Dalam mengelola informasi yang didapatkan dari twitter, salah satu kendala yang ditemui adalah melakukan klasifikasi terhadap informasi yang didapatkan, karena tidak semua informasi yang didapatkan dari twitter dapat digunakan sebagai masukan. Terkadang informasi yang didapatkan hanya berupa candaan yang kebetulan menggunakan kata BPJS.

Salah satu metode yang dapat digunakan untuk melakukan klasifikasi adalah menggunakan algoritma machine learning, berbagai algoritma machine learning dapat digunakan dalam membantu melakukan klasifikasi tetapi dengan tingkat akurasi yang berbeda-beda. Dalam penelitian ini, penulis ingin melakukan pengujian tiga algoritma machine learning yaitu Naïve Bayes, K-Nearest Neighbors dan Random Forest, untuk membandingkan tingkat akurasi dari masing-masing algoritma tersebut dalam melakukan klasifikasi terhadap sentimen masyarakat terhadap BPJS Kesehatan melalui media twitter.

Naïve Bayes

Algoritma Naïve Bayes merupakan metode klasifikasi yang mengacu pada teorema bayes yang pertama kali dikemukakan oleh Thomas Bayes seorang ilmuwan yang berasal dari inggris. Metode ini menggunakan metode klasifikasi statistik. Ciri utama dari metode ini adalah asumsi akan independensi dari masing-masing kondisi[1][4][12][13][15][16]. Teorema bayes dapat digambarkan dalam model matematika sebagai berikut.

$$P(H|X) = \frac{P(X|H)}{\sum_{i=1}^{n} P(H_i|X)} . P(H)$$

K-Nearest Neighbors

Cara kerja Algoritma KNN adalah melakukan klasifikasi data dengan mengambil sejumlah data terdekat sebagai acuan untuk menentukan kelas data baru. Algoritma ini mengklasifikasikan data berdasarkan kedekatannya terhadap data lainnya.

Untuk menghitung jarak antara dua titik, algoritma KNN menggunakan metode Euclidean Distance[8][10][12][14]. Metode Euclidean distance menggunakan formula berikut.

$$dis = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + \cdots}$$

Random Forest

Random Forest adalah salah satu algoritma yang dipakai untuk melakukan klasifikasi data. Klasifikasi random forest dilakukan dengan menggabungkan pohon (tree) melalui proses training pada sampel data yang dimiliki. Jumlah pohon (tree) yang digunakan akan mempengaruhi tingkat akurasi yang akan didapatkan. Penentuan klasifikasi dengan random forest diambil berdasarkan hasil voting dari tree yang terbentuk[7][9][11].

2. METODE

Langkah-langkah yang dilakukan dalam melakukan penelitian ini adalah sebagai berikut :

1. Pengumpulan Data

   Pengumpulan data dilakukan dengan teknik scrapping data pada Twitter dengan menggunakan Twitter API. Scrapping dilakukan dengan menggunakan kata kunci “BPJS”.

2. Seleksi dan Pelabelan Data

   Data hasil scrapping kemudian dianalisis untuk melakukan seleksi agar data yang digunakan adalah tweet yang menggunakan bahasa Indonesia dan bukan merupakan data berulang dari hasil retweet.

   Selanjutnya dilakukan pelabelan data secara manual yang terdiri dari Positif, Negatif dan Netral. Setelah proses seleksi dan pelabelan data dilakukan, didapatkan dataset sebanyak 150 data tweet yang terdiri dari 50 data positif, 50 data negatif dan 50 data netral.

3. Preprocessing

   Prepocessing merupakan sebuah proses transformasi data agar dapat lebih mudah dipahami[1], data yang telah melalui proses preprocessing akan lebih mudah untuk diolah dalam pengujian.

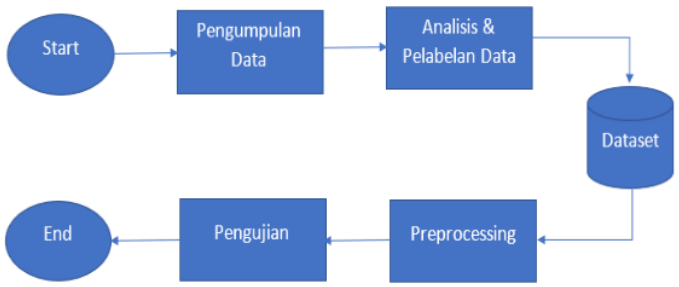
4. Term Weighting

Term Weighting atau pembobotan kata merupakan proses menghitung bobot setiap kata pada sebuah dokumen agar diketahui kemiripan antar kata pada dokumen tersebut[6]. Pada penelitian ini, term weighting dilakukan dengan menghitung Term Frequency-Inverse Document Frequency (TF-IDF).

