TESIS

ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN PELAKSANAAN PEGAWAI PEMERINTAH DENGAN PERJANJIAN KERJA (P3K) GURU DENGAN ALGORITMA NAIVE BAYES, DECISION TREE DAN K-NN

Disusun oleh:

Nama : **Fitriani**
NIM : **21.55.1036**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2022

TESIS

ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN PELAKSANAAN PEGAWAI PEMERINTAH DENGAN PERJANJIAN KERJA P3K GURU DENGAN ALGORITMA NAIVE BAYES, DECISION TREE DAN K-NN

THE ANALYSIS OF COMMUNITY SENTIMENT ON THE IMPLEMENTATION POLICY OF GOVERMENT EMPLOYEES WITH WORK AGREEMENT (P3K) TEACHERS WITH NAIVE BAYES ALGORITHM, DECISION TREE AND K-NN

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama	**: Fitriani**
NIM	**: 21.55.1036**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2022

HALAMAN PENGESAHAN

ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN PELAKSANAAN PEGAWAI PEMERINTAH DENGAN PERJANJIAN KERJA (P3K) GURU DENGAN ALGORITMA NAIVE BAYES, DECISION TREE DAN K-NN

THE ANALYSIS OF COMMUNITY SENTIMENT ON THE IMPLEMENTATION POLICY OF GOVERMENT EMPLOYEES WITH WORK AGREEMENT (P3K) TEACHERS WITH NAIVE BAYES ALGORITHM, DECISION TREE AND K-NN

Dipersiapkan dan Disusun oleh

Fitriani

21.55.1036

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Rabu, 07 Desember 2022

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 07 Dwesember 2022
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN PELAKSANAAN PEGAWAI PEMERINTAH DENGAN PERJANJIAN KERJA (P3K) GURU DENGAN ALGORITMA NAIVE BAYES, DECISION TREE DAN K-NN

THE ANALYSIS OF COMMUNITY SENTIMENT ON THE IMPLEMENTATION POLICY OF GOVERMENT EMPLOYEES WITH WORK AGREEMENT (P3K) TEACHERS WITH NAIVE BAYES ALGORITHM, DECISION TREE AND K-NN

Dipersiapkan dan Disusun oleh

Fitriani

21.55.1036

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Rabu, 07 Desember 2022

Pembimbing Utama	**Anggota Tim Penguji**
Prof. Dr. Ema Utami, S.Si., M.Kom, Ph.D	**Alva Hendi Muhammad, S.T., M.Eng.**
NIK. 190302037	**NIK. 190302493**
	Dhani Ariatmanto, M.Kom., Ph.D.
Pembimbing Pendamping	**NIK. 190302197**
Anggit Dwi Hartanto, M.Kom.	**Dr. Kusrini, M.Kom.**
NIK. 190302163	**NIK. 190302106**

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, Rabu, 07 Desember 2022
Direktur Program Pascasarjana

Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Fitriani
NIM : 21.55.1036
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN PELAKSANAAN PEGAWAI PEMERINTAH DENGAN PERJANJIAN KERJA (P3K) GURU DENGAN ALGORITMA NAIVE BAYES, DECISION TREE DAN K-NN

Dosen Pembimbing Utama : **Prof. Dr. Ema Utami, S.Si., M.Kom**
Dosen Pembimbing Pendamping : **Anggit Dwi Hartanto, M.Kom**

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 07 Desember 2022
Yang Menyatakan,

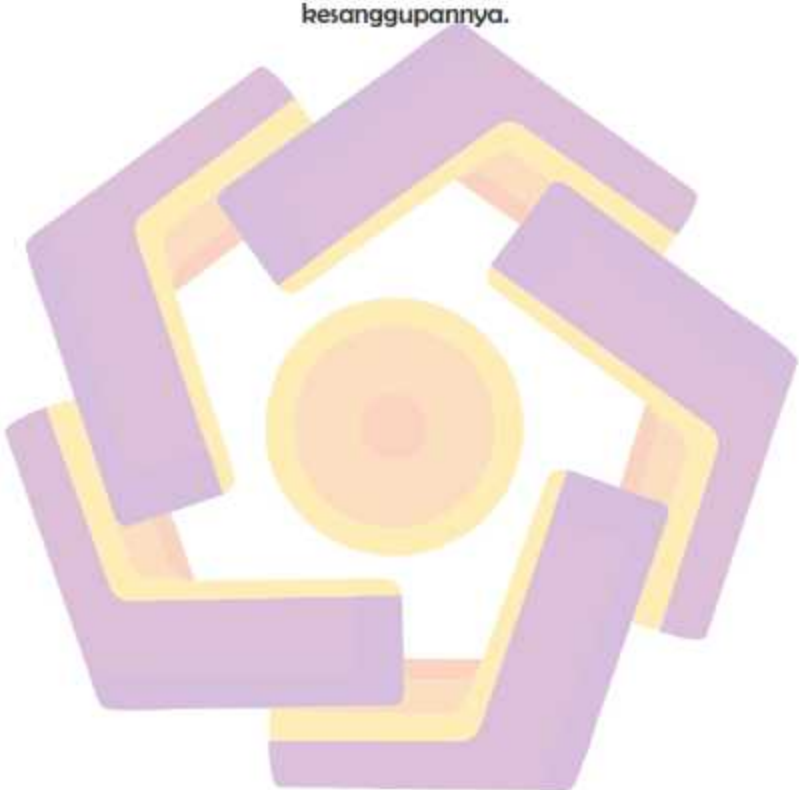
Fitriani

HALAMAN PERSEMBAHAN

Kedua orang tuaku, suami dan anak-anakku, Semoga ridho dan do'amu menghantarkanku meraih gelar Magister ini. Semoga apa yang kupersembahkan ini dapat membanggakan kalian semua.

HALAMAN MOTTO

Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya.



KATA PENGANTAR

Puji syukur Alhamdulillah kehadirat Allah Swt, atas segala rahmat-NYA sehingga penulis dapat menyelesaikan Tesis yang berjudul "**ANALISIS SENTIMEN MASYARAKAT TERHADAP PELAKSANAAN P3K GURU DENGAN ALGORITMA NAIVE BAYES, DECISION TREE DAN K-NN**".

Laporan Tesis ini disusun untuk memenuhi salah satu syarat akademis dalam menyelesaikan Program Studi S2 Teknik Informatika Program Pascasarjana Universitas AMIKOM Yogyakarta.

Pada kesempatan ini, penulis menyampaikan terima kasih kepada :

1. Allah SWT, yang telah memberikan semua kebutuhan sehingga terselesaikannya laporan Tesis ini.
2. Kedua orang tuaku dan suamiku atas semua dukungan, doa dan kesabarannya sehingga penulis dapat menyelesaikan Tesis ini.
3. Bapak TGH. LL.G. Muhammad Khairul Fatihin,S.Kom.I., M.M. dan Bapak Marzuki Adami , Selaku ketua dan puket 1 dan semua civitas akademik STMIK Syaikh Zainuddin NW Anjani yang telah membiayai dan memberikan dorongan dan inspirasi agar penulis mantap mengambil kuliah S2 ini serta dorongan agar segera menyelesaikan laporan Tesis ini.
4. Sahabat sekaligus adikku Hannatul Ma'wa yang telah memberikan dukungan atas terselesaikannya tesis ini.

5. Prof. Dr. Ema Utami, S.Si., M.Kom, Professor inspirator saya, yang memberikan wawasan dan dukungan atas terselesaikan Tesis ini. Semoga sehat selalu Prof.
6. Bapak Anggit Dwi Hartanto, M.Kom, yang telah memberikan ide judul Tesis serta dukungan atas terselesaikannya tugas ini.
7. Semua civitas akademik STMIK Syaikh Zainuddin NW Anjani.

Menyadari bahwa dalam laporan Tesis ini masih terdapat banyak sekali kekurangan, penulis sangat mengharapkan saran dan kritik yang membangun dari semua pihak yang akan sangat berguna untuk perbaikan dan penyempurnaan laporan Tesis ini.

Yogyakarta, Rabu, 07 Desember 2022

Penulis

DAFTAR ISI

HALAMAN JUDUL ii
HALAMAN PENGESAHAN iii
HALAMAN PERSETUJUAN iv
HALAMAN PERNYATAAN KEASLIAN TESIS viii
HALAMAN PERSEMBAHAN ix
HALAMAN MOTTO x
KATA PENGANTAR xii
DAFTAR ISI x
DAFTAR TABEL xiii
DAFTAR GAMBAR xiv
INTISARI xvi
ABSTRACT xvii
BAB I PENDAHULUAN 1
1.1.Latar Belakang Masalah 1
1.2. Rumusan Masalah 5
1.3. Batasan Masalah 6
1.4. Tujuan Penelitian 6
1.5. Manfaat Penelitian 6
BAB II TINJAUAN PUSTAKA 8
2.1. Tinjauan Pustaka 8
2.2. Keaslian Penelitian 12

2.3. Landasan Teori 22
2.3.1. PPPK Guru 22
2.3.2. Natural Language Processing 23
2.3.3. Sentiment Analisis 24
2.3.4. Twitter 25
2.3.5. Data Mining 26
2.3.6. Klasifikasi 27
2.3.7. Naive Bayes 27
2.3.8. KNN 28
2.3.9. Decision Tree 30
2.3.10. Text Processing 30
2.3.11. Cross Validation 31
2.3.12. Confusion Matrik 31
2.3.13. Rafidminer 32
BAB III METODE PENELITIAN 34
3.1. Jenis, Sifat dan Pendekatan Penelitian 34
3.2. Metode Pengumpulan Data 35
3.2.1. Crawling 35
3.2.2. Twitter API 37
3.3. Metode Analisis Data 39
3.4. Alur Penelitian 40
3.4.1. Crawling Data Twetter 40
3.4.2. Analisis Sumber Data 41

3.4.3. Preprocessing Data 41
3.4.4. Naive Bayes Decision Tree Dan KNN 46
3.4.5. Cross Validation Dan Confusion Matrix 48
BAB IV HASIL DAN PEMBAHASAN 51
4.1. Crawling data Twet 51
4.2. Analisa Data 52
4.2.1. Penghapusan Data Kembar dan Pemilihan Data Yang Relevan 52
4.2.2. Cleansing 52
4.2.3. Labeling 53
4.3. Preprocessing Data 53
4.4. Klasifikasi dan Akurasi 55
4.4.1. Klasifikasi Pada Naive Bayes 55
4.4.2. Klasifikasi Pada Decision Tree 62
4.4.3. Klasifikasi Pada KNN 66
BAB V PENUTUP 67
5.1. Kesimpulan 67
5.2. Saran 68
DAFTAR PUSTAKA
LAMPIRAN

DAFTAR TABEL

Tabel 3.1. Analisis Sumber Data 42
Tabel 3.2. Pelabelan Data Tweet 45
Tabel 4.1. Contoh Dokumen Yang Sudah Dilabeli 58
Tabel 4.2. Hasil akurasi Naïve Bayes 59
Tabel 4.3. Perhitungan Manual Naive Bayes 61
Tabel 4.4. Hasil Prediksi Manual Naive Bayes 62
Tabel 4.5. Contoh Dokumen Yang Sudah Dilabeli 63
Tabel 4.6. Hasil Akurasi Decision Tree 66
Tabel 4.7. Hasil Akurasi KNN 69
Tabel 4.8. Hasil Manual KNN 70
Tabel 4.9. Hasil Perhitungaan Manual KNN dengan $K=2$ 70
Tabel 4.10. Perbandingan Naive Bayes, Decision Tree dan KNN 71

DAFTAR GAMBAR

Gambar 2.1. Rumus Theorema Bayes 28
Gambar 2.2. Ilustrasi Kedekatan Kasus 29
Gambar 2.3. Rumus KNN 30
Gambar 2.4. Cara Kerja Cross Validation 33
Gambar 2.5. Confusion Matrik 34
Gambar 3.1. Skema Web Crawling 37
Gambar 3.2. Alur Penelitian 41
Gambar 3.3. Tahapan Preprocessing 44
Gambar 3.4. Contoh Penerapan Cleansing 44
Gambar 3.5. Hasil Tokenizing 46
Gambar 3.6. Contoh Case Folding 47
Gambar 3.7. Contoh Stopword Removal 47
Gambar 3.8. Contoh Token Filtering 48
Gambar 3.9. Rumus NBC 49
Gambar 3.10. Rumus Kedekatan Pada KNN 50
Gambar 3.11. Confusion Matrik 52
Gambar 4.1. Hasil Crawling Dari Repidminer 53
Gambar 4.2. Remove Duplicate dan Select Attribut 54
Gambar 4.3. Hasil Proses Crawling, Cleansing, Labeling 55
Gambar 4.4. Nominal To Text 58
Gambar 4.5. Proses Document From Data 57

Gambar 4.6. Cross Validation 59
Gambar 4.7. Cross Validation Dengan Naive Bayes 59
Gambar 4.8. Prediksi Naive Bayes 61
Gambar 4.9. Cross Validation 64
Gambar 4.10. Cross Validation Dengan Decision Tree 64
Gambar 4.11. Hasil Prediksi Decision Tree 66
Gambar 4.12. Cross Validation 67
Gambar 4.13. Cross Validation Dengan KNN 67
Gambar 4.14. Grafik Hasil Prediksi KNN 69
Gambar 4.15. Hasil Naive Bayes, Decision Tree dan KNN 71

INTISARI

Penyelenggaraan pendidikan di Indonesia hingga saat ini masih belum lepas dari persoalan tata kelola guru, guru honorer, dan reformasi birokrasi yang berpengaruh terhadap kualitas pendidikan dan iklim kerja di dalamnya. Dalam upaya peningkatan kualitas layanan publik oleh Aparatur Sipil Negara (ASN), Kementerian Pendidikan dan Kebudayaan sepakat dengan Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi dan Kementerian Keuangan untuk mengubah sistem rekrutmen guru pegawai pemerintah dari penerimaan Calon Pegawai Negeri Sipil (CPNS) menjadi Pegawai Pemerintah dengan Perjanjian Kerja (PPPK) yang dalam pelaksanaannya masih menyisakan beberapa masalah dan pro kontra. Oleh karena itu peneliti melakukan analisis sentimen dalam bidang data mining terhadap Pelaksanaan PPPK Guru pada media sosial Twitter sebanyak 871 data yang kemudian dilakukan *preprocessing* data menjadi 519. Penulis menggunakan metode *Naive bayes, Decision Tree* dan KNN untuk mengatahui hasil prediksi algoritma naive bayes, decision tree dan KNN mengenai opini masyarakat terhadap pelaksanaan PPPK Guru serta memperbandingkan tingkat akurasi dari ketiga metode tersebut. Peneliti menggunakan *tools RapidMiner* versi 9.10.1. Hasil prediksi naive bayes yaitu 328 data bersentimen positif dan 191 data bersentimen negatif, Selanjutnya hasil prediksi decision tree yaitu 165 data bersentimen positif dan 354 data bersentimen negatif dan yang terakhir yaitu hasil prediksi dari KNN yaitu 315 data bersentimen positif dan 204 data bersentimen negatif. Analisis sentimen masyarakat terhadap pelaksanaan PPPK guru pada media sosial Twitter dengan algoritma naive bayes mencapai tingkat akurasi 75,53%. Decision Tree tingkat akurasi mencapai 61,85%. Dan yang terakhir adalah algoritma KNN mencapai akurasi 73,41%. Dalam penelitian ini, dapat diketahui bahwa metode Naive Bayes adalah metode yang tingkat akurasinya lebih tinggi dibandingkan kedua metode lainnya dengan tingkat akurasi sebesar 75,53%.

Kata kunci: *Analisis Sentimen*; PPPK; Twitter; *Data Mining*

ABSTRACT

The implementation of education in Indonesia is still inseparable from the problems of teacher management, honorary teachers, and bureaucratic reforms that affect the quality of education and the work climate in it. In an effort to improve the quality of public services by the State Civil Apparatus (ASN), the Ministry of Education and Culture agreed with the Ministry of Empowerment of State Apparatus and Bureaucratic Reform and the Ministry of Finance to change the government employee teacher recruitment system from accepting Civil Servant Candidates (CPNS) to Government Employees by Work Agreement (PPPK) which in its implementation still leaves several problems and pros and cons. Therefore, the researchers conducted a sentiment analysis in the field of data mining on the implementation of the teacher's first aid program on Twitter social media as much as 871 data which was then preprocessed into 519 data. The author used the Naive Bayes, Decision Tree and KNN methods to find out the prediction results of the Naive Bayes algorithm, decision tree and KNN regarding public opinion on the implementation of Teacher Teacher Training and Education as well as comparing the level of accuracy of the three methods. The researcher uses RapidMiner version 9.10.1 tools. The results of the Naive Bayes prediction are 328 data with positive sentiment and 191 data with negative sentiment. Furthermore, the decision tree prediction results are 165 data with positive sentiment and 354 data with negative sentiment and the last is the prediction result from KNN, which is 315 data with positive sentiment and 204 data with negative sentiment. The analysis of public sentiment towards the implementation of teacher CPE on Twitter social media with the Naive Bayes algorithm reached an accuracy rate of 75.53%. Decision Tree accuracy rate reaches 61.85%. And finally, the KNN algorithm achieves 73.41% accuracy. In this study, it can be seen that the Naive Bayes method is a method that has a higher accuracy rate than the other two methods with an accuracy rate of 75.53%.

