KLASIFIKASI KOMENTAR *YOUTUBE* MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

SKRIPSI

Oleh :
IMADA WAHYU NATALIA
NIM. 19650046

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

KLASIFIKASI KOMENTAR *YOUTUBE* MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

SKRIPSI

Oleh:
IMADA WAHYU NATALIA
NIM. 19650046

Diajukan kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2023

HALAMAN PERSETUJUAN

KLASIFIKASI KOMENTAR *YOUTUBE* MENGGUNAKAN METODE *NAIVE BAYES CLASSIFIER*

SKRIPSI

Oleh:
IMADA WAHYU NATALIA
NIM. 19650046

Telah diperiksa dan disetujui untuk Diuji
Tanggal: 24 Maret 2023

Pembimbing I	Pembimbing II
Fajar Rohman Hariri, M.Kom	Okta Qomaruddin Aziz, M.Kom
NIP. 19890515 201801 1 001	NIP.19911019201903 1 013

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT.,IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

KLASIFIKASI KOMENTAR *YOUTUBE* MENGGUNAKAN METODE *NAIVE BAYES CLASSIFIER*

SKRIPSI

Oleh:
IMADA WAHYU NATALIA
NIM. 19650046

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal : 24 Maret 2023

Susunan Dewan Penguji

Ketua Penguji	:	Dr. Ririen Kusumawati, S.Si., M.Kom NIP. 19720309 200501 2 002	( )
Anggota Penguji I	:	Puspa Miladin N. S. A. Basid, M.Kom NIP. 19930828 201903 2 018	( )
Anggota Penguji II	:	Fajar Rohman Hariri, M.Kom NIP. 19890515 201801 1 001	( )
Anggota Penguji III	:	Okta Qomaruddin Aziz, M.Kom NIP. 19911019 201903 1 013	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Cahyo Crysdian, M.MT., IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama	: Imada Wahyu Natalia
NIM	: 19650046
Fakultas/Program Studi	: Sains dan Teknologi/Teknik Informatika
Judul Skripsi	: Klasifikasi Komentar *Youtube* Menggunakan Metode *Naïve Bayes Classifier*

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila di kemudian hari terbukti atau dapat dibuktikan Skripsi ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 20 Maret 2023

Yang membuat pernyataan,

Imada Wahyu Natalia

NIM. 19650046

HALAMAN MOTTO

“Sesuatu yang udah dimulai harus diselesaikan, tepat waktu adalah lebih baik”

“Yakin dulu”

HALAMAN PERSEMBAHAN

Skripsi ini saya persembahkan untuk
Orang Tua, Keluarga,
Teman-teman seperjuangan dan
Diri saya sendiri

Terima kasih

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh

Syukur alhamdulillah, atas berkat rahmatnya, kekuatan dan kasih Tuhan semesta alam Allah SWT yang telah berikan, sehingga penulis diberi kemudahan dalam setiap menyelesaikan skripsi yang berjudul "Klasifikasi Komentar *Youtube* Menggunakan Metode *Naïve Bayes*" sebagai salah satu persyaratan untuk memperoleh gelar sarjana pada program studi Teknik Informatika Jenjang Strata-1 Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang.

Penulis sangat menyadari minimnya ilmu dan pengetahuan. Keberhasilan penulisan skripsi ini tidak lepas dari dorongan dan bimbingan dari berbagai pihak yang telah membantu. Ucapan terima kasih penulis sampaikan kepada seluruh pihak yang sudah membantu baik berupa kritik maupun saran sehingga skripsi dapat terselesaikan, ucapan terima kasih ditujukan kepada yang terhormat:

1. Ibu, Ayah dan Kakak tercinta yang selalu mendukung dan mendoakan dalam setiap langkah hingga akhirnya skripsi ini dapat terselesaikan tepat waktu.
2. Prof. Dr. H. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Sri Hariani, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. Fachrul Kurniawan, M.MT selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang

5. Fajar Rohma Hariri, M.Kom dan Okta Qomarrudin Aziz, M.Kom selaku dosen pembimbing Skripsi, yang telah meluangkan waktunya untuk memberikan ilmu serta arahannya dalam setiap langkah menyelesaikan skripsi ini.
6. Roro Inda Melani, M.T, M.Sc selaku wali dosen yang selalu membimbing serta memotivasi kepada penulis.
7. Anggota Keluarga yang selalu mendoakan dan mendukung penulis
8. M. Iqbal Choirul yang telah membantu dan menemani penulis selama ini
9. Sahabat tercinta yang mendukung penulis
10. Seluruh teman “Alien” Teknik Informatika 2019 untuk semua rangkaian cerita untuk menyemangati penulis
11. Penulis sendiri yang berusaha menyelesaikan skripsi ini tepat waktu
12. Semua pihak yang telah membantu penulis menyelesaikan skripsi ini.

Akhir kata, penulis skripsi ini menyadari masih jauh dari ketidaksempurnaan. Oleh karena itu, penulis sangat mengharapkan saran dan kritik yang membangun untuk perbaikan selanjutnya. Semoga dalam penulisan skripsi ini banyak memberikan manfaat bagi berbagai pihak.