5. Pengujian

Pengujian dilakukan dengan menggunakan algoritma Naive Bayes, K-Nearest Neighbors dan Random Forest. Masing-masing algoritma akan dilakukan pengujian sebanyak 3 kali dengan ketentuan pengujian pertama menggunakan 50% sebagai data training dan 50% sebagai data testing, pengujian kedua menggunakan 75% sebagai data training dan 25% sebagai data testing serta pengujian ketiga menggunakan 90% sebagai data training dan 10% sebagai data testing.

Langkah-langkah penelitian yang dilakukan dapat digambarkan dalam gambar berikut.



Gambar 1. Alur Penelitian

Penelitian dilakukan menggunakan bahasa pemrogaman phyton 3 pada platform anaconda.

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan merupakan data hasil scrapping yang telah diseleksi dan dilakukan pelabelan data. Kemudian dilakukan preprocessing dengan langkah-langkah.

1. Stopword Removal

Merupakan langkah untuk melakukan filter terhadap dataset untuk mengurangi penggunaan kata yang mempunyai frekuensi kemunculan yang tinggi.

2. Punctuation Filter

Punctuation filter merupakan proses untuk menghilangkan tanda baca ataupun simbol yang terdapat pada dataset. Tanda baca atau symbol dihapus karena tidak mempengaruhi hasil sentimen analisis.

3. Stemming

Merupakan proses untuk mengembalikan kata yang digunakan ke kata dasarnya[4][5]. Proses stemming pada penelitian ini menggunakan library sastrawi.

Setelah melakukan preprocessing dengan melakukan stopword removal, punctuation filter serta stemming, dilakukan pembobotan kata dengan melakukan perhitungan TF-IDF. Secara umum model matematika yang digunakan untuk melakukan perhitungan TF-IDF adalah sebagai berikut :

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D).$$

Pada penelitian ini perhitungan TF-IDF dilakukan dengan memanfaatkan fungsi Tfidfvektorizer yang telah disediakan oleh library scikit learn.

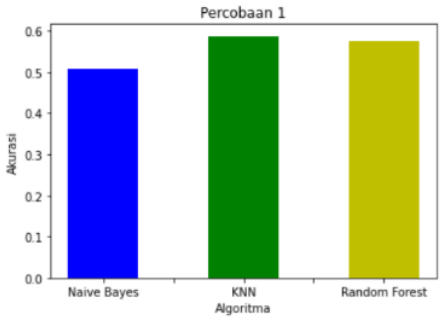
Setelah nilai TF-IDF didapatkan, dilakukan percobaan untuk mengatahui tingkat akurasi dari 3 algoritma yang digunakan, serta membandingkannya. Proses ujicoba dilakukan dengan menggunakan modul Naive Bayes, KNN dan Random Forest yang telah disediakan oleh library scikit learn.

1. Percobaan Pertama

Pada percobaan pertama dataset yang digunakan dibagi atas 50% sebagai data training dan 50% sebagai data testing. Hasil dari ujicoba pertama dapat dilihat pada tabel dibawah ini

Tabel 1. Hasil Percobaan Pertama

Algoritma	Akurasi
Naïve Bayes	51%
K-Nearest Neighbors	59%
Random Forest	57%



Gambar 2. Grafik Percobaan 1

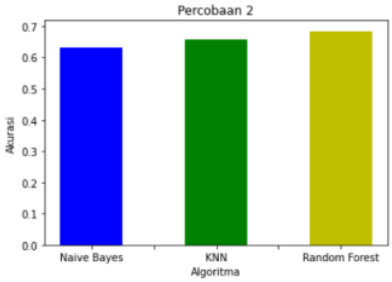
Pada percobaan pertama tingkat akurasi tertinggi didapatkan pada algoritma KNN yang mencapai 59% sedangkan tingkat akurasi paling rendah pada algoritma naive bayes sebesar 51%

2. Percobaan Kedua

Pada percobaan kedua dataset yang digunakan dibagi atas 75% sebagai data training dan 25% sebagai data testing. Hasil dari ujicoba kedua dapat dilihat pada tabel dibawah ini.