Kata kunci: *Analisis Sentimen*; PPPK; *Twitter; Data Mining*

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Peran guru sangat menentukan kualitas generasi yang dihasilkan dunia pendidikan (Goldman, Ian. and Pabari, 2021). Penyelenggaraan pendidikan di Indonesia hingga saat ini belum lepas dari persoalan tata kelola guru, guru honorer, dan reformasi birokrasi yang berpengaruh terhadap kualitas pendidikan dan iklim kerja di dalamnya (Goldman, Ian. and Pabari, 2021). Dalam upaya peningkatan kualitas layanan publik oleh Aparatur Sipil Negara (ASN), Kementerian Pendidikan dan Kebudayaan sepakat dengan Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi dan Kementerian Keuangan untuk mengubah sistem rekrutmen guru pegawai pemerintah dari penerimaan Calon Pegawai Negeri Sipil (CPNS) menjadi Pegawai Pemerintah dengan Perjanjian Kerja (PPPK(Goldman, Ian. and Pabari, 2021). Pemerintah memiliki rencana besar untuk guru honorer di Indonesia yaitu pengangkatan satu juta guru honorer dalam program rekrutmen pegawai pemerintah dengan perjanjian kerja atau yang disingkat menjadi PPPK pada periode 2021 (Goldman, Ian. and Pabari, 2021). PPPK adalah warga negara Indonesia yang memenuhi syarat tertentu, yang diangkat berdasarkan perjanjian kerja untuk jangka waktu tertentu dalam rangka melaksanakan tugas pemerintahan. (Fahmi et al., 2021). Berdasarkan Undang-Undang (UU) Nomor 5 Tahun 2014 tentang Aparatur Sipil Negara (ASN), PPPK dikontrak minimal satu tahun, dan dapat diperpanjang

Paling lama 30 tahun, Ini semua tergantung situasi dan kondisi yang ada(Goldman, Ian. and Pabari, 2021). Yang dalam pelaksanaannya masih menyisakan beberapa masalah dan pro kontra, ada yang setuju dengan adanya PPPK ini ada juga yang tidak setuju.

Maka dari itu perlu diketahui opini dari masyarakat Indonesia mengenai PPPK. Biasanya masyarakat mengemukakan pendapatnya melalui sosial media, salah satu manfaat dari media sosial adalah sebagai media komunikasi massa sehingga mampu memberikan popularitas kepada siapa saja yang muncul di media massa(Adipradana, 2020). Salah satu aplikasi yang sering digunakan masyarakat Indonesia adalah *Twitter*. Menurut data internal Twitter pada kuartal pertama 2018, pengguna Twitter Indonesia tumbuh sebanyak 11 persen, sementara angka global hanya 10 persen (Normawati & Prayogi, 2021). Pada kuartal kedua, jumlah pengguna Twitter Indonesia tumbuh 31 persen, sedangkan pertumbuhan global hanya 11 persen. Pada kuartal ketiga, Twitter Indonesia mencatat pertumbuhan pengguna aktif harian sebanyak 33 persen, naik tajam dibandingkan dengan pertumbuhan global yang hanya 9 persen, puncaknya pada kuartal keempat 2018, rerata pertumbuhan pengguna Twitter Indonesia sebesar 41 persen, sedangkan pertumbuhan global tetap 9 persen (Adipradana, 2020). *Twitter* adalah sebuah *platform* untuk menyampaikan opini atau pendapat seseorang(Puspita & Widodo, 2021). Pertumbuhan pengguna Twitter di Indonesia sangat pesat dan menduduki peringkat 5 di dunia(Mardiana et al., 2019). Twitter banyak digunakan orang untuk menyampaikan keluh kesahnya mulai dari keluh kesah mengenai kehidupan sehari-hari ataupun keluh kesah

terhadap kebijakan-kebijakan pemerintah(Romadloni et al., 2019). Oleh karena itu sangat efisien jika menggunakan Twitter sebagai media untuk mengambil data mengenai opini masyarakat Indonesia terkait PPPK.

Perlu diketahui bahwa sentimen analisis adalah cabang dari *data mining*. *Data Mining* merupakan sebuah, proses, yang dapat mengekstrak informasi, sehingga menghasilkan informasi yang sangat berharga(Sang et al., 2021). Dengan kata lain dapat juga dikatakan bahwa *data mining merupakan* proses untuk mencari informasi mengenai teknik tertentu. Teknik dan metode dalam data mining sangat banyak (Thakkar et al., 2021). Oleh karena itu, dalam pemilihan teknik atau algoritma yang tepat akan sangat bergantung pada tujuan yang diinginkan (Thakkar et al., 2021). Dalam hal ini, peneliti menggunakan tiga metode dalam analisis sentimen untuk mengkatagorikan hasil komentar netizen dengan melihat prediksi dari ketiga algoritma yang dipakai terkait pelaksanaan PPPK Guru dan untuk memperbandingkan tingkat akurasi dari ketiga metode tersebut diantaranya adalah metode *Naïve Bayes, Decision Tree* dan KNN. Algoritma KNN adalah salah satu algoritma yang sudah *popular*. KNN ini termasuk ke dalam grup *instance-based, learning*. Metode KNN merupakan teknik *lazy, learning* (Puspita & Widodo, 2021). Maksudnya adalah metode ini digunakan dalam klasifikasi data yang jaraknya dekat. Ada juga yang berpendapat bahwa algoritma KNN adalah algoritma pembelajaran yang banyak digunakan dalam sistem *cyber*-fisik-sosial (CPSS) untuk menganalisis dan menambang data (*main* data) (Romadloni et al., 2019). Pada penelitian sebelumnya tentang analisis sentimen, algoritma SVM memiliki akurasi yang lebih tinggi daripada *Naive*

Bayes dan KNN dengan rata-rata akurasinya sebesar 90,01% pada SVM dengan kernel *linear*, 79,20% pada *Naive Bayes* dengan jumlah *laplace* adalah 1, dan 62,10% pada KNN dengan jumlah K adalah 20 dan menggunakan kernel *optimal*(Sodik & Kharisudin, 2021).

Selain metode KNN. Peneliti juga menggunakan metode Decision Tree. Algoritma Tree biasa dipakai untuk pengenalan pola statistik(Romadloni et al., 2019). *Decision Tree* terbuat dari tiga simpul yaitu *leaf*, lalu terdiri juga dari simpul *root* yang merupakan titik awal dari suatu *decision tree*, dan yang terakhir adalah simpul perantara yang berhubungan dengan suatu pengujian(Mardiana et al., 2019). Pada penelitian sebelumnya, Melalui klasifikasi diperoleh hasil akurasi sebesar 90.20% untuk metode *Support Vector Machine* sedangkan 89.80% untuk metode *Decision Tree*. Jadi bisa disimpulkan untuk metode *Support Vector Machine* nilai akurasinya lebih tinggi dibandingkan metode Decision Tree(Purwokerto & Kunci, 2021). Selain menggunakan metode KNN dan *Decision Tree*, peneliti juga menggunakan metode *Naïve Bayes*. *Naïve, Bayes*, adalah metode, *machine learning* untuk probabilitas. Dalam kata lain, *Naïve, Bayes,* merupakan, metode untuk, klasifikasi, text, dengan, kecepatan pemrosesan yang tinggi jika dalam data besar(Demircan et al., 2021a). Ada juga yang berpendapat bahwa *Naïve, Bayes,* adalah metode yang digunakan untuk prediksi karena mengandung probabilistik sederhana yang diterapkan pada *teorema bayes* dengan ketergantungan yang kuat.(Suryono et al., 2018)

Berdasarkan latar belakang diatas, maka dilakukannya penelitian ini untuk ke tiga metode tersebut dengan mengintegrasikan Twitter sebagai *platform* untuk

peneliti melakukan pengolahan data untuk mengetahui analisis sentimen masarakat terhadap PPPK guru dengan *data mining*. Masalah dalam penelitian ini adalah bagaimana hasil prediksi algoritma naive bayes, decision tree dan KNN mengenai data opini masyarakat terhadap pelaksanaan PPPK guru dan berapa tingkat akurasi dari metode KNN, *Decision Tree dan Naïve Bayes*. Selain itu, tujuannya adalah untuk mengatahui hasil prediksi algoritma naive bayes, decision tree dan KNN mengenai data opini masyarakat terhadap pelaksanaan PPPK guru dan berapa tingkat akurasi dari metode KNN, *Decision Tree* dan *Naïve Bayes*.

1.2. Rumusan Masalah

Rumusan masalah dari penelitian ini adalah:

a. Bagaimana hasil prediksi algoritma naive bayes, decision tree dan KNN mengenai data opini masyarakat terhadap pelaksanaan PPPK guru?
b. Bagaimana tingkat akurasi dari metode *Naive Bayes, Decision Tree* dan KNN dalam mengklasifikasikan opini masyarakat terhadap pelaksanaan PPPK Guru?

1.3. Batasan Masalah

Batasan masalah pada penelitian ini adalah :

a. Obyek penelitian adalah opini pengguna *twitter* pada pelaksanaan PPPK Guru.
b. Tahapan Preprocessing yang dilakukan adalah *cleansing, labeling, tokenizing, case folding, stopword removal* dan *token filtering*.
c. Teknik validasi yang digunakan adalah *cross validation* dengan metode *k-fold cross validation*.

d. Algoritma yang digunakan dalam penelitian ini adalah *Naive Bayes Classifier, Decision Tree* dan *K-Nearest Neighbor (K-NN).*

e. Penggunaan ketiga algoritma bertujuan untuk mengkategorikan hasil komentar netizen terkait pelaksanaan PPPK guru dengan melihat hasil prediksi dari metode *naive bayes, decision tree* dan KNN.

f. Media sosial yang digunakan sebagai penelitian ini adalah *Twitter.*

g. Kurun waktu pengambilan data pada Twitter yaitu pada bulan januari 2022.

h. Pelabelan data dilakukan secara manual dengan melibatkan ahli bahasa.

i. Pengelompokan opini pada Twitter dibedakan menjadi sentimen positif dan negatif.

j. Penelitian ini tidak mengambil data faktor pendorong netizen dalam mengeluarkan sentimen.

k. Penelitian ini tidak meneliti *buzzer* yang melakukan *tweet* di *twitter.*

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk:

a. Untuk mengatahui hasil prediksi algoritma *naive bayes, decision tree* dan KNN mengenai data opini masyarakat terhadap pelaksanaan PPPK guru

b. Untuk mengatahui tingkat akurasi dari metode *Naive Bayes, Decision Tree* dan KNN dalam mengklasifikasikan opini masyarakat terhadap pelaksanaan PPPK Guru.

1.5. Manfaat Penelitian

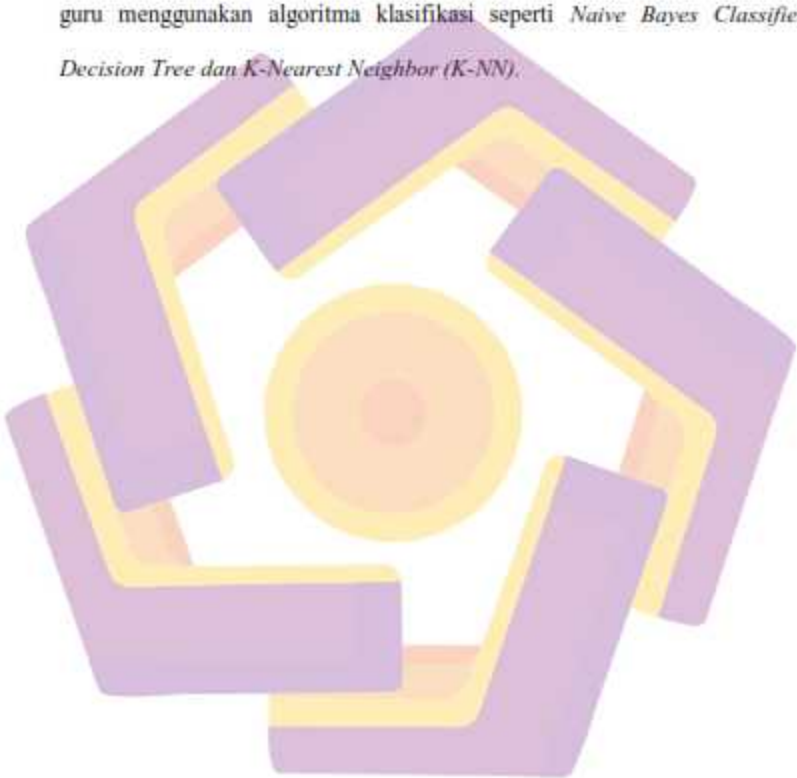
Manfaat penelitian ini adalah :

a. Dapat menjadi pedoman penelitian dalam menganalisa suatu sentimen publik

pada media sosial Twitter.

b. Dapat menjadi bahan acuan menghitung tingkat akurasi penggunaan suatu metode penelitian pada data mining.

c. Untuk mengatahu analisis sentimen masyarakat tentang pelaksanaan PPPK guru menggunakan algoritma klasifikasi seperti *Naive Bayes Classifier, Decision Tree dan K-Nearest Neighbor (K-NN).*



BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Berdasarkan studi literatur yang peneliti amati, penelitian mengenai sentimen analisis yang dilakukan oleh Nova Tri Romadloni, Imam Santoso, Sularso Budilaksono dari Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta, Indonesia tahun 2019, judul penelitian yaitu Perbandingan Metode *Naive Bayes* , KNN Dan *Decision Tree* Terhadap Analisis Sentimen Transportasi Krl, dengan hasil penelitian yaitu Perbedaan hasil akurasi pada metode *Naive Bayes* akurasi sebesar 80%, *precision* 66,67%, *sensitivity* 100%, *specificity* 66,67%. Pada metode KNN akurasi sebesar 80%, *precision* 100%, *sensitivity* 50%, *specificity* 100% dan pada metode *Decision Tree* akurasi sebesar 100%, precision 100%, *sensitivity* 100%, *specificity* 100%(Romadloni et al., 2019). Penelitian yang dilakukan oleh Tobias Daudert tahun 2021 dari *National University of Ireland Galway* dengan judul *Exploiting textual and relationship information for fine-grained financial sentiment analysis*, hasil penelitiannya yaitu Solusi yang di usulkan dapat meningkatkan kinerja sebanyak 15% dan 234% relatif terhadap beberapa *baseline*. Pekerjaan kami menunjukkan dampak sentimen *implisit* serta pentingnya hubungan yang berbeda untuk prediksi sentimen pada laporan perusahaan dan analis, artikel berita, dan mikroblog. Misalnya, kami mengidentifikasi informasi stempel waktu sebagai tidak penting untuk analisis sentimen terperinci dari laporan perusahaan dan analis. Meskipun kami dapat

menunjukkan peningkatan dalam analisis sentimen keuangan, penularan sentimen dan konteks terbatas adalah dua masalah umum yang terus terjadi. Oleh karena itu, dengan mendefinisikan ulang analisis sentimen sebagai masalah multi-teks, solusi yang kami usulkan dapat diterapkan di beberapa domain dan jenis teks, seperti ulasan produk(Daudert, 2021). Dan peneltian yang dilakukan oleh Murat Demircan , Adem Seller , Fatih Abut , Mehmet Fatih Akay tahun 2021 *dari Department of Geomatics Engineering, Istanbul Technical University, Istanbul, Turkey, Universal Yazılım A.Ş., İstanbul, Turkey dan Department of Computer Engineering, Çukurova University, Adana, Turkey. Dengan* judul peneltian *Developing Turkish Sentiment Analysis Models Using Machine Learning and E-Commerce Data,* hasil penelitiannya *yaitu* Hasil validasi silang pada data uji independen yang diambil dari situs e-commerce yang sama, menunjukkan bahwa model analisis sentimen berbasis SVM dan berbasis RF mengungguli model lainnya. Secara lebih rinci, tidak ada urutan yang ketat antara model prediksi berbasis SVM dan berbasis RF, tetapi hasil model berbasis SVM dan berbasis RF, secara umum, adalah yang tertinggi atau, dalam kasus terburuk, serupa jika kita bandingkan. mereka dengan skor yang diperoleh dengan menggunakan model berbasis DT, berbasis LR, dan berbasis KNN. Dapat disimpulkan bahwa SVM dan RF adalah metode yang layak yang dapat digunakan untuk mengklasifikasikan ulasan produk menjadi tiga kelompok sebagai positif, negatif, dan netral dalam tingkat kesalahan yang dapat diterima(Demircan et al., 2021a).

Penelitian yang dilakukan oleh Eka Wahyu Sholeha, Selviana Yunita, Rifqi Hammad, Veny Cahya Hardita dan Kaharuddin tahun 2022, dengan judul

penelitian Analisis Sentimen Pada Agen Perjalanan Online Menggunakan *Naïve Bayes* dan *K-Nearest Neighbor*, hasil dari penelitian ini yaitu: Penelitian ini bertujuan ganda, yaitu untuk membandingkan algoritma, serta menemukan efek dari huruf kapital serta tanda baca. Perbandingan akurasi antara *Naïve Bayes* dan *K-Nearest Neighbor* diberikan terhadap kumpulan data. Penelitian ini mengumpulkan data dari komentar para pengguna Facebook terhadap tiga agen perjalanan online terbesar di Indonesia. Komentar di kelompokkan menjadi tiga kategori, yaitu positif, negatif, dan netral. Hasil dari penelitian ini menemukan jika *K-Nearest Neighbor* memiliki akurasi yang sedikit lebih tinggi dibandingkan *Naïve Bayes*. Selain itu, penggunaan huruf kecil tanpa tanda baca memiliki akurasi yang lebih baik untk kedua algoritma(Sholeha et al., 2022).