Malang, 20 Maret 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Natalia, Imada Wahyu. 2023. Klasifikasi Komentar Youtube Menggunakan Metode Naïve Bayes Classifier. Skripsi. Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang.
Pembimbing: (I) Fajar Rohman Hariri, M.Kom (II) Okta Qomaruddin Aziz, M.Kom

Kata Kunci: *Klasifikasi Teks, Pengujian cross validasi, Naïve Bayes*

Salah satu dampak dari digitalisasi di era modern ini yakni meluasnya penggunaan media sosial. Berbagai kalangan dari berbagai latar belakang dan usia dapat dengan mudah mengakses media sosial. Hal ini memungkinkan pengguna dapat berinteraksi dengan pengguna lain. Salah satu platform yang paling banyak digunakan untuk berinteraksi dengan pengguna lain terkait dengan video yang trending/populer yakni *youtube*. Melalui fitur komentar di *youtube*, pengguna dapat mengungkapkan pendapat atau tanggapan mereka terhadap suatu video. Tujuan penelitian ini untuk mengklasifikasikan komentar *youtube* menjadi kelas positif, negatif dan netral. Pengklasifikasian tersebut untuk mengetahui gambaran tentang sentiment atau pendapat pengguna terhadap suatu video menggunakan algoritma *Naïve Bayes Classifier* (NBC). NBC dipilih karena algoritma ini sebagai metode klasifikasi teks sederhana yang dihitung dari probabilitas yang didasarkan pada kemunculan kata terhadap komentar *youtube*. Kemudian untuk mengetahui seberapa baik model digunakan untuk klasifikasi dilakukan pengukuran kinerja model untuk mendapatkan nilai akurasi, presisi dan *recall*. Sebanyak 1000 data komentar *youtube* dilakukan uji coba hasilnya menunjukkan bahwa nilai akurasi yang diperoleh sebesar 78% pada nilai *k-20* yakni *fold* ke-1, nilai presisi di kelas positif diperoleh sebesar 69%, kelas negatif sebesar 87%, kelas netral sebesar 67% dan nilai *recall* pada kelas positif sebesar 86%, kelas negatif 84% dan kelas netral sebesar 40%. Berdasarkan hasil pengujian evaluasi model dengan *k-fold cross validation* dapat disimpulkan bahwa metode tersebut mempengaruhi nilai akurasi dan meningkatkan performa dalam sistem.

ABSTRACT

Natalia, Imada Wahyu. 2023. Youtube Comment Classification Using Naïve Bayes Classifier Method. Thesis. Informatics Engineering Study Program, Faculty of Science and Technology, State Islamic University of Maulana Malik Ibrahim Malang.

Supervisors: (I) Fajar Rohman Hariri, M.Kom (II) Okta Qomaruddin Aziz, M.Kom

Keywords: *Text Classificationl, Testing K-fold Cross Validation, Naïve Bayes*

One of many impact from digitalization in modern era is the expanding use of social media. There's various groups with various age and background that can easily access the social media, that makes people can connect with others and interact related with trending video on youtube. Through youtube comments feature, the user can write their opinion against the video. The purpose of this study is to classsified the commentary on youtube within classes, with the class is, positive, negative and netral. This classification will show the overview about sentiment or opinion about an video using Naive Bayes Classifier (NBC). NBC methods has been selected because this methods contains simple text classifying that be counted by probability or incident based on word appearance in the commentary. Next is the calculation to find out how well the model is used for classification to get the value of accuration, precision and recall. About 1000 youtube commentar data the results showed that the accuracy value obtained was 78% at the 1 th k-20 fold value, the precision value in the positive class was obtained by 69%, the negative class was 87%, the neutral class was 67% and the recall value in the positive class is 86%, the negative class is 84% and the neutral class is 40%. Based on the results of model evaluation testing with k-fold cross validation, it can be concluded that this method affects the value of accuracy and improves performance in the system.

خلاصة

ناتاليا ، إيمادا الوحي. 2023. تنفيذ برنامج مأكول لاختبار نظم المعلومات الأكاديمية في مدرسة بحر المغفرة الإسلامية الداخلية. أُطرُوحَة. برنامج دراسة هندسة المعلوماتية ، كلية العلوم والتكنولوجيا ، جامعة الولاية الإسلامية مولانا مالك إبراهيم مالانج.

المشرفون: (I) فجر رحمن الحريري، م.كوم (II) أوكتا قمر الدين عزيز، م. كوم

تصنيف النص ، اختبار التحقق من الصحة عبر K-fold ، Naïve Bayes

أحد آثار الرقمنة في هذا العصر الحديث هو الاستخدام الواسع لوسائل التواصل الاجتماعي .يمكن للعديد من الأشخا من مختلف الخلفيات والأعمار الوصول بسهولة إلى وسائل التواصل الاجتماعي .يتيح ذلك للمستخدمين الاتصال والتفاعل مع مستخدمين أحد أكثر المنصات استخدامًا للتفاعل مع المستخدمين الآخرين YouTube آخرين لمشاركة المعلومات أو تبادل الآراء .يعد موقع يمكن للمستخدمين التعبير عن آرائهم ، YouTube المرتبطين بمقاطع الفيديو الشائعة /الشائعة .من خلال ميزة التعليقات على أحد أكثر المنصات استخدامًا للتفاعل مع المستخدمين الآخرين المرتبطين YouTube أو ردودهم على مقطع فيديو .يعد موقع يمكن للمستخدمين التعبير عن آرائهم أو ردودهم ، YouTube بمقاطع الفيديو الشائعة /الشائعة .من خلال ميزة