Tabel 2. Hasil Percobaan Kedua

Algoritma	Akurasi
Naïve Bayes	63%
K-Nearest Neighbors	66%
Random Forest	68%



Gambar 3. Grafik Percobaan 2

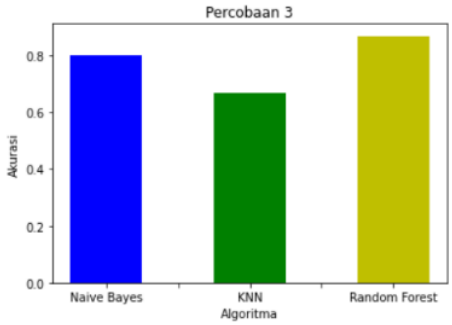
Pada percobaan kedua tingkat akurasi tertinggi didapatkan pada algoritma random forest yang mencapai 68% sedangkan tingkat akurasi paling rendah pada algoritma naive bayes sebesar 63%

3. Percobaan Ketiga

Pada percobaan ketiga dataset yang digunakan dibagi atas 90% sebagai data training dan 10% sebagai data testing. Hasil dari ujicoba ketiga dapat dilihat pada tabel dibawah ini.

Tabel 3. Hasil Percobaan Ketiga

Algoritma	Akurasi
Naïve Bayes	80%
K-Nearest Neighbors	67%
Random Forest	87%



Gambar 4. Grafik Percobaan 3

Pada percobaan ketiga tingkat akurasi tertinggi didapatkan pada algoritma random forest yang mencapai 87% sedangkan tingkat akurasi paling rendah pada algoritma naive bayes sebesar 67%.

Dari ketiga ujicoba yang dilakukan, secara umum peningkatan jumlah data training dapat meningkatkan akurasi dari algoritma yang diujicoba. Pengecualian pada algoritma KNN, percobaan kedua dan ketiga tingkat akurasinya relatif tidak mengalami peningkatan.

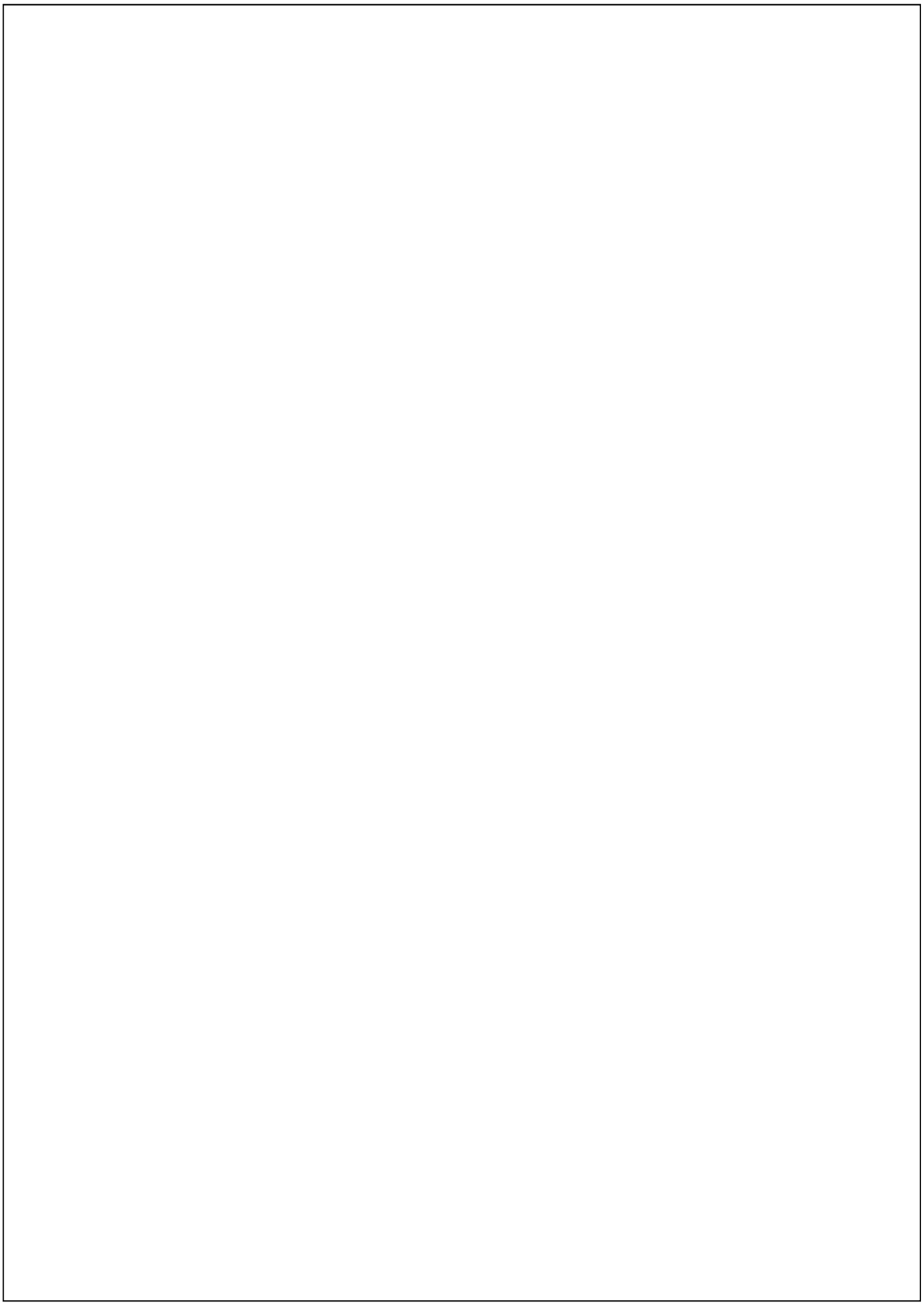
Hasil percobaan menunjukkan algoritma KNN mempunyai tingkat akurasi tertinggi pada percobaan pertama sedangkan random forest mempunyai tingkat akurasi tertinggi pada percobaan kedua dan ketiga.