Penelitian yang dilakukan oleh M. Khairul Anam, Bunga Nanti Pikir, Muhammad Bambang Firdaus, Susi Erlinda dan Agustin tahun 2021 dengan judul penelitian Penerapan *Na¨ive Bayes Classifier*, *K-Nearest Neighbor (KNN)* dan *Decision Tree* untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah, hasil penelitiannya yaitu *Twitter* menjadi tempat untuk mendapatkan data yang diungkapkan masyarakat melalui tweet yang diposting ke *timeline*. Analisa sentimen dilakukan untuk melihat pendapat atau kecenderungan opini netizen terhadap pemerintah Pekanbaru yang mengandung sentimen positif, negatif, dan netral. Data yang digunakan adalah tweet dengan jumlah dataset sebanyak 150 tweets. Data tersebut kemudian di analisa agar menjadi informasi. Analisa dilakukan menggunakan metode data mining yaitu *Na¨ive Bayes Classifier*, *K-Nearest Neighbor (KNN)*, dan *Decision tree*. Penggunaan ketiga

pendekatan ini berupaya untuk mengkategorikan hasil komentar netizen terkait penggunaan teknologi yang telah melalui proses analisis sentimen dan membandingkan keakuratan ketiga cara tersebut. Hasil akurasi yang didapatkan cukup beragam yaitu dari metode *Na¨ive Bayes* akurasi 100%, metode KKN akurasi 98,25%, dan metode *decision tree* akurasi 62,28% (Anam et al., 2021)

Penelitian yang dilakukan oleh Fajar Sodik Pamungkasa dan Iqbal Kharisudin tahun 2021 dengan judul penelitian Analisis Sentimen dengan SVM, *Naive Bayes* dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial *Twitter*. Hasil penelitiannya yaitu: Pada penelitian ini dilakukan analisis sentimen tanggapan masyarakat Indonesia terhadap pandemi Covid-19 pada media sosial Twitter menggunakan algoritma *Support Vector Machine* (SVM), *Naive Bayes*, dan K-*Nearest Neighbor*, yang kemudian ketiga algoritma tersebut dibandingkan mana yang paling baik untuk mengklasifikasikan data tanggapan. Berdasarkan tingkat rata-rata akurasi dengan menggunakan evaluasi model 10-*Fold Cross Validation*, diperoleh kesimpulan bahwa algoritma SVM memiliki akurasi yang lebih tinggi daripada *Naive Bayes* dan KNN dengan rata-rata akurasinya sebesar 90,01% pada SVM dengan kernel *linear*, 79,20% pada *Naive Bayes* dengan jumlah *laplace* adalah 1, dan 62,10% pada KNN dengan jumlah K adalah 20 dan menggunakan kernel *optima* (Sodik & Kharisudin, 2021).

2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian

Analisis Sentimen Masyarakat Terhadap Pelaksanaan P3k Guru Dengan Algoritma Naive Bayes, Decision Tree Dan K-NN

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Sentiment analylisis of COVID-19 vaccines in indonesia on twitter using pre-tarined and self-training word embeddings(Agustiningsih et al., 2022)	Kartikasari Kusuma Agustiningsih, Ema Utami, Omar Muhammad Altoumi Alsyaibani. Jurnal ilmu komputer dan informasi., (JIKI),e-ISSN:2502-9274, printed ISSN:2338-3070,Vol. 8, No 1,March 2022	Bertujuan untuk memeriksa kinerja penyisipan kata, menggunakan vektor penyematan kata yang telah dilatih sebelumnya dan yang dilatih sendiri. Serta menguji pengaruh proses stemming baik dalam vektor pra-terlatih dan vektor mandiri.	Dari hasil uji klasifikasi didapatkan diketahui bahwa penyematan kata yang dilatih sendiri dapat membuat model menghasilkan akurasi yang lebih tinggi daripada menggunakan penyisipan kata pra-latihan. Model yang menggunakan penyematan kata GloVe menghasilkan akurasi yang lebih tinggi pada kumpulan data yang tidak distem. Hasil yang berbeda dihasilkan oleh model yang menggunakan penyisipan kata fastText di mana akurasi yang lebih tinggi diperoleh saat menggunakan kumpulan data bertangkai. Perbedaan akurasi antara penyematan kata fastText dan GloVe sangat kecil di mana penyematan kata GloVe menghasilkan akurasi yang sedikit lebih tinggi daripada FastText.	Analisis sentimen pada dataset bahasa non-formal seperti deteksi emosi, klasifikasi sentimen atau deteksi sarkasme yang menggunakan penyisipan kata untuk menghasilkan vektor penyisipan kata yang dilatih sendiri. Perbedaan akurasi antara penyematan kata fastText dan GloVe sangat kecil di mana penyematan kata GloVe menghasilkan akurasi yang sedikit lebih tinggi daripada FastText.	Pada penelitian ini peneliti menggunakan Bidirectional LSTM yang akan dikombinasikan dengan word embedding, GloVe dan fastText. Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix.

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

2	Sentiment Analysis and topic modelling of the COVID-19 vaccine in indonesia on twitter social media using word embedding(Agustiningsih et al., 2021)	Kartikasari Kusuma Agustiningsih, Ema Utami, Omar Muhammad Altoumi Alsyaibani. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), e-ISSN: 2338-3062, printed ISSN: 2338-3070, Vol. 8, No. 1, March 2022.	Penelitian ini bertujuan untuk menganalisis sentimen masyarakat Indonesia terhadap vaksin COVID-19 di Twitter.	Dari data yang dipkai dapat disimpulkan bahwa meskipun ada masyarakat yang memiliki sentimen negatif terhadap vaksin yang hanya 8%, namun penerimaan masyarakat Indonesia terhadap vaksin cukup tinggi yaitu 42%. Sementara itu, 50% sentimen diskusi bersifat netral. Diskusi publik tentang vaksin dimulai pada September 2020. Jumlah tweet terbanyak muncul pada Januari 2021, yaitu 23492 tweet. Berdasarkan hasil pengukuran menggunakan Skor Koherensi, membagi data menjadi 3 topik menghasilkan Skor Koherensi tertinggi, yaitu 0,4824. Topik pertama memiliki nilai persentase token 51,8% mengarah ke sentimen positif, sedangkan topik kedua dan ketiga mengarah pada berita tentang kegiatan vaksinasi dan protokol kesehatan yang netral dengan nilai persentase token masing-masing 24,5% dan 23,7%. Secara umum performansi model LSTM Bidirectional	Menangani masalah ketidakseimbangan dalam data yang dikumpulkan, pemodelan topik juga perlu dilakukan pada data bulan-bulan tertentu untuk memahami lebih detail topik apa yang dibahas pada waktu-waktu tertentu, belum mencoba menggunakan model berbasis transformator yang merupakan penyematan kata mutakhir saat ini. Belum bisa mendapatkan akurasi yang lebih baik dari pengujian menggunakan SVM dengan pembobotan TF-IDF	Peneliti menggunakan SNScrape tool yang digunakan untuk pengumpulan data, GloVe dan fastText, Latent Dirichlet Allocation (LDA) untuk menentukan topic modeling, selain itu klasifikasi dan training monitoring juga digunakan dalam penelitian ini. Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

				pada penelitian ini hanya mencapai sekitar 73% - 75% bahkan dengan berbagai skenario. Uji akurasi uji tertinggi dihasilkan oleh model yang menggunakan penyisipan kata Fasttext. Penggunaan kata-kata slang tidak dapat meningkatkan akurasi tes dalam penelitian ini		
3	An optimization of a lexicon based sentiment analysis method on Indonesian app review(Pratama et al., 2019)	Bayu Trisna Pratama, Ema Utami, Andi Sunyoto. 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019	Penelitian ini bertujuan untuk mengoptimalkan metode yang diusulkan oleh penelitian sebelumnya dengan berfokus pada dua peluang optimasi: (1) sumber daya leksikon yang digunakan, dan (2) penggunaan fitur spesifik domain	Hasil penelitian menunjukkan bahwa SentiWordNet dapat mengungguli sumber daya leksikon lainnya dan penambahan fitur khusus domain berdampak baik pada kinerja pengklasifikasi baik pada akurasi keseluruhan dan parameter evaluasi rata-rata f-measure dibandingkan tidak hanya hasil penelitian ini tetapi juga dibandingkan dengan hasil penelitian sebelumnya.	Fokus pada penggunaan dataset yang lebih besar untuk menguji konsistensi kinerja classifier. Data yang digunakan tidak terlalu banyak yaitu 533 ulasan pengguna diambil dari Apple App Store dan Google Play dan memiliki 241 komentar negatif, 259 komentar positif, dan 53 komentar netral	Penelitian ini menggunakan pendekatan berbasis leksikon yang tidak membutuhkan data berlabel untuk pelatihan. Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix
4	Natural Language Processing on Marketplace Product Review Sentiment Analysis(Rohman et al., 2020)	Arif Nur Rohman, Rizqa Luviana Musyarofah, Ema Utami, Suwanto Raharjo. 2020 2nd International Conference on	Penelitian ini bertujuan untuk melakukan analisis sentimen pada ulasan produk pasar.	Algoritma K-NN memberikan akurasi yang lebih baik, presisi dan recall dibandingkan dengan Naive Bayes. Paling atas Akurasi K-NN pada uji coba dataset Bigram dengan K = 9	Penelitian selanjutnya sebaiknya menerapkan deteksi spam untuk menemukan review yang tidak relevan tentang suatu produk sehingga dapat	Penelitian ini menggunakan metode K-NN dan Naive Bayes. Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

5	Sentiment Analysis of COVID-19 Vaccine on Twitter Social Media: Systematic Literature Review(Agustiningsih & Utami, 2020)	Kartikasari Kusuma Agustiningsih, Ema Utami, Hanif Al Fatta. The 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2021	Penelitian ini merangkum beberapa penelitian mengenai analisis sentimen pada program vaksinasi COVID-19 di Twitter. Hal ini bertujuan untuk menggambarkan bidang studi ini bagi peneliti selanjutnya dalam melakukan penelitian lebih lanjut.	Sebagian besar penelitian analisis sentimen dilakukan dalam bahasa Inggris. Sementara itu, analisis sentimen pada bahasa lain masih relatif sedikit. Hasil penelitian yang diulas menunjukkan bahwa sentimen positif masyarakat terhadap vaksin COVID-19 lebih besar daripada sentimen negatif. Penyebab sentimen negatif dapat ditemukan dengan mendeteksi topik dalam tweet sentimen negatif.	Perlu dilakukan analisis sentimen pada bahasa lain, mengingat penelitian tentang analisis sentimen akan terus bermunculan dari peneliti lain dari berbagai negara seiring dengan penerapan vaksinasi di seluruh dunia dan perkembangan vaksin COVID-19 di masing-masing Negara Belum mengkaji metode klasifikasi secara detail, sehingga masih diperlukan penelitian lebih lanjut untuk memberikan gambaran yang komprehensif tentang metode dan variabel atau parameter yang digunakan kepada pembaca yang ingin mempelajari tentang analisis sentimen menggunakan data dari Twitter	Penelitian ini menggunakan metode Systematic Literature Review (SLR) adalah metode pengumpulan hasil penelitian dari berbagai sumber. Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

6	Comparative study of various approaches, applications and classifiers for sentiment analysis(Sudhir & Deshakulkarni, 2021)	Prajval Sudhir, Varun Deshakulkarni Suresh. by Elsevier B.V. on behalf of KeAi. Communications Global Transitions Proceedings 2, 2021	Penelitian ini bertujuan untuk mengetahuai berbagai pendekatan dan model klasifikasi yang digunakan dalam tugas analisis sentimen.	Hasil akurasi untuk model berdasarkan dataset IMDB menggambarkan bahwa pendekatan pembelajaran mesin seperti SVM, GRU dan BERT menunjukkan akurasi yang luar biasa. Khususnya, model yang lebih baru seperti GRU dan BERT menunjukkan akurasi yang melebihi model klasifikasi konvensional seperti Naive Bayes, Decision Tree dll.	Tidak menggunakan confusion matrix sebagai alat pengukuran hasil evaluasi. Tidak menggunakan model klasifikasi KNN, Naive Bayes dan Decision Tree	Penelitian ini hanya membahas keuntungan dan kerugian dari berbagai aplikasi dan model klasifikasi, tetapi tidak menggunakan confusion matrix sebagai alat pengukuran hasil evaluasi.Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix.
7	Public sentiment analysis and topic modeling regarding COVID-19vaccines on the Reddit social media platform: A call to action forstrengthening vaccine confidence(Melton et al., 2021)	Chad A. Meltona, Olufunto A. Olusanyab, Nariman Ammarb, Arash Shaban-Nejad. Journal of Infection and Public Health 14, 2021	Untuk melakukan analisis sentimen dan pemodelan topik Alokasi Dirichlet Laten pada data tekstual yang dikumpulkan dari 13 komunitas Reddit yang berfokus pada vaksin COVID-19 dari 1 Desember 2020, hingga 15 Mei 2021	Kesimpulan Analisis konten terkait vaksin Covid-19 dari 13 subred-dit menunjukkan bahwa sentimen yang diungkapkan dalam komunitas media sosial ini secara keseluruhan lebih positif daripada negatif tetapi tidak berubah secara berarti sejak Desember 2020.	Penelitian ini menggunakan unsupervised learning. Tidak menggunakan algoritma Naive Bayes, Decision Tree dan K-NN	Penelitian ini menggunakan unsupervised learningPenelitian berikutnyamenggunakan supervised learning dengan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

8	Developing Turkish sentiment analysis models using machine learning and e-commerce data(Demircan et al., 2021b)	Murat Demircan ,Adem Seller , Fatih Abut , Mehmet Fatih Akay. International Journal of Cognitive Computing in Engineering 2, 2021.	Penelitian ini bertujuan untuk mengetahui sentimen yang diungkapkan melalui teks di media sosial dengan menggunakan metode machine learning, menggunakan mesin vektor dukungan (SVM), hutan acak (RF), pohon keputusan (DT), regresi logistik (LR), dan k-nearest neighbor (KNN).	Dapat disimpulkan bahwa SVM dan RF adalah metode yang layak yang dapat digunakan untuk mengklasifikasikan ulasan produk menjadi tiga kelompok sebagai positif, negatif, dan netral dalam tingkat kesalahan yang dapat diterima.	Tidak menggunakan twitter utk crawling data. Tidak menggunakan confusion matrix sebagai alat pengukuran hasil evaluasi. Tidak menggunakan algoritma Naive Bayes, Decision Tree	Tidak menggunakan twitter utk crawling data. Tidak menggunakan confusion matrix sebagai alat pengukuran hasil evaluasi. Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix.
9	Restaurant recommender system based on sentiment analysis(Asani et al., 2021)	Elham Asani , Hamed Vahdat-Nejad , Javad Sadri. Machine Learning with Applications 6, 2021	Untuk sistem rekomendasi makanan dan restoran berdasarkan hasil ekstraksi dari komentar pengguna dan menganalisis sentimen mereka.	Memanfaatkan kriteria Wu–Palmer untuk pengelompokan dan kosinus kriteria untuk menghitung kesamaan antara preferensi dan menu menghasilkan hasil terbaik mengenai presisi, recall dan f-measure. Itu hasil juga mengungkapkan bahwa sistem yang diusulkan dapat menyediakan pengguna dengan	Data yang digunakan sangat sedikit yaitu 100 komentar yang di ambil dari situs Web TripAdvisor. Tidak menggunakan algoritma Naive Bayes, Decision Tree dan K-NN	Penelitian berikutnya menggunakan data yang lebih besar. Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix.

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

				Presisi 92,8% dalam mode Top5. Akhirnya, membandingkan yang diusulkan sistem dengan pekerjaan terkait sebelumnya telah menunjukkan peningkatan di sebagian besar dari kriteria di atas.		
10	Decision Support System for Covid19 Affected Family Cash Aid Recipients Using the Naïve Bayes Algorithm and the Weight Product Method(Sa'ad et al., 2020).	Muhammad Ibnu Sa'ad, Doni Bryan, Kusrini, Supriatin. 2020 3rd International Conference On Information and Communications Technology (ICOIACT) 2020	Tujuan dari penelitian ini adalah untuk memprediksi penerima bantuan tunai dan untuk mengevaluasi Naïve Bayes dalam memprediksi penerima bantuan tunai dari keluarga terdampak Covid19.	Dari hasil penelitian ini, dapat disimpulkan bahwa penerima bantuan tunai di Desa X dapat diprediksi Naïve Bayes menggunakan a nilai pelatihan 10%. Berdasarkan hasil evaluasi menggunakan matriks kebingungan dan menguji akurasi Naïve Bayes adalah 67%. Untuk perhitungan metode Weighted Product menggunakan variabel Age, Pendapatan, Pendidikan, Status Pekerjaan, Status Keluarga dan ada dua alternatif yaitu Cannot dan Can. Dari perhitungan Weighted Product menghasilkan Nilai rangking Vektor S sebesar 2,24 dan Vektor V sebesar 0,66, yang menyatakan bahwa keluarga yang terkena dampak Covid19 memiliki hak untuk menerima bantuan tunai.	Tidak menggunakan algoritma Decision Tree dan K-NN	Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix.

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

11	Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach(Muktafin & Kusrini, 2021)	Elik Hari Muktafin, Pramono, Kusrini. TELKOMNIKA Telecommunication, Computing, Electronics and Control Vol. 19, No. 1, February 2021	sistem ini bertujuan untuk mengubah percakapan suara menjadi teks dan menganalisis kepuasan pelanggan untuk mendapatkan informasi untuk evaluasi dan peningkatan layanan.	Penelitian ini menggunakan algoritma K-nearest neighbor (KNN) dan term frequency-inverse document frequency (TF-IDF) dengan pendekatan natural language processing (NLP) untuk mengklasifikasikan percakapan menjadi 2 kelas yaitu "puas" dan "tidak puas". Hasil penelitian ini mendapatkan akurasi 74,00%, presisi 76,00% dan recall 73,08%.	Dataset diambil dari hasil percakapan yang diubah menjadi teks dan diklasifikasikan menggunakan KNN. Tidak menggunakan algoritma Naive Bayes dan Decision Tree	Penelitian berikutnya mengambil data dari tweeter, menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix.
12	Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial *Twitter*	Fajar Sodik Pamungkasa, Iqbal Kharisudina. PRISMA 4 (2021): 628-634 PRISMA, Prosiding Seminar Nasional Matematika, 2021	Untuk mengetahui sentimen tanggapan masyarakat terhadap COVID-19, analisis sentimen dengan algoritma machine learning. Pada penelitian ini dilakukan analisis sentimen tanggapan masyarakat Indonesia terhadap pandemi Covid-19 pada media sosial Twitter menggunakan algoritma *Support*	Berdasarkan tingkat rata-rata akurasi dengan menggunakan evaluasi model 10-*Fold Cross Validation*, diperoleh kesimpulan bahwa algoritma SVM memiliki akurasi yang lebih tinggi daripada *Naive Bayes* dan KNN dengan rata-rata akurasinya sebesar 90,01% pada SVM dengan kernel *linear*, 79,20% pada *Naive Bayes* dengan jumlah *laplace* adalah 1, dan 62,10% pada KNN dengan jumlah K adalah 20 dan menggunakan kernel *optimal*.	Tidak menggunakan algoritma Decision Tree	Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix.

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

			Vector Machine (SVM), *Naive Bayes*, dan K-*Nearest Neighbor*	.		
13	Perbandingan Metode *Support Vector Machine* Dan *Decision Tree* Untuk Analisis Sentimen *Review* Komentar Pada Aplikasi Transportasi Online	Khoirul Abbi Rokhman, Berlilana, Primandani Arsi. *Joism : Jurnal Of Information System Management E-Issn: 2715-3088 Vol. 2, No. 2. (2021)*	Tujuan dari penelitian ini yaitu untuk melakukan analisis sentiment dengan menggunakan data ulasan yang terdapat pada *Google Play* guna mengetahui perbandingan keakurasian antara metode *Support Vector Machine* untuk mengklasifikasikan ulasan dari dua ketegori yaitu ulasan positif dan negative. Kemudian dibandingkan dengan metode *Decision Tree*	Melalui klasifikasi diperoleh hasil akurasi sebesar 90.20% untuk metode *Support Vector Machine* sedangkan 89.80% untuk metode *Decision Tree*. Jadi bisa disimpulkan untuk metode *Support Vector Machine* nilai akurasinya lebih tinggi dibandingkan metode *Decision Tree*.	Tidak menggunakan algoritma Naive Bayes dan KNN	Penelitian berikutnya menggunakan algoritma Naive Bayes, Decision Tree dan K-NN, software yang digunakan adalah RapidMiner. Dan evaluasi menggunakan confusion Matrix.

2.3. Landasan Teori

2.3.1 PPPK Guru

Penyelenggaraan pendidikan di Indonesia hingga saat ini belum lepas dari persoalan tata kelola guru, guru honorer, dan reformasi birokrasi yang berpengaruh terhadap kualitas pendidikan dan iklim kerja didalamnya (Goldman, Ian. and Pabari, 2021). Dalam upaya peningkatan kualitas layanan publik oleh Aparatur Sipil Negara (ASN), Kementerian Pendidikan dan Kebudayaan sepakat dengan Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi dan Kementerian Keuangan untuk mengubah sistem rekrutmen guru pegawai pemerintah dari penerimaan Calon Pegawai Negeri Sipil (CPNS) menjadi Pegawai Pemerintah dengan Perjanjian Kerja (PPPK) (Fahmi et al., 2021). Pemerintah memiliki rencana besar untuk guru honorer di Indonesia yaitu pengangkatan satu juta guru honorer dalam program rekrutmen pegawai pemerintah dengan perjanjian kerja atau yang disingkat menjadi PPPK pada periode 2021 (Fahmi et al., 2021). PPPK adalah warga negara Indonesia yang memenuhi syarat tertentu, yang diangkat berdasarkan perjanjian kerja untuk jangka waktu tertentu dalam rangka melaksanakan tugas pemerintahan sesuai dengan bunyi Pasal 1 ayat (4) UU Nomor 5 Tahun 2014. Berdasarkan Undang-Undang (UU) Nomor 5 Tahun 2014 tentang Aparatur Sipil Negara (ASN), PPPK dikontrak minimal satu tahun, dan dapat diperpanjang paling lama 30 tahun, Ini semua tergantung situasi dan kondisi yang ada (Fahmi et al., 2021). Yang dalam pelaksanaannya masih menyisakan beberapa masalah dan prokontra, ada yang setuju dengan adanya PPPK ini ada juga yang tidak setuju.

2.3.2 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) atau yang juga dikenal dengan pemrosesan bahasa Alami merupakan salah satu cabang ilmu dari *artificial intelligence* atau kecerdasan buatan yang memberi pelajaran terhadap komputer agar dapat memahami, menafsirkan serta memanipulasi bahasa alami manusia (Sang et al., 2021). NLP merupakan teknologi yang merepresentasikan kemampuan dari komputer atau sistem yang dirancang untuk menirukan cara berpikir manusia dalam memahami dan memproses bahasa manusia secara alami (Setiawan, 2021).

Pada NLP dilakukan pembuatan model komputasi bahasa yang memungkinkan adanya interaksi manusia dan komputer dengan perantara bahasa alami yang digunakan oleh manusia (Setiawan, 2021). NLP bertujuan memecahkan masalah untuk memahami bahasa manusia dengan berbagai aturan gramatika dan aturan semantiknya serta mengubah bahasa tersebut menjadi representasi formal yang dapat dipahami dan diproses oleh komputer (Thakkar et al., 2021). Pada praktek dan penerapannya, NLP ini memiliki tantangan, yang diantaranya adalah sebagai berikut (Adipradana, 2020):

a. *Part-of-speech Tagging*

Part-of-speech tagging atau penandaan kelas kata ini sulit dilakukan karena pengelasan kata sangat bergantung pada konteks penggunaan kata tersebut.

b. *Text Segmentation*

Text segmentation atau segmentasi teks sulit dilakukan terhadap bahasa tulis yang tidak memiliki pembatas kata yang spesifik, seperti misalnya pada bahasa

Thailand, Mandarin dan Jepang. Segmentasi teks juga sulit dilakukan terhadap bahasan lisan yang seringkali membaurkan bunyi antar kata.

c. *Word Sense Disambiguation*

Word sense disambiguation atau makna kata yang ambigu. Seringkali ditemukan suatu kata yang memiliki lebih dari satu makna baik dalam bentuk homonim atau polisemi. Untuk dapat membedakan makna kata tersebut harus terlebih dahulu dilihat dari konteks penggunaannya.

d. *Syntatic Ambiguity*

Syntatic ambiguity atau ambiguitas sintaksis. Suatu bahasa memiliki berbagai macam struktur kalilmat yang berbeda. Untuk memilih struktur kalimat yang tepat dibutuhkan penggabungan informasi semantik dan juga kontekstual.

e. *Imperfect or Irregular Input*

Imperfect or irregular input atau masukan yang tidak sempurna atau tidak biasa. Kesalahan penulisan baik secara pengetikan atau gramatikal juga mempersulit proses NLP.

f. *Speech Act*

Speech act atau penuturan. Terkadang struktur kalimat tidak dengan tepat menggambarkan tujuan dan maksud dari penulis, gaya bahasa dan konteks kalimat menentukan maksud yang dinginkan (Adipradana, 2020).

2.3.3 Sentiment Analysis

Sentiment analysis merupakan merupakan kajian tentang cara menyelesaikan dan memecahkan masalah dari berdasarkan opini masyarakat, sikap serta emosi suatu entitas, dimana entitas tersebut dapat mewakili individu

(Sugianto, 2015)*Sentiment analysis* atau yang juga disebut *opinion mining* merupakan proses memahami, mengekstrak serta mengolah data tekstual secara otomatis guna mendapatkan informasi yang terkandung dalam suatu kalimat opini. Dilakukannya analisis sentimen ini bertujuan untuk melihat pendapat atau kecenderungan opini terhadap suatu masalah ataupun objek oleh seseorang, apa memiliki kecenderungan positif, negatif, atau netral (Sang et al., 2021). *Sentiment analysis* merupakan bagian dari NLP yang membangun sistem untuk mengenali serta mengekstraksi opini dalam bentuk teks. Di internet banyak terdapat informasi berbentuk teks yang tidak terstruktur, seperti misalnya pada blog, forum, media sosial, situs yang berisi *review*, dan lain-lain. Informasi yang tidak terstruktur ini kemudian diubah menjadi data yang terstruktur dengan *sentiment analysis*. Terdapat tiga jenis *sentiment analysis* yaitu *Fine-Grained Sentiment Analysis*, *Intent Sentiment Analysis* dan *Aspect-Based Sentiment Analysis*. Perbedaan ketiganya adalah pada *Fine-Grained Sentiment Analysis* akan mengelompokkan respon menjadi beberapa kategori seperti positif, negatif dan netral. *Intent Sentiment Analysis* akan mengidentifikasi sebuah pesan atau teks termasuk kedalam golongan pesan atau teks keluhan, saran, pendapat, pertanyaan, atau pujian terhadap sesuatu, misalnya terhadap produk atau layanan. Sementara pada *Aspect-Based Sentiment Analysis* dapat lebih fokus pada elemen-elemen yang lebih spesifik dari suatu teks(Demircan et al., 2021a).

2.3.4 Twitter

Twitter merupakan sebuah situs media sosial yang mulai dikembangkan pada tahun 2006. Situs ini pertama kali ditemukan oleh Jack Dorsey dan Evan

Williams. Twitter merupakan *social networking* dimana memungkinkan penggunanya dapat saling berkomunikasi satu sama lain melalui fitur yang bernama *tweet*(Indriani, 2019). Dengan fitur *tweet* pengguna dapat membuat tulisan atau teks sebanyak 280 karakter(Puspita & Widodo, 2021). Tidak hanya *tweet*, saat ini *Twitter* memiliki banyak fitur lainnya seperti *direct message* yang memungkinkan pengguna berkomunikasi satu sama lain dengan lebih privat, *story* yang memungkinkan pengguna dapat merekam momen baik itu foto atau video secara langsung atau *realtime*, *live* yang memungkinkan pengguna melakukan siaran langsung (melalui aplikasi pihak ketiga *Periscope*), *voice note* memungkinkan pengguna untuk merekam suara, dan masih banyak fitur lainnya. *Twitter* menyediakan API (*Application Programming Interface*). *Twitter* API diperuntukkan bagi pengembang. Dengan *Twitter* API memungkinkan pengguna dapat membaca, menulis dan mengambil data dari *Twitter*. Penggunaan *Twitter* API ini juga memungkinkan pengembang untuk mengambil informasi atau data pengguna di *Twitter* atau suatu subjek di lokasi tertentu (Romadloni et al., 2019)

2.3.5 Data Mining

Data mining adalah proses menganalisis data dari berbagai perpektif dan merangkum menjadi informasi yang berguna, informasi tersebut dapat digunakan untuk meningkatkan pendapatan dan memangkas biaya, hal ini memungkinkan pengguna untuk menganalisis data dari berbagai dimensi, mengkategorikan, dan merangkum hubungan yang diidentifikasi. Tetapi dalam beberapa tahun terakhir, selain untuk mengolah data, teknik *data mining* dapat berubah sesuai kecanggihan

dan kemudahan penggunaan alat untuk menganalisis data sesuai dengan peningkatan jumlah peneliti untuk menerapkan *data mining*(Syakuro, 2017)

2.3.6. Klasifikasi

Klasifikasi merupakan kategori *supervised learning*. Metode klasifikasi mempelajari data yang ada dan memprediksi data-data lainnya. Ada beberapa algoritma yang sering digunakan dalam klasifikasi misalnya *Naïve Bayes, K-Nearest Neighbor, Support Vector Machine* dan lain-lain (Sang et al., 2021). Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya kedalam kelas tertentu dari sejumlah kelas yang tersedia, dalam klasifikasi ada dua pekerjaan utama yang dilakukan yaitu satu pembangunan model sebagai prototype untuk disimpan sebagai memori dan yang kedua penggunaan model tersebut untuk klasifikasi prediksi pada suatu objek data lain agar diketahui dikelas mana data objek tersebut.(Indriani, 2019)

2.3.7 Naïve Bayes

Merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes . Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes , yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes (Syakuro, 2017) . Ciri utama dr Naïve Bayes Classifier ini adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian *Naïve Bayes* untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah

independen (Puspita & Widodo, 2021). *Naive Bayes Classifier* bekerja sangat baik dibanding dengan model *classifier* lainnya. Sebelum menjelaskan *Naïve Bayes Classifier* ini, akan dijelaskan terlebih dahulu Teorema *Bayes* yang menjadi dasar dari metoda tersebut. Pada Teorema Bayes, bila terdapat dua kejadian yang terpisah (misalkan x dan H), maka Teorema Bayes dirumuskan sebagai berikut (Syakuro, 2017).

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \quad (2.1)$$

Dimana :

X : Data dengan *class* yang belum diketahui

H : Hipotesis data merupakan suatu *class* spesifik

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (*posteriori probabilitas*)

P(H) : Probabilitas hipotesis H (*prior probabilitas*)

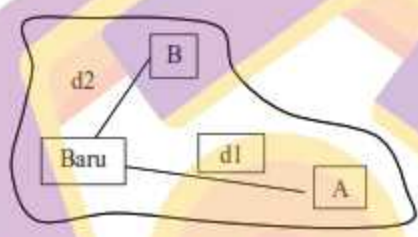
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) : Probabilitas X

2.3.8 K-Nearest Neighbor (KNN)

Algoritma K-Nearest Neighbor (KNN) adalah merupakan sebuah metode untuk melakukan klasifikasi terhadap objek baru berdasarkan (K) tetangga terdekatnya. KNN termasuk algoritma supervised learning, dimana hasil dari query instance yang baru, dklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Kelas yang paling banyak muncul yang akan menjadi kelas hasil klasifikasi (Sang et al., 2021). *Nearest Neighbor* adalah suatu pendekatan untuk

menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Ilustrasi kedekatan kasus memberikan gambaran tentang proses mencari solusi terhadap seorang pasien baru dengan mengacu pada pasien terdahulu, untuk mencari kasus pasien mana yang akan dgunakan maka dihitung kedekatan antara kasus fasien baru dengan semua kasus pasien lama. Kasus pasien lama dengan kedekatan terbesarlah yang akan diambil solusinya untuk digunakan pada kasus pasien baru.



Gambar 2.1. Ilustrasi kedekatan kasus

(Sumber: Kusrini & Emha Taufiq Luthfi, 2009)

Seperti tampak pada Gambar. Ada 2 pasien lama A dan B. Ketika ada pasien Baru, maka solusi yang akan diambil adalah solusi dari pasien terdekat dari pasien Baru. Seandainya d1 adalah kedekatan antara pasien Baru dan pasien A, sedangkan d2 adalah kedekatan antara pasien Baru dengan pasien B. Karena d2 lebih dekat dari d1 maka solusi dari pasien B lah yang akan digunakan untuk memberikan solusi pasien Baru. Rumus kedekatan Kedekatan biasanya berada pada nilai antara 0 s/d 1. Nilai 0 artinya kedua kasus mutlak tidak mirip, sebaliknya untuk nilai 1 kasus mirip dengan mutlak (Kusrini & Emha Taufiq Luthfi, 2009).

$$Similarity(T,S) = \frac{\sum_{i=1}^{n} F(Ti,Si)x\,Wi}{\sum_{i=1}^{n} Wi} \quad (2.2)$$

Dimana:

T : Kasus baru

S : Kasus yang ada dalam penyimpanan

n : Jumlah atribut dalam masing – masing kasus

i : Atribut individu antara 1 s/d n

f : Fungsi similarity atribut i antara kasus T dan kasus S

W : Bobot yang diberikan pada atribut ke-i

2.3.9 Decision Tree

Decision Tree (Pohon Keputusan) adalah pohon dimana setiap cabangnya menunjukkan pilihan diantara sejumlah alternatif pilihan yang ada, dan setiap daunnya menunjukkan keputusan yang dipilih(Puspita & Widodo, 2021). Decision tree biasa digunakan untuk mendapatkan informasi untuk tujuan pengambilan sebuah keputusan. Decision Tree digunakan untuk mempelajari klasifikasi dan prediksi pola dari data dan menggambarkan relasi dari variabel attribut x dan variabel target y dalam bentuk pohon(Romadloni et al., 2019) .Kelebihan Decision Tree yaitu Daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik, Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode pohon keputusan maka sampel diuji hanya berdasarkan criteria atau kelas tertentu, Fleksibel untuk memilih fitur dari node internal yang berbeda, fitur yang terpilih akan membedakan suatu kriteria dibandingkan kriteria yang lain dalam node yang sama. Kefleksibelan metode pohon keputusan ini

meningkatkan kualitas keputusan yang dihasilkan jika dibandingkan ketika menggunakan metode penghitungan satu tahap yang lebih konvensional. Dalam analisis multivarian, dengan kriteria dan kelas yang jumlahnya sangat banyak, seorang penguji biasanya perlu mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut(Sang et al., 2021).