Dari tiga kali percobaan yang dilakukan, perbedaan tingkat akurasi dari ketiga algoritma relatif tidak terlalu jauh, kecuali algoritma KNN pada percobaan ketiga memiliki selisih tingkat akurasi mencapai 13% dibandingkan algoritma naïve bayes dan selisih akurasi mencapai 20% dibandingkan dengan algoritma random forest.

4. KESIMPULAN

Dari tiga kali percobaan yang dilakukan algoritma random forest menunjukkan tingkat akurasi terbaik pada percobaan kedua sebesar 68% dengan data training 75% dari dataset dan percobaan ketiga sebesar 87% dengan data training 90% dari dataset. Sehingga algoritma tersebut dapat dipertimbangkan untuk digunakan dalam melakukan klasifikasi sentimen masyarakat terhadap BPJS Kesehatan pada media twitter.

Selanjutnya pengujian dapat dilakukan dengan menggunakan dataset yang lebih besar untuk menilai efektifitas algoritma tersebut atau dengan menggunakan algoritma lain sebagai pembanding tingkat akurasi.



Perbandingan Algoritma Naïve Bayes, K-Nearest Neighbors dan Random Forest untuk Klasifikasi Sentimen Terhadap BPJS Kesehatan pada Media Twitter

ORIGINALITY REPORT

19%	18%	5%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	id.wikipedia.org Internet Source	2%
2	pt.scribd.com Internet Source	2%
3	e-journal.janabadra.ac.id Internet Source	1%
4	es.scribd.com Internet Source	1%
5	simki.unpkediri.ac.id Internet Source	1%
6	www.slideshare.net Internet Source	1%
7	AM Tamrizal, Ainul Yaqin. "Decision Support System of Stock Selection Using Promethee Method", 2021 4th International Conference on Information and Communications Technology (ICOIACT), 2021 Publication	1%

8	nanopdf.com Internet Source	1 %
9	core.ac.uk Internet Source	1 %
10	learningbox.coffeecup.com Internet Source	1 %
11	123dok.com Internet Source	1 %
12	eprints.uns.ac.id Internet Source	1 %
13	lsisi.id Internet Source	1 %
14	repository.its.ac.id Internet Source	1 %
15	www.jamsosindonesia.com Internet Source	1 %
16	www.scribd.com Internet Source	1 %
17	Submitted to Universitas Trunojoyo Student Paper	1 %
18	a-research.upi.edu Internet Source	1 %
19	repository.uinjkt.ac.id Internet Source	1 %

20	Andreyestha Andreyestha, Agus Subekti. "ANALISA SENTIMENT PADA ULASAN FILM DENGAN OPTIMASI ENSEMBLE LEARNING", Jurnal Informatika, 2020 Publication	<1%
21	doku.pub Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		