2.3.10 *Text Processing*

Tahap ini merupakan tahapan pertama dalam pemrosesan teks. Tahap ini juga merupakan salah satu tahapan yang paling penting dalam *text mining*. *Text* Preprocessing ini merupakan tahap dimana sistem melakukan seleksi data yang kemudian akan diproses pada setiap dokumen (Thakkar et al., 2021).

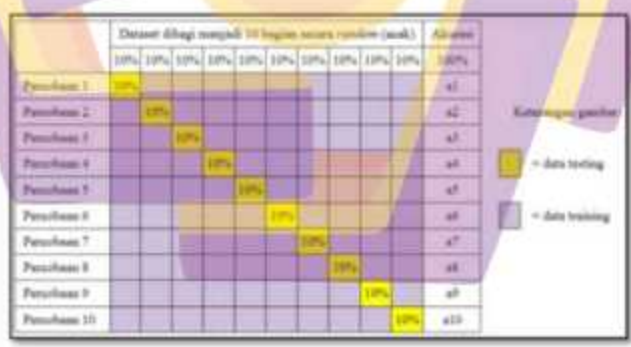
2.3.11 *Cross Validation*

Cross validation adalah suatu metode tambahan dari teknik data mining yang bertujuan untuk memperoleh hasil akurasi yang maksimal metode ini juga sering disebut k-fold cross validation dimana percobaan sebanyak k kali untuk satu model dengan parameter yang sama.(Suryono et al., 2018). Fungsi dari penggunaan cross validation adalah:

a. Untuk mengatahui performa dari suatu model algoritma dengan melakukan percobaan sebanyak *k* kali.
b. Untuk meningkatkan tingkat performance dari model tersebut.
c. Untuk mengolah dataset dengan kelas yang seimbang.

Dalam kasus klasifikasi ada yang perlu diperhatikan dalam pembagian set data ke sejumlah *k* partisi, yaitu harus melakukan *stratification* yang artinya kita akan mempartisi atau membagi set data tersebut ke *k* partisi dengan komposisi

kelas yang seimbang disetiap partisinya. Cross validation merupakan teknik validasi dari pengembangan model split validation dimana cross validation mampu bekerja dengan cepat dengan pengambilan sampel yang lebih tersetruktur, jadi dalam jumlah pengujian beberapapun set data latih dan set data uji akan diambil dengan data yang berbeda dengan percobaan atau literasi sebelumnya. Yang nantinya hasil percobaan tersebut akan dicatat nilai evaluasi ferporma darri sebuah model dengan menggunakan confussion matrix. Dalam beberapa penelitian yang sudah dilakukan oleh pakar pakar data mining, model pengujian atau validasi model dari suatu algoritma klasifikasi, *cross validation* lebih sering dipakai ketimbang *split validition* karna model validasi dengan menerapkan *10-cross validation* sudah merupakan standar dan suatu metode validasi yang canggih atau lebih praktis dan efisien serta mempu meningkatkan sedikit nilai performansinya.



Gambar 2.2. Cara kerja cross validation

2.3.12. Confusion Matrix

Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. Evaluasi dimaksudkan untuk menguji model klasifikasi data mining untuk mengetahui kinerja sistem. Salah satu metode untuk mengukur evaluasi performansi adalah *confusion matrix*(Sang et al., 2021). *Confusion matrix* adalah alat ukur berbentuk matrix yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi terhadap kelas dengan algoritma yang dipakai(Sang et al., 2021). Evaluasi performansi dapat diukur dengan *precision* untuk menunjukkan tingkat ketepatan atau ketelitian dalam pengklasifikasian, *recall* berfungsi untuk mengukur proporsi positif aktual yang benar diidentifikasi, *F1-Measure* dapat didefinisikan sebagai alternatif dari metode akurasi yang diperoleh dari hasil perhitunganantara presisi dan *recall*, dan akurasi dapat didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai actual (Sang et al., 2021). Masing-masing perhitungan memiliki persamaan:

$$
\begin{aligned}
precision &= \frac{TP}{TP + FP} \\
recall &= \frac{TP}{TP + FN} \\
accuracy &= \frac{TP + TN}{TP + TN + FP + FN}
\end{aligned}
\qquad (2.3)
$$

2.3.13 Rapid Miner

RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan rediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik (Sugianto, 2015).

RapidMiner berkembang sejak tahun 2001, sebelumnya disebut dengan nama YALE (Yet Another Learning Environment). Salah satu ciri utama RapidMiner adalah kemampuannya yang canggih untuk memprogram eksekusi alur kerja yang kompleks, semua dilakukan dalam antarmuka pengguna visual, tanpa perlu tradisional keterampilan pemrograman [10]. Software ini dikembangkan oleh Ralf Klinkenberg, Ingo Mierswa, serta Simon Fischer pada Unit Artificial Intelligence dari Technical University of Dortmund. RapidMiner merupakan platform analisis modern yang meliputi data mining, mechine learning, analisis prediktif, text mining dan analisis bisnis [9]. Berikut area Kerja dari rapidminer diantaranya adalah view operator Merupakan view yang paling penting, semua operator atau langkah kerja dari rapidminer disajikan dalam bentuk kelompok hirarki di operator view ini sehingga operator-operator tersebut dapat digunakan pada proses analisis. ViewRepository View ini dapat Anda gunakan untuk mengelola dan menata proses Analisis Anda menjadi proyek dan pada saat yang sama juga dapat digunakan sebagai sumber data dan yang berkaitan dengan meta data. ViewParameters yaitu menu yang secara default akan tampil di bagian kanan tepatnya pada pojok atas yang ada dalam aplikasi rapidminer. Beberapa operator dalam RapidMiner membutuhkan satu atau lebih parameter agar dapat diindikasikan sebagai fungsionalitas yang benar. Namun terkadang parameter tidak mutlak dibutuhkan, meskipun eksekusi operator dapat dikendalikan dengan menunjukkan nilai parameter tertentu. Parameter view memiliki toolbar sendiri sama seperti view-view yang lain.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Adapun Jenis, Sifat dan Pendekatan pada Penelitian ini adalah

a. Identifikasi Masalah

Pada tahap ini dilakukan proses identifikasi masalah atas hal-hal/tema yang berhubungan dengan tweet para netter.

b. Analisis dan Pemetaan Masalah

Pada tahap ini analisis terhadap dataset serta model algoritma Naive Bayes, Decision Tree dan KNN yang akan diuji serta menghitung validitas nilainya menggunakan *cross validation*.

c. Studi Literatur

Pada tahap ini, penelitian dilakukan dengan mencari literatur-literatur yang berhubungan dengan penelitian terdahulu. Literatur tersebut dapat berupa Buku, Skripsi, Tesis, Jurnal, Proceeding, Hasil seminar baik tingkat lokal, nasional maupun international.

d. Preparasi Data

Pada tahap ini, dilakukan suatu persiapan data sebagai data uji yang akan digunakan dalam penelitian. Pada tahap ini pula data-data dari API Twitter tersebut dikumpulkan.

e. Pengelompokan Data

Pada tahap ini, dilakukan pengelompokan data dari API Twitter yang telah

dikumpulkan. Pengelompokan data sesuai dengan kriteria pemetaan masalah yang telah disusun.

f. Pemodelan Data

Pada tahap ini, dilakukan proses pemodelan data pada tools RapidMiner sehingga dapat dihitung nilai akurasi dan persentasenya.

g. Evaluasi

Pada tahap ini, dilakukan evaluasi atas data yang telah dimodelkan sebelumnya. Data-data yang telah dihitung kemudian dilakukan evaluasi untuk menghindari kesalahan perhitungan dan pengelompokan.

h. Hasil Evaluasi

Pada tahap ini, dilakukan dokumentasi atas hasil evaluasi untuk dijabarkan dalam bentuk statistik, tabel serta kesimpulan.

3.2. Metode Pengumpulan Data

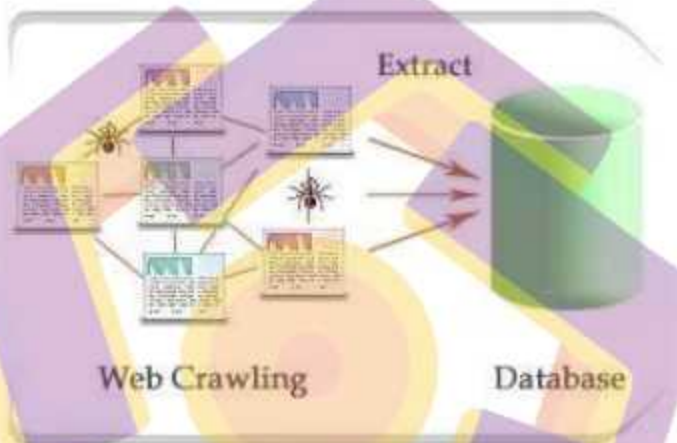
Adapun metode pengumpulan data pada penelitian ini adalah:

3.2.1 Crawling

Crawling adalah suatu teknik yang digunakan untuk mengumpulkan informasi yang ada dalam web. Crawling bekerja secara otomatis, dimana informasi yang dikumpulkan berdasarkan atas kata kunci yang diberikan oleh pengguna. Alat yang digunakan untuk melakukan crawling disebut dengan crawler(Anam et al., 2021).

Crawler berbentuk program yang diprogram dengan algoritma tertentu, sehingga dapat melakukan pemindaian ke halaman-halaman web, sesuai dengan alamat web ataupun kata kunci yang diberikan pengguna. Pada saat melakukan

pemindaian, crawler akan membaca teks yang ada, hyperlink dan berbagai tag yang digunakan di halaman web tersebut. Berdasarkan informasi tersebut, crawler akan mengindeks informasinya atau pun menyimpan informasi tersebut ke dalam sebuah file atau ke dalam database. Skema Web Crawling ditunjukkan pada gambar dibawah ini :



Gambar 3.1. Skema Web Crawling

Twitter crawling merupakan penerapan teknik crawling pada media sosial Twitter. Teknik ini bertujuan untuk mengumpulkan data tweet dari media sosial Twitter. Dari data yang telah dikumpulkan, dapat dianalisa dan diambil informasi penting yang bermanfaat bagi pengguna. Untuk menerapkan teknik twitter crawling ini, pihak Twitter telah memberikan akses bagi pengguna untuk memanfaatkan Twitter API. Sehingga dengan memanfaatkan Twitter API tersebut, pengguna bisa dengan mudah memperoleh data-data seperti tweet, data pengguna dan lain-lain. Untuk selanjutnya dikumpulkan dan disimpan dalam sebuah file atau basis data. Dalam penelitian ini Crawling ini akan dilakukan oleh

program RapidMiner dan hasil crawling akan disimpan kedalam file csv.

3.2.2 Twitter API

Twitter API, merupakan API (*Application Programming Interface*) yang disediakan oleh Twitter untuk memfasilitasi pengguna agar dapat berinteraksi dengan data-data yang ada pada Twitter. Data-data ini misalnya adalah tweet, id pengguna, lokasi, waktu pembuatan tweet dan lain-lain. Untuk memanfaatkan Twitter API, pengguna harus menggunakan bahasa server side scripting seperti php, python, Repidminer dan lain-lain(Purwokerto & Kunci, 2021).

Dengan menggunakan bahasa-bahasa tersebut, pengguna dapat melakukan request kepada Twitter API, dan respon hasilnya dirupakan dalam format JSON. Agar komunikasi pengguna dengan Twitter API aman, maka Twitter menerapkan *OAuth* atau *Open Authorization*. *OAuth* merupakan protokol terbuka yang memungkinkan pengguna untuk berbagi resource pribadi seperti foto, video, data pengguna dan lain-lain yang tersimpan di suatu situs web, dengan situs lain tanpa memberikan nama pengguna dan kata sandi pengguna tersebut. OAuth mengizinkan pengguna untuk memberikan akses kepada situs pihak ketiga untuk mengakses informasi mereka yang disimpan di penyedia layanan lain tanpa harus membagi izin akses atau keseluruhan data mereka. *OAuth* bergantung pada tiga set dari token dan secret yang didapat dari server dan client:

a. *Consumer key dan consumer secret*

Consumer key adalah *unique identifier* untuk client yang digunakan client untuk melakukan request untuk mendapatkan request tokens.

b. *Request token dan request token secret*

Request token adalah temporary *onetime identifier* yang diberikan oleh server untuk tujuan permintaan kepada user untuk melakukan grant permission pada client. *Token secret* digunakan untuk melakukan *sign request* untuk mendapatkan sebuah *access token.*

c. Access token dan access token secret

Access token adalah identifier untuk digunakan oleh client untuk melakukan akses ke resources milik user. Sebuah client bisa melakukan akses ke resources milik user selama token nya masih valid. Server bisa melakukan revoke kapanpun karena sudah expire atau user melakukan revoke secara manual. Secret digunakan untuk sign request ke resources yang di proteksi oleh akses user. Cara untuk mendapatkan consumer key, consumer secret, access token dan access secret adalah sebagai berikut:

1. Client melakukan request ke server menggunakan sebuah consumer key.
2. Client menggunakan consumer key untuk mendapatkan sebuah request token dan secret.
3. Client melakukan redirect pada user ke server untuk *grant permission* untuk client melakukan akses ke resources milik user. Proses ini bisa terjadi jika request token telah diautentikasi.
4. Client melakukan request ke server untuk memberikan access token dan secret.

Hasilnya merepresentasikan sebuah identifier dan shared secret yang client nya bisa gunakan untuk mengakses resources atas nama user. Ketika membuat

sebuah request untuk akses ke resource terproteksi, client menyertakan *Authorization header* yang berisi *consumer key, access token , signature method dan sebuah signature , timestamp* , sebuah *nonce* , dan untuk opsionalnya adalah versi dari OAuth yang digunakan(Sang et al., 2021). Dalam penelitian ini proses OAuth dijalankan melalui program RapidMiner.

3.3. Metode Analisis Data

Adapun metode analisis data pada penelitian ini adalah

a. Analisis konten

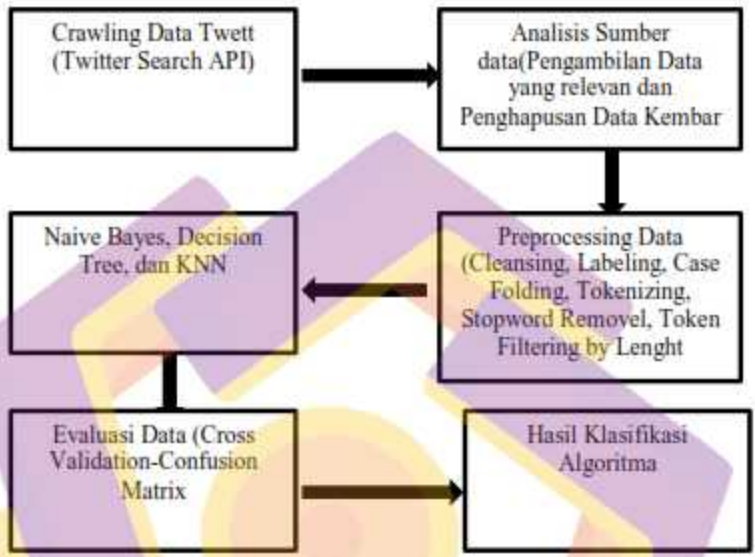
Metode ini membantu memahami keseluruhan komentar yang ada pada tweet. Metode ini membantu mengurai data tekstual yang ada agar dapat dikelompokkan menjadi kelompok sentimen.

b. Analisis naratif

Jenis metode ini berfokus pada cara bagaimana sebuah komentar/ide yang kemudian dikomunikasikan pada hal-hal yang terkait dengan PPPK Guru. Analisis ini menggunakan metode Naive Bayes, Decision Tree dan K-NN serta validasinya menggunakan metode *cross validation* yang metode *k-fold cross validation.*

3.4. Alur Penelitian

Berikut adalah alur penelitian yang dbuat:



Gambar 3.2. Alur Penelitian

Berikut tahapan dari alur penelitian adalah:

3.4.1 Crawling Data Tweet (Twitter search API)

Pada tahap ini dilakukan crawling data tweet dengan memanfaatkan Twitter search API (*Application Programming Interface*) untuk tweet yang berhubungan dengan Pelaksanaan PPPK. Langkah awal yang dilakukan adalah Instalasi tools RapidMiner serta membuat pola crawling data twitter dengan akun twitter yang telah terhubung. Setelah menghubungkan akun twitter tersebut, maka selanjutnya melakukan registrasi ke https://dev.twitter.com/apps/new untuk mendapatkan kode akses API twitter. *Crawling* ini dilakukan dengan menentukan kata kunci

yaitu PPPK GURU, atau apa yang sedang dicari serta menentukan jumlah tweet yang diinginkan. Setelah itu, data hasil *crawling* tersebut di *scrapping* dan disimpan dalam bentuk dokumen CSV.

3.4.2 Analisis Sumber Data (Penghapusan data kembar dan pemilahan data yang relevan)

Pada tahap analisis sumber data ini dilakukan proses Penghapusan data yang kembar atau sering disebut *remove duplicate*, agar menambah tingkat akurasi hasil penelitian ini nantinya. Selanjutnya pada pemilahan data yang relevan adalah melakukan pemilahan kolom atau atribut yang dibutuhkan pada penelitian ini dan membuang kolom yang tidak relevan serta membuang data yang tidak berkaitan dengan PPPK Guru. Berikut contohnya.

Tabel 3.1. Analisis sumber data

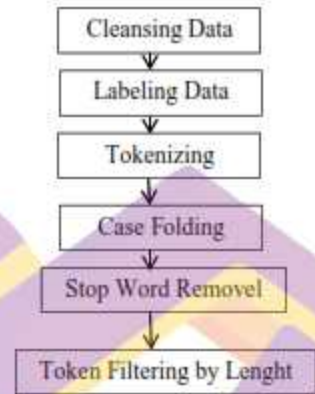
No	Text	Sentimen
1	Semua guru honorer bisa ikut PPPK Guru	Positif
2	PPPK Guru Tahap 2 Picu Migrasi Swasta ke Sekolah Negeri Ada Buktinya	Negatif
3	Peserta PPPK Guru Tahap 2 yang Tenang Ada Kabar Menyejukkan dari Kemendikbudristek	Poisitif
4	Jadwal Pengumuman Kelulusan PPPK Guru Tahap 2 Diubah Lagi Parah	Negatif
5	Semua guru honorer bisa ikut PPPK.	Positif
6	PPPK meningkatkan kesejahteraan guru.	Positif
7	PPPK itu tujuanya apa si pak apa mau membunuh sekolah swasta secara perlahan Banyak dr guru	Negatif

Tabel 3.1. Analisis sumber data (Lanjutan)

7	kami yg sangat berpengaruh dan berpotensi di sekolahan keterima PPPK Dan diambil smua ke sekolah negeri	Negatif

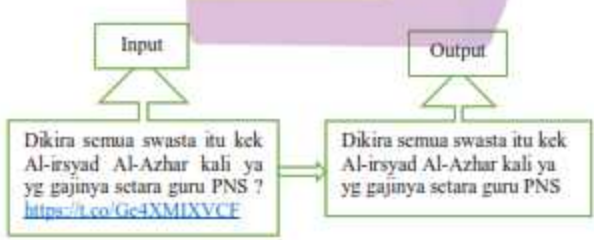
3.4.3 Preprocessing Data (Cleansing, Labeling, Case Folding, Tokenizing, Stop Word Removal, token filtering by lenght).

Pada tahap ini dilakukan proses ekstraksi dokumen, teks pada proses *text mining* ini memiliki resiko noise data yang cukup tinggi serta memiliki struktur teks bahasa Indonesia yang kurang baik. Dalam hal ini data yang berasal dari Twitter memiliki kerumitan yang cukup tinggi karena ketidaksesuaian dengan ejaan yang baku dan kesalahan pada ejaan yang cukup tinggi pada penulisan tweetnya(Sang et al., 2021). Cara memahami suatu teks adalah dengan terlebih dahulu menentukan fitur-fitur yang mewakili setiap kata untuk setiap fitur yang ada pada dokumen. Sebelum menentukan fitur-fitur yang mewakili diperlukan tahap Preprocessing yang dilakukan secara umum dalam *text mining* pada dokumen, yaitu cleansing, labeling, case folding, tokenizing, stop words removal, Token filtering by lenght(Purwokerto & Kunci, 2021). Berikut tahapan preprocessing data.



Gambar 3.3. Tahapan *Preprocessing*

a. Tahap *Cleansing* masih termasuk dalam *preprocess* text dimana dalam tahap ini dilakukan penghapusan semua RT, karakter html, atau web yang tidak lagi memiliki makna yang berhubungan dalam pengklasifikasian dokumen. Hal ini disebabkan karena terkadang sebuah tweet menyertakan suatu alamat web didalamnya, sehingga jika tidak dihapus akan menggangu proses klasifikasi. Pembersihan semua karakter html atau web ini seperti spesial karakter, URL link, username, serta *emoticon*, *emoticon* disini dihilangkan karena bisa menurunkan kualitas akurasi data yang signifikan(Imron, 2019). Contoh karakter yang dihapus seperti @,#,$,", & dan lain- lain. Berikut contoh hasil cleansing.



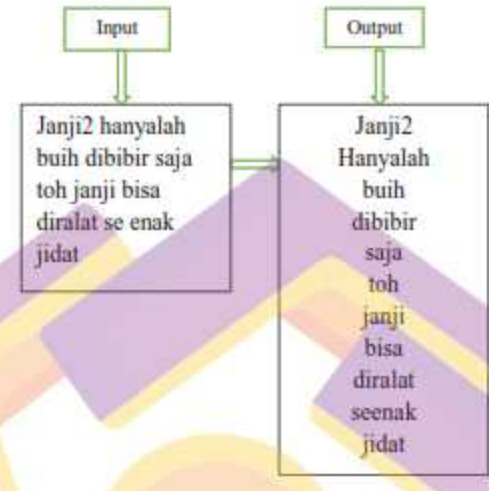
Gambar 3.4. Contoh Penerapan *Cleansing*

b. Labeling Pada tahap ini dilakukan proses labeling data dengan cara mengklasifikasikan data tweet ke dalam sentimen positif dan negatif. Proses ini dapat dilakukan secara manual maupun menggunakan aplikasi pihak ketiga. Berikut contoh hasil pelabelan secara manual dengan melibatkan ahli bahasa.

Tabel 3.2. Contoh Pelabelan Data Tweet

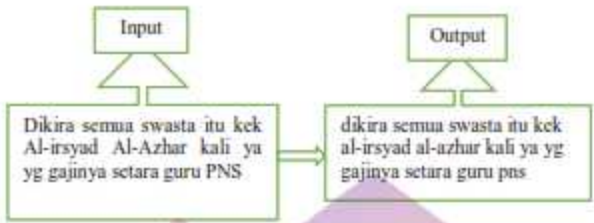
Pernyataan	Labeling
ALHAMDULILLAH Semua Guru Akan Gembira Baca Berita Ini Guru Honorer yang Lulus Passing Grade Langsung Jadi ASN PPPK.	Positif
Alhamdulillah ada kabar baik hari ini PPPK untuk guru bahasa daerah sudah ada.	Positif
Jelas ini perbuatan curang utk kasus guru honorer fiktif dia tdk pernah mengajar.	Negatif
Janji2 hanyalah buih dibibir saja toh janji bisa diralat se enak jidat.	Negatif

c. *Tokenizing* adalah sebuah proses pemotongan string input/tweet berdasarkan tiap kata yang menyusunnya. Proses ini memisahkan setiap kata yang menyusunnya. Secara umum sebuah kata dipisah dengan spasi. Oleh sebab itu karakter spasi menjadi karakter penting dalam penentuan pemisahan suatu kata. Pada saat bersamaan tokenizing juga berfungsi untuk membuang bagian karakter tertentu yang dianggap sebagai tanda baca(Imron, 2019). Berikut contoh hasil *tokenizing*.



Gambar 3.5. Hasil *Tokenizing*

d. *Case Folding* adalah proses merubah huruf kapital menjadi huruf kecil pada data tweet. Hal ini digunakan untuk mempermudah pencarian. Karena tidak semua tweet menggunakan huruf kapital. Oleh sebab itu lakukan keseragaman bentuk tulisan dalam model *lowercase*. Dalam tahap ini hanya huruf „a" sampai „z" yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter* seperti tanda baca dan angka. *Delimiter* merupakan karakter yang digunakan untuk membatasi atau memisahkan data yang disajikan dalam *plain text*. Berikut contoh hasil Case Folding.



Gambar 3.6. Contoh Case Folding

e. *Stop words Removal* adalah langkah penghilangan kata-kata yang tidak terkait dan berkontribusi pada topik data. Kata- kata yang termasuk dalam stop words tersebut memberikan pengaruh yang tidak baik pada proses text mining seperti pada kata-kata "bagaimana", "juga", "agar", "jadi" dan lain-lain (Imron, 2019). Pada proses stop words ini dibutuhkan kamus stop words untuk mendata kata-kata apa saja yang harus dibuang.

Gambar 3.7. Contoh Stop Word Removal

f. *Token Filtering (By Lenght)* yaitu membuang kata yang kurang dari tiga huruf. Berikut contoh hasil token filtering.

Gambar 3.8. Contoh Token Filtering

3.4.4. Cross Validation (K-fold cross validation) dan Confusion Matrix

Salah satu cara mengatahui kinerja model adalah dengan mengukur akurasinya (meskipun akurasi bukan satu-satunya parameter yang digunakan untuk mengukur kinerja suatu model). *Cross validation* adalah suatu metode tambahan dari teknik data mining yang bertujuan untuk memperoleh hasil akurasi yang maksimal metode ini juga sering disebut *k-fold cross validation* dimana percobaan sebanyak k kali untuk satu model dengan parameter yang sama.(Suryono et al., 2018). Fungsi dari penggunaan cross validation adalah:

a. Untuk mengatahui performa dari suatu model algoritma dengan melakukan percobaan sebanyak *k* kali.
b. Untuk meningkatkan tingkat performance dari model tersebut.
c. Untuk mengolah dataset dengan kelas yang seimbang.

Dalam kasus klasifikasi ada yang perlu diperhatikan dalam pembagian set data ke sejumlah *k* partisi, yaitu harus melakukan *stratification* yang artinya kita

akan mempartisi atau membagi set data tersebut ke *k* partisi dengan komposisi kelas yang seimbang disetiap partisinya. Cross validation merupakan teknik validasi dari pengembangan model split validation dimana cross validation mampu bekerja dengan cepat dengan pengambilan sampel yang lebih tersetruktur, jadi dalam jumlah pengujian beberapapun set data latih dan set data uji akan diambil dengan data yang berbeda dengan percobaan atau literasi sebelumnya. Yang nantinya hasil percobaan tersebut akan dicatat nilai evaluasi ferporma darri sebuah model dengan menggunakan confussion matrix. Dalam beberapa penelitian yang sudah dilakukan oleh pakar pakar data mining, model pengujian atau validasi model dari suatu algoritma klasifikasi, *cross validation* lebih sering dipakai ketimbang *split validition* karna model validasi dengan menerapkan *10-cross validation* sudah merupakan standar dan suatu metode validasi yang canggih atau lebih praktis dan efisien serta mempu meningkatkan sedikit nilai performansinya. Confusion Matrix. Pengukuran akurasi algoritma Pengukuran akurasi merupakan tahapan untuk membuktikan tingkat performa suatu algoritma terhadap dataset yang digunakan. Dalam penelitian ini digunakan confusion matrix sebagai alat ukur performa algoritma klasifikasi. Confussion matrix atau matrik kebingungan merupakan sebuah perhitungan yang membandingkan dataset dengan hasil klasifikasi sesuai dengan data sebenarnya dengan jumlah keseluruhan data. Hasil akhir dari matrik ini adalah tingkat akurasi dengan satuan persen (%). Tingkat akurasi ini yang nantinya dijadikan acuan para peneliti terkait performa algoritma klasifikasi tersebut. Confusion Matrix adalah evaluasi dari sebuah klasifikasi data mining yang direpresentasikan menjadi tabel [22]

Convussion matrix berisi informasi perbandingan label hasil klasifikasi dengan label sebenarnya. Berikut confusion matrik dengan rafidminer.

Gambar 3.9. Confussion matrix model dalam aplikasi rapid miner

Dalam data training digunakan algoritma KNN dan dalam testing digunakan apply model serta prformance untuk menghasilkan matrik kebingungan atau confussion matrix.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Crawling Data Tweet (Twitter search API)

Pada tahap ini dilakukan crawling data atau pengumpulan data selama bualn januari 2022 yaitu dari tanggal 1 sampai 31 januari 2022 dengan kata kunci PPPK Guru dengan menggunakan tools RapidMiner yang telah terhubung dengan akun twitter. Pengumpulan data pada sosial media twitter dengan jumlah data yang kita dapatkan sebesar 871 data, kemudian melakukan proses *filtering* data dengan operator *select attributs* yang digunakan untuk memilih text pada data karna attributs lain tidak dibutuhkan. Berikut ini contoh hasil *crawling* pada RapidMiner ditunjukkan pada gambar 4.1 dibawah ini :

Row No.	"Text"	"Id"
1	[illegible] Kalo ASN [illegible]	"1472[illegible]...
2	[illegible] guru [illegible] PNS [illegible] pppk [illegible]	"1472[illegible]...
3	Forum Honorer PGRI Jatim Minta Formasi Pengangkatan ASN...	?
4	?	"1472486801...
5	Kabar Menyejukkan hari ini bagi peserta pppk guru 2021 tahap ...	"1472[illegible]...
6	PPPK Guru Tahap 2 [illegible] ke Sekolah Negeri [illegible]...	?
7	Nilai 717 tetapi Tidak Lolos Formasi PPPK Guru Tahap 2 [illegible]...	?
8	?	"1472477011...
9	Ketua Paguyuban GTT/PTT Kabupaten Kebumen [illegible]...	?

ExampleSet (871 examples, 0 special attributes, 2 regular attributes)

Gambar 4.1. Hasil Crawling data Twitter dari RapidMiner

Untuk diketahui bahwa crawling twitter ini hanya dapat menarik data selama 1 minggu kebelakang sehingga apabila ingin menarik data lebih banyak secara berkala maka proses crawling dapat dilakukan secara harian. Dalam pengujian ini data yang telah dicrawling disimpan kedalam file csv.

4.2 Analisis Sumber Data

Pada tahap ini akan dilakukan penghapusan data kembar dan pemilahan kolom data yang relevan dengan kebutuhan penelitian ini. Tahap ini berguna untuk memastikan bahwa data yang akan kita olah merupakan data yang sudah relevan dengan studi kasus yang diambil dalam penelitian ini yaitu sentimen analisis terhadap pelaksanaan PPPK Guru.

4.2.1 Penghapusan data kembar dan pemilahan data yang relevan

Tahap pengambilan data yang relevan disini dimaksudkan adalah membuang data hasil crawling yang memiliki kesamaan kalimat dan kemudian memilah kolom apa saja yang dibutuhkan pada penelitian ini serta menghapus data yang kosong serta menghapus data yang tidak berkaitan dengan PPPK guru. Pada tahap ini digunakan komponen operator *Remove Duplicates* dan *Select Attributes*. Dari data 871 setelah dihapus data yang duplicate atau ganda, menghapus data yang tidak berkaitan dengan PPPK guru, dan menghapus data yang kosong maka hasilnya menjadi 519 data.

Gambar 4.2. Remove duplicate dan Select attributes

4.2.2 Cleansing

Proses ini merupakan tahap penghapusan karakter-karakter yang mengiringi kata-kata tersebut yang hilangnya karakter-karakter tersebut tidak mengurangi makna dari penelitian ini. Contoh dari karakter-karakter yang ikut terbuang dari penelitian ini adalah tanda kurung "(", petik dua ", url, #, titik dua

":", tanda titik ".", tanda "@", dan lain-lain. Untuk itu ditambahkan operator sub proses dimana dibagian parameter kita isi dengan replace RT, URL, hastag, mention dan symbol setelah komponen select attribut.

4.2.3 Labeling

Tahap labeling ini adalah tahap pemberian klasifikasi kalimat positif dan negative. Pada tahap ini dilakukan pemberian label positif dan negatif ini baik secara otomatis maupun secara manual. Pemberian label secara otomatis dilakukan oleh operator dari pihak ketiga yaitu *Rosette Text Analytics*. Sedangkan secara manual juga masih dilakukan untuk mencegah kesalahan otomatisasi pelabelan terkait atas makna bahasa yang digunakan serta memastikan jumlah bentuk sentimen sesuai dengan skenario penelitian.

Selanjutnya hasil rangkaian proses crawling hingga tersimpan di file csv. Berikut contoh hasil crawling, cleansing dan labeling ditunjukkan oleh gambar 4.3.

Gambar 4.3. Hasil proses Crawling, cleaning dan labeling

4.3. PreProcessing Data

Dalam tahap ini dilakukan proses case folding, tokenizing, stop word removal. Proses ini berguna untuk membantu menghasilkan data yang berkualitas

untuk dapat diklasifikasikan ke dalam 3 algoritma klasifikasi nantinya Adapun langkah - langkah pada tahap ini adalah :

1. Case Folding

Proses yang dilakukan pada tahap ini adalah merubah huruf kapital menjadi huruf kecil. Proses case folding ini memastikan bahwa format huruf yang akan diklasifikasikan tidak memiliki perbedaan. Hal ini membantu kualitas pengklasifikasian kata yang digunakan. Implementasi case folding ini dilakukan setelah data berbentuk csv.

2. Tokenizing

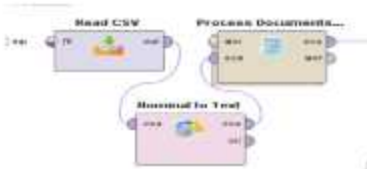
Proses ini memisahkan kata-kata yang telah dibersihkan dari karakter dan kata-kata yang tidak berhubungan dengan penelitian ini menjadi kata-kata yang tersendiri. Dalam tahap ini proses diawali dari proses import data dari csv ke dalam aplikasi RapidMiner.

3. Stop Words Removal

Pada proses ini dilakukan penghapusan atas kata-kata yang ketiadaan fungsinya tidak membuat berkurangnya makna & arti dari penelitian yang dilakukan. Contoh kata-kata yang dibuang dalam penelitian ini adalah yang, maka, itu, daripada, begitu, ini, dan lain-lain. Pada penelitian ini filter stop words dilakukan dengan menggunakan file kamus *Stopwords* yang penulis ambil dari kaggle .

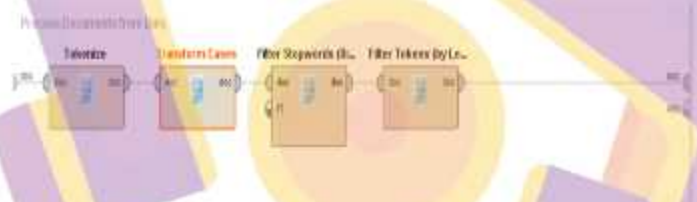
4. Token Filtering Yaitu menghapus kata-kata yang kurang dari tiga huruf.

Berikut gambar proses preprocessing data.



Gambar 4.4. Nominal to text

Data training berupa csv yang akan di*convert* menjadi *Nominal to text* untuk mengubah sebuah tipe data yang ada di setiap kalimat, kemudian melalui tahap *Process Document From Data* yang berfungsi untuk mengolah data dari *transform case, Tokenize, Filter Token,*dan *Filter Stopword,*. Berikut gambar proses dokument yang diolah.



Gambar 4.5. Proses Document From Data

4.4 Klasifikasi dan Akurasi

Membuat suatu klasifikasi dan akurasi atas hasil preprocessing menggunakan Algoritma Naive Bayes, Decission Tree dan K-NN.

4.4.1 Klasifikasi pada Naïve Bayes.

Pada proses pengklasifikasian text mining menggunakan Algoritma Naïve Bayes ini dibutuhkan langkah-langkah sebagai berikut : Persiapan dokumen yang telah berlabel (Positif dan Negatif) Pada langkah pertama ini disiapkan data dokumen yang telah berlabel sentiment yang sudah dilakukan pada proses awal penelitian ini sebagaimana yang ditunjukkan pada tabel 4.1. dibawah ini:

Tabel 4.1 Contoh dokumen yang sudah dilabeli

Pernyataan	Labeling
ALHAMDULILLAH SELURUH Guru Bisa Tenang	Positif
ALHAMDULILLAH Kabar Baik	Positif
Pekerjaan paling menyebalkan: menunggu Kapan ya proses PPPK Guru bisa gak molor Dari awal pelaksanaan selalu molor Padahal jelas dilaksanakan saat pandemi jadi alasannya apa sampai bisa molor Kalau memang butuh waktu lebih kenapa ngebuat jadwalnya bisa ga sesuai.	Negatif
Mentang mentang sekolah ini pelosok dinas pendidikan seakan tutup mata akan kualitas sekolah Mau melaporkan di laporan kampus mengajar tp laporannya wajib direview dulu sama guru dan dinas pendidikan setempat jadi percuma juga mau pppk cpns kalau masih ada yang nakal yang sama	Negatif

Data yang telah dilabeli dan dilakukan preprocessing kemudian hasil preprocessing akan dibagi menjadi dua yaitu data latih dan data uji. Dalam kasus klasifikasi ada yang perlu diperhatikan dalam pembagian set data ke sejumlah *k* partisi, yaitu harus melakukan *stratification* yang artinya kita akan mempartisi atau membagi set data tersebut ke *k* partisi dengan komposisi kelas yang seimbang disetiap partisinya. *Cross validation* merupakan teknik validasi dari pengembangan model *xplit validation* dimana *cross validation* mampu bekerja dengan cepat dengan pengambilan sampel yang lebih tersetruktur, jadi dalam jumlah pengujian beberapapun set data latih dan set data uji akan diambil dengan data yang berbeda dengan percobaan atau literasi sebelumnya. Yang nantinya hasil percobaan tersebut akan dicatat nilai evaluasi ferporma dari sebuah model dengan menggunakan *confussion matrix*. Dalam beberapa penelitian yang sudah

dilakukan oleh pakar pakar data mining, model pengujian atau validasi model dari suatu algoritma klasifikasi, *cross validation* lebih sering dipakai ketimbang *split validition* karna model validasi dengan menerapkan *10-cross validation* sudah merupakan standar dan suatu metode validasi yang canggih atau lebih praktis dan efisien serta mempu meningkatkan sedikit nilai performansinya. Berikut gambar *cross validation* dengan *repidminer* dimana jumlak number of fold pada parameter adalah 10.

Gambar 4.6. *Cross Validation*

Pada gambar 4.6. dengan menggunakan repidminer, model pengujian atau validasi model dari suatu algoritma klasifikasi menerapkan *10-cross validation* sehingga pada parameters disebelah kanan number of folds diisi dengan 10, karena akan dilakukan 10 kali pengujian. Dimana pada repidminer split dilakukan secara otomatis dengan validasi silang.

Selanjutnya kita lihat *cross validation* dengan algoritma naive bayes.

Gambar 4.7. *Cross validation* dengan *Naive Bayes*

Berikut hasil akurasi dan prediksi naive bayes. Accuracy merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Akurasi menjawab pertanyaan berapa persen sentimen yang benar diprediksi positif dan negatif dari keseluruhan data tes. Berikut gambar hasil akurasi dari metode naive bayes:

Akurasi = (TP + TN) / (TP + FP + FN + TN)

Akurasi = (240 + 152) / (240 + 88 + 39 + 152)

= 392 / 519

= 0.75

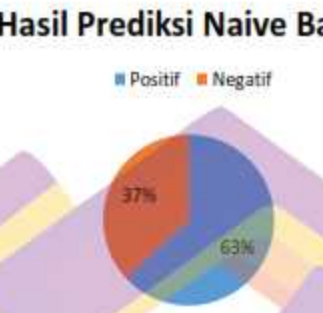
Berikut tabel hasil akurasi algoritma Naïve Bayes

Tabel 4. 2 Hasil akurasi Naïve Bayes

Prediction	True Negtif	True Positif	Class Precision
Pred.Negatif	152	39	79,58%
Pred. Positf	88	240	73,17%
Class Recall	63,33%	86,02%	
Akurasi = 75,53% +/- 5,38%			

Dari data diatas analisis menggunakan aplikasi repidminer dengan algoritma naive bayes didapatkan hasil dengan tingkat akurasi 75,53%. Dari 519 data yang ada, hasil prediksi naive bayes yaitu 328 diantaranya merupakan tanggapan masarakat yang memiliki sentimen positif terhadap pelaksaan PPPK Guru, sedangkan sisanya 191 merupakan tanggapan masarakat yang memiliki

sentimen negatif terhadap pelaksaan PPPK Guru. Berikut grafik hasil prediksi naive bayes:



Gambar 4.8. Prediksi Naive Bayes

Selanjutnya kita akan melakukan perhitungan manual dengan menggunakan Ms. Excel, bagaimana cara kerja algoritma Naive Bayes dengan rumus pada equation (2.1). Hasilnya adalah sebagai berikut.

Tabel 4.3. Perhitungan Manual Naive Bayes

Text	Posittif	Negatif
Sentimen	54%	46%
Sentimen	279	240

Dari semua data diperoleh data yang Positif berjumlah 279 atau sebanyak 45% dan yang negatif berjumlah 240 atau 46%. Untuk mendapatkan jumlah data positif dan negatif manualnya dengan menggunakan Formula " =COUNTIF(semua data sentimen, data bernila positif)". Dan " =COUNTIF(semua data sentimen, data bernilai negatif").

Tabel 4.4. Hasil Prediksi Manual

Text	Positif	Negatif
Sentimen	29%	21%
Sentimen	0.29	0.21

Sedangkan pada kelas prediksi menggunakan formula "=VLOOKUP(Text sentimen,hasil data bernilai Positif (yakni 54% atau279), 2 (Menggunakan angka 2 karena nila data sentimen bernila positif berada pada kolom ke – 2) * VLOOKUP((Text sentimen,hasil data bernilai Positif (yakni 54% atau279), 2 (Menggunakan angka 2 karena nila data sentimen bernila positif berada pada kolom ke – 2). Diperoleh hasil Class Predition sentimen bernila positif sebanyak 29% atau 0.29 dan hasil prediksi sentimen bernilai negatif menggunakan formula "=VLOOKUP(Text sentimen,hasil data bernilai Negatif (yakni 46% atau 240), 3 (Menggunakan angka 3 karena nila data sentimen bernila Negatif berada pada kolom ke – 3) * VLOOKUP((Text sentimen,hasil data bernilai Negatif (yakni 46% atau 240), 3 (Menggunakan angka 3 karena nila data sentimen bernila positif berada pada kolom ke – 3). sebanyak 21% atau 0.21. Hasil dari Perhitungan dengan Algoritma Naïve Bayes adalah : Hasil dari perhitungan % pada perhitungan manual yakni 54% ditambahkan dengan Hasil pada Class Predicion sebesar 21% = 75%.

4.4.2. Klasifikasi pada decision tree

Pada proses pengklasifikasian text mining menggunakan Algoritma Decision tree ini dibutuhkan langkah-langkah sebagai berikut : Persiapan dokumen yang telah berlabel (Positif dan Negatif) Pada langkah pertama ini disiapkan data

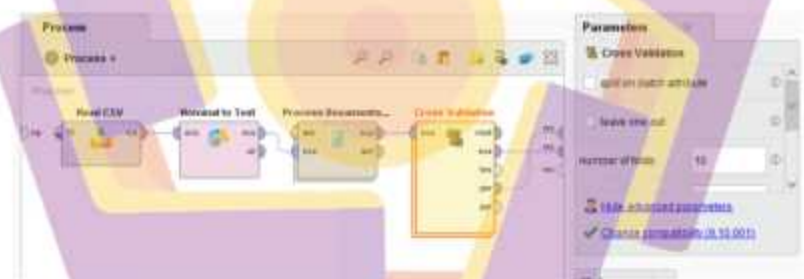
dokumen yang telah berlabel sentiment yang sudah dilakukan pada proses awal penelitian ini sebagaimana yang ditunjukkan pada tabel 4.4 dibawah ini:

Tabel 4.5. Contoh dokumen yang sudah dilabeli

Pernyataan	Labeling
ALHAMDULILLAH SELURUH Guru Bisa Tenang	Positif
ALHAMDULILLAH Kabar Baik	Positif
Pekerjaan paling menyebalkan: menunggu Kapan ya proses PPPK Guru bisa gak molor Dari awal pelaksanaan selalu molor Padahal jelas dilaksanakan saat pandemi jadi alasannya apa sampai bisa molor Kalau memang butuh waktu lebih kenapa ngebuat jadwalnya bisa ga sesuai.	Negatif
Mentang mentang sekolah ini pelosok dinas pendidikan seakan tutup mata akan kualitas sekolah Mau melaporkan di laporan kampus mengajar tp laporannya wajib direview dulu sama guru dan dinas pendidikan setempat jadi percuma juga mau pppk cpns kalau masih ada yang nakal yang sama	Negatif

Data yang telah dilabeli dan dilakukan preprocessing yang selanjutnya akan dibagi menjadi dua yaitu data latih dan data uji. Dalam kasus klasifikasi ada yang perlu diperhatikan dalam pembagian set data ke sejumlah *k* partisi, yaitu harus melakukan *stratification* yang artinya kita akan mempartisi atau membagi set data tersebut ke *k* partisi dengan komposisi kelas yang seimbang disetiap partisinya. *Cross validation* merupakan teknik validasi dari pengembangan model *split validation* dimana *cross validation* mampu bekerja dengan cepat dengan

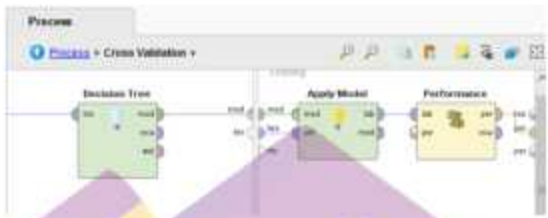
pengambilan sampel yang lebih tersetruktur, jadi dalam jumlah pengujian beberapapun set data latih dan set data uji akan diambil dengan data yang berbeda dengan percobaan atau literasi sebelumnya. Yang nantinya hasil percobaan tersebut akan dicatat nilai evaluasi ferporma darri sebuah model dengan menggunakan *confussion matrix*. Dalam beberapa penelitian yang sudah dilakukan oleh pakar pakar data mining, model pengujian atau validasi model dari suatu algoritma klasifikasi, *cross validation* lebih sering dipakai ketimbang *split validition* karna model validasi dengan menerapkan *10-cross validation* sudah merupakan standar dan suatu metode validasi yang canggih atau lebih praktis dan efisien serta mempu meningkatkan sedikit nilai performansinya. Berikut gambar *cross validation* pada *repidminer*.



Gambar 4.9. *Cross Validation*

Pada gambar 4.9. dengan menggunakan repidminer, model pengujian atau validasi model dari suatu algoritma klasifikasi menerapkan *10-cross validation* sehingga pada parameters disebelah kanan number of folds diisi dengan 10, karena akan dilakukan 10 kali pengujian. Dimana pada repidminer split dilakukan secara otomatis dengan validasi silang.

Selanjutnya kita akan melakukan *cross validation* dengan algoritma *decision tree*. Berikut gambarnya



Gambar 4.10. Cross Validation dengan Decision Tree

Selanjutnya kita akan lihat hasil akurasi dari algoritma decision tree. Accuracy merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Akurasi menjawab pertanyaan berapa persen sentimen yang benar diprediksi positif dan negatif dari keseluruhan data tes.

Akurasi = (TP + TN) / (TP + FP + FN + TN)

Akurasi = (123 + 198) / (123 + 42 + 156 + 198)

= 321/519

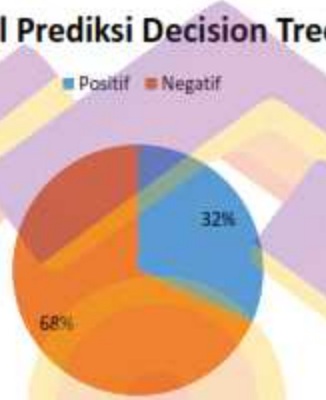
= 0.61

Tabel 4.6. Hasil Akursi Decision Tree

Prediction	True Negtif	True Positif	Class Precision
Pred.Negatif	198	156	55,93%
Pred. Positf	42	123	74,55%
Class Recall	82,50%	44,09%	
Akurasi = 61,85% +/- 3,46%			

Analisis menggunakan aplikasi repidminer dengan algoritma Decision Tree didapatkan hasil dengan tingkat akurasi 61,85%.

Dari 519 data yang digunakan 165 diantaranya merupakan tanggapan masarakat yang memiliki sentimen positif terhadap pelaksaan PPPK Guru, sedangkan sisanya 354 merupakan tanggapan masarakat yang memiliki sentimen negatif terhadap pelaksaan PPPK Guru. Berikut grafik prediksi decision tree:

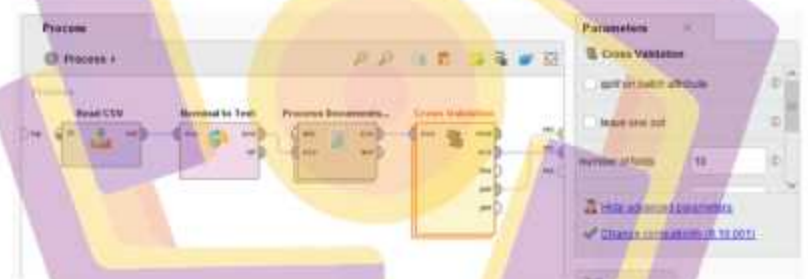


Gambar 4.11. Hasil Prediksi decision Tree

4.4.3. Klasifikasi pada KNN

Data yang telah dilabeli dan dilakukan preprocessing yang selanjutnya akan dibagi menjadi dua yaitu data latih dan data uji. Dalam kasus klasifikasi ada yang perlu diperhatikan dalam pembagian set data ke sejumlah *k* partisi, yaitu harus melakukan *stratification* yang artinya kita akan mempartisi atau membagi set data tersebut ke *k* partisi dengan komposisi kelas yang seimbang disetiap partisinya. *Cross validation* merupakan teknik validasi dari pengembangan model *split validation* dimana *cross validation* mampu bekerja dengan cepat dengan pengambilan sampel yang lebih tersetruktur, jadi dalam jumlah pengujian beberapapun set data latih dan set data uji akan diambil dengan data yang berbeda

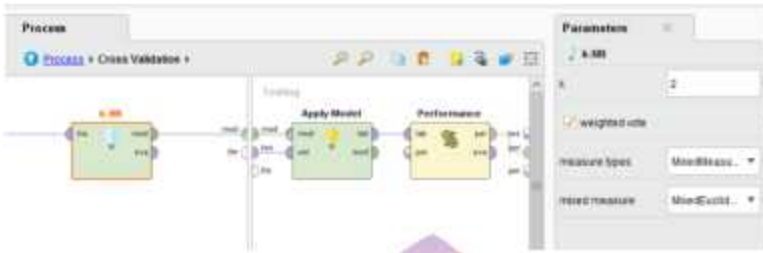
dengan percobaan atau literasi sebelumnya. Yang nantinya hasil percobaan tersebut akan dicatat nilai evaluasi ferporma darri sebuah model dengan menggunakan *confussion matrix*. Dalam beberapa penelitian yang sudah dilakukan oleh pakar pakar data mining, model pengujian atau validasi model dari suatu algoritma klasifikasi, *cross validation* lebih sering dipakai ketimbang *split validition* karna model validasi dengan menerapkan *10-cross validation* sudah merupakan standar dan suatu metode validasi yang canggih atau lebih praktis dan efisien serta mempu meningkatkan sedikit nilai performansinya. Dalam hal ini penulis menggunakan tool repidminer, dengan jumlah k=10. Berikut gambar *cross validation* di *repidminer*.



Gambar 4.12. Cross Validation

Pada gambar 4.12. dengan menggunakan repidminer, model pengujian atau validasi model dari suatu algoritma klasifikasi menerapkan *10-cross validation* sehingga pada parameters disebelah kanan number of folds diisi dengan 10, karena akan dilakukan 10 kali pengujian. Dimana pada repidminer split dilakukan secara otomatis dengan validasi silang.

Selanjutnya kita akan lihat gambar cross validation dengan algoritma KNN dengan jumlah K yang akan penulis gunakan adalah 2.



Gambar 4.13. Cross Validation

Pada gambar 4.13. yaitu cross validation dengan KNN dimana model KNN didasarkan pada jarak tetangga terdekat sebagai nilai prediksi dari instance yang baru, pada penelitian ini nilai *k* ditentukan k=3, yang bisa kita ubah di parameter di sebelah kanan.

. Berikut hasil akurasi dan prediksi dari KNN dengan jumlah K=2. Accuracy merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Akurasi menjawab pertanyaan berapa persen sentimen yang benar diprediksi positif dan negatif dari keseluruhan data tes.

Akurasi = (TP + TN) / (TP + FP + FN + TN)

Akurasi = (228 + 153) / (228 + 87 + 51 + 153)

= 381 / 519

= 0.73

Tabel 4.7. Hasil Akurasi KNN

Prediction	True Negtif	True Positif	Class Precision
Pred.Negatif	153	51	75,00%
Pred. Positf	87	228	72,38%
Class Recall	63,75%	81,72%	
Akurasi = 73,41% +/- 5.26%			

Analisis menggunakan aplikasi repidminer dengan algoritma KNN didapatkan hasil dengan tingkat akurasi 73,41%.

Data yang diproleh dari twetter dilakukan preprocessing dan juga dilabeli secara manual dengan dua kategori yaitu positif dan negatif, yang kemudian dilakukan perbandingan hasil klasifikasi dengan hasil prediksi dari tiga algoritma yaitu naive bayes, decision tree dan KNN. Hasil prediksi dari metode KNN dapat dilihat pada grafik dibawah ini. Berikut grafik hasil prediksi KNN

Gambar 4.14. Grafik Hasil Prediksi KNN

Berdasarkan dari gambar diatas dari 519 data yang digunakan 315 diantaranya merupakan tanggapan masarakat yang memiliki sentimen positif

terhadap pelaksaan PPPK Guru, sedangkan sisanya 204 merupakan tanggapan masarakat yang memiliki sentimen negatif terhadap pelaksaan PPPK Guru. Selanjutnya kita akan melakukan proses perhitungan manual dengan menggunakan MS.Excel, berikut hasinya.

Tabel 4.8. Hasil Manual KNN

Text	Positif	Negatif
Sentimen	54%	46%
Sentimen	279	240

Dari semua data diperoleh data yang Positif berjumlah 279 atau sebanyak 45% dan yang negatif berjumlah 240 atau 46%. Untuk mendapatkan jumlah data positif dan negatif manualnya dengan menggunakan Formula " =COUNTIF(semua data sentimen, data bernila positif)". Dan " =COUNTIF(semua data sentimen, data bernilai negatif").

Setelah diperoleh nilai manual data maka dicari nilai Distance dari sentimen data positif dan negatifnya dengan menggunakan formula "=SQRT(Hasil data yang bernilai positif (yakni 54% atau 279) dipangkat 2 atau kali 2") . Dan untuk data Distance Negatif dengan menggunakan "=SQRT(Hasil data yang bernilai negatif (yakni 46% atau 240) dipangkat 2 atau kali 2"). Dari Formula tersebut diperoleh hasil Distance sentimen Positif dan Negatif sebagai berikut :

Tabel 4.9. Hasil Perhitungan Manual KNN dengan K=2

KNN	
Distance	K=2
Positif	40,42
Negatif	32,99
Total	73,41

Berikut tabel perbandingan hasil akurasi dari ketiga algoritma yang kita gunakan yaitu naive bayes,decision tree dan KNN.

Tabel 4.10. Perbandingan Hasil Naive Bayes, Decision Tree dan KNN

Algoritma	**Accuracy**
Naive Bayes	75,53%
Decision Tree	61,85%
KNN	73,41%

Berdasarkan tabel diatas, algoritma pertama yang digunakan untuk mengklasifikasikan sentimen masyarakat terhadap pelaksanaan PPPK Guru pada media sosial twitter adalah naive bayes dengan hasil akurasi sebesar 75,53%. Hasil tersebut merupakan akurasi tertinggi, oleh karena itu naive bayes sangat cocok untuk mengklasifikasikan data tentang analisis sentimen masyarakat terhadap pelaksanaan PPPK Guru. Algoritma selanjutnya adalah decision tree, pada algoritma ini hasil akurasi yang diproleh adalah 61,85%, dimana hasil akurasi decision tree berada dibawah naive bayes. Algoritma yang terakhir untuk mengklasifikasikan sentimen masyarakat terhadap pelaksanaan PPPK Guru pada

media sosial twitter adalah K-nearst neighbor atau KNN dengan nilai K=2, hasil akurasi dari algoritma ini adalah 73,41%, dimana hasil ini berada dibawah naive bayes. Berikut grafik hasil Naive Bayes, Decision Tree dan KNN di tujukkkan oleh gambar dibawah ini.

Gambar 4.15. Hasil Naive Bayes, Decision Tree dan KNN

Gambar 4.15 menampilkan grafik hasil akurasi dari tiga algoritma yaitu naïve bayes, decision tree dan KNN, algoritma pertama yang digunakan untuk mengklasifikasikan sentimen masyarakat terhadap pelaksanaan PPPK Guru pada media sosial twitter adalah naive bayes dengan hasil akurasi sebesar 75,53%. Hasil tersebut merupakan akurasi tertinggi, oleh karena itu naive bayes sangat cocok untuk mengklasifikasikan data tentang analisis sentimen masyarakat terhadap pelaksanaan PPPK Guru. Algoritma selanjutnya adalah decision tree, pada algoritma ini hasil akurasi yang diproleh adalah 61,85%, dimana hasil akurasi decision tree berada dibawah naive bayes. Algoritma

yang terakhir untuk mengklasifikasikan sentimen masyarakat terhadap pelaksanaan PPPK Guru pada media sosial twitter adalah K-nearst neighbor atau KNN dengan nilai K=2, hasil akurasi dari algoritma ini adalah 73,41%, dimana hasil ini berada dibawah naive bayes.

BAB V
PENUTUP

5.1. Kesimpulan

Berdasarkan hasil dan pembahasan diatas dapat disimpulkan bahwa dari 519 data yang telah dilabeli secara manual dan dilakukan preprocessing yang selanjutnya diklasifikasikan dan dilakukan validasi data dengan *k-fold cross validation* dan *confusion matrik* dengan jumlah *k=10*. Hasil prediksi naive bayes yaitu 328 data bersentimen positif dan 191 data bersentimen negatif, Selanjutnya hasil prediksi decision tree yaitu 165 data bersentimen positif dan 354 data bersentimen negatif dan yang terakhir yaitu hasil prediksi dari KNN yaitu 315 data bersentimen positif dan 204 data bersentimen negatif. Analisis sentimen masyarakat terhadap pelaksanaan PPPK guru pada media sosial Twitter dengan algoritma naive bayes mencapai tingkat akurasi 75,53%. Decision Tree tingkat akurasi mencapai 61,85%. Dan yang terakhir adalah algoritma KNN mencapai akurasi 73,41%. Dalam penelitian ini, dapat diketahui bahwa metode Naive Bayes adalah metode yang tingkat akurasinya lebih tinggi dibandingkan kedua metode lainnya dengan tingkat akurasi sebesar 75,53%.

5.2. Saran

a. Data yang digunakan dalam penelitian ini masih sangat sedikit, diharapkan kedepannya bisa menggunakan lebih banyak data.
b. Hasil akurasi yang didapatkan dengan algoritma naïve bayes, Decision Tree dan KNN masih rendah sehingga perlu dilakukan penelitian dengan algoritma yang lain.

c. Pelabelan dalam penelitian ini dilakukan secara manual diharapkan kedepannya dicoba dengan pelabelan tidak manual.

d. Hasil akurasi dalam penelitian ini masih sangat rendah perlu dilakukan penelitian lagi utk meningkatkan akurasi.

DAFTAR PUSTAKA

PUSTAKA BUKU

Kusrini, Luthfi, E. T., 2009, *Algorima Data Mining*. Yogyakarta: Andi Offset

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Agustiningsih, K. K., & Utami, E. (2020). *Sentiment Analysis Towards COVID-19 Vaccine on Twitter Social Media : Systematic Literature Riview*.

Agustiningsih, K. K., Utami, E., Muhammad, O., & Alsyaibani, A. (2021). *Sentiment Analysis and Topic Modelling of The COVID-19 Vaccine in Indonesia on Twitter Social Media Using Word Embedding*. *7*(3), 1–12. https://doi.org/10.26555/jiteki.v7i3.xxxx

Agustiningsih, K. K., Utami, E., Muhammad, O., & Alsyaibani, A. (2022). Sentiment Analysis of COVID-19 Vaccines in Indonesia on Twitter Using Pre-Trained and Self-Training Word Embeddings. *Jurnal Ilmu Komputer Dan Informasi (Journal of Computer Science and Information)*, *15*(1), 39–46.

Anam, M. K., Pikir, B. N., Firdaus, M. B., & Erlinda, S. (2021). *Penerapan Na ¨ ıve Bayes Classifier , K-Nearest Neighbor (KNN) dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah Applications of Na ¨ ıve Bayes Classifier , K-Nearest Neighbor and Decision Tree to Analyze Sentiment on Ne*. *21*(1). https://doi.org/10.30812/matrik.v21i1.1092

Asani, E., Vahdat-Nejad, H., & Sadri, J. (2021). Restaurant recommender system based on sentiment analysis. *Machine Learning with Applications*, *6*(July), 100114. https://doi.org/10.1016/j.mlwa.2021.100114

Daudert, T. (2021). Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, *230*, 107389. https://doi.org/10.1016/j.knosys.2021.107389

Demircan, M., Seller, A., Abut, F., & Akay, M. F. (2021a). Developing Turkish sentiment analysis models using machine learning and e-commerce data. *International Journal of Cognitive Computing in Engineering*, *2*, 202–207. https://doi.org/10.1016/j.ijcce.2021.11.003

Demircan, M., Seller, A., Abut, F., & Akay, M. F. (2021b). Developing Turkish sentiment analysis models using machine learning and e-commerce data. *International Journal of Cognitive Computing in Engineering*, *2*(July), 202–207. https://doi.org/10.1016/j.ijcce.2021.11.003

Fahmi, S., Faridhi, A., & Hendayana, N. (2021). *Website : https://journal.unilak.ac.id/index.php/Respublica PELAKSANAAN UNDANG-UNDANG NOMOR TENTANG APARATUR SIPIL NEGARA TERHADAP PPPK YANG DILAKUKAN DI SMP Perubahan Undang-Undang Dasar Negara Kesatuan Republik Indonesi Tahun 1945 membawa perubahan dise*. 1–15.

Goldman, Ian. and Pabari, M. (2021). *PERMASALAHAN GURU HONORER TERKAIT KEBIJAKAN PENGHENTIAN REKRUTMEN GURU PNS MENJADI PPPK*, *2*(4).

Mardiana, T., Syahreva, H., & Tuslaela, T. (2019). Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter. *Jurnal Pilar Nusa Mandiri*, *15*(2), 267–274. https://doi.org/10.33480/pilar.v15i2.752

Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-nejad, A. (2021). Journal of Infection and Public Health Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform : A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, *14*(10), 1505–1512. https://doi.org/10.1016/j.jiph.2021.08.010

Muktafin, E. H., & Kusrini, P. (2021). Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach. *Telkomnika (Telecommunication Computing Electronics and Control)*, *19*(1), 146–154. https://doi.org/10.12928/TELKOMNIKA.V19I1.17417

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika (J-Sakti)*, *5*(2), 697–711.

Pratama, B. T., Utami, E., & Sunyoto, A. (2019). An optimization of a lexicon based sentiment analysis method on Indonesian app review. *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019*, 341–346. https://doi.org/10.1109/ICITISEE48480.2019.9003900

Purwokerto, U. A., & Kunci, K. (2021). *PERBANDINGAN METODE SUPPORT VECTOR MACHINE DAN DECISION TREE UNTUK ANALISIS SENTIMEN REVIEW KOMENTAR PADA APLIKASI TRANSPORTASI ONLINE Abstraksi Keywords : Pendahuluan Tinjauan Pustaka*. *2*(2).

Puspita, R., & Widodo, A. (2021). Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS. *Jurnal Informatika Universitas Pamulang*, *5*(4), 646. https://doi.org/10.32493/informatika.v5i4.7622

Rohman, A. N., Luviana Musyarofah, R., Utami, E., & Raharjo, S. (2020). Natural Language Processing on Marketplace Product Review Sentiment Analysis. *2020 2nd International Conference on Cybernetics and Intelligent System, ICORIS 2020*, 6–10. https://doi.org/10.1109/ICORIS50180.2020.9320827

Romadloni, N. T., Santoso, I., & Budilaksono, S. (2019). Perbandingan Metode Naive Bayes , Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl. *Jurnal IKRA-ITH Informatika*, *3*(2), 1–9.

Sa'ad, M. I., Bryan, D., Kusrini, & Supriatin. (2020). Decision Support System for Covid19 Affected Family Cash Aid Recipients Using the Naïve Bayes Algorithm and the Weight Product Method. *2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020*, 120–125. https://doi.org/10.1109/ICOIACT50329.2020.9331964

Sang, A. I., Sutoyo, E., & Darmawan, I. (2021). *Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma Decision Tree Dan Support Vector Machine Data Minning Analysis for Classification of Air Quality Data Dki Jakarta Using Decision Tree Algorthm and Support Vector*. *8*(5), 8954–8963.

Sholeha, E. W., Yunita, S., Hammad, R., Hardita, V. C., & Kaharuddin, K. (2022). Analisis Sentimen Pada Agen Perjalanan Online Menggunakan Naïve Bayes dan K-Nearest Neighbor. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *3*(4), 203–208. https://doi.org/10.35746/jtim.v3i4.178

Sodik, F., & Kharisudin, I. (2021). *Analisis Sentimen dengan SVM , NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter*. *4*, 628–634.

Sudhir, P., & Deshakulkarni, V. (2021). Comparative study of various approaches , applications and classifiers for sentiment analysis. *Global Transitions*

Proceedings, *2*(2), 205–211. https://doi.org/10.1016/j.gltp.2021.08.004

Sugianto, C. A. (2015). Analisis Komparasi Algoritma Klasifikasi Untuk Menangani Data Tidak Seimbang Pada Data Kebakaran Hutan. *Techno.Com*, *14*(4), 336–342. http://publikasi.dinus.ac.id/index.php/technoc/article/view/992

Suryono, S., Utami, E., Luthfi, E. T., Magister, M., Informatika, T., Amikom, U., Yogyakarta, A., Classifier, N. B., Language, N., & Mining, O. (2018). *Analisis Sentiment Pada Twitter Dengan Menggunakan*. 9–15.

Syakuro, A. (2017). Pada Media Sosial Menggunakan Metode Naïve Bayes Classifier (NBC) Dengan Seleksi Fitur Information Gain (IG) Halaman Judul Skripsi Oleh : Abdan Syakuro. *Analisis Sentimen Masyarakat Terhadap E-Commerce Pada Media Sosial Menggunakan Metode Naive Bayes Classifier (NBC) Dengan Seleksi Fitur Information Gain (IG)*, 1–89.

Thakkar, H., Shah, V., Yagnik, H., & Shah, M. (2021). Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. *Clinical eHealth*, *4*, 12–23. https://doi.org/10.1016/j.ceh.2020.11.001

PUSTAKA LAPORAN PENELITIAN

Adipradana, C. (2020). *Pengukuran Kinerja Optimasi Algoritma Bat Pada Algoritma Naive Bayes, Decision Tree Dan K-Nn Untuk Sentimen Analisis Di Lini Masa Twitter*.

Setiawan, A. B. (2021). PEMODELAN PERCAKAPAN BAHASA BANJAR MENGGUNAKAN ARSITEKTUR SEQ2SEQ DENGAN MEKANISME ATENSI. In *Paper Knowledge . Toward a Media History of Documents*.

Imron, A. (2019). *ANALISIS SENTIMEN TERHADAP TEMPAT WISATA DI KABUPATEN REMBANG MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER*.

Indriani, D. (2019). *ANALISIS SENTIMEN PADA TWEET DENGAN TAGAR KPU JANGAN CURANG MENGGUNAKAN NAIVE BAYES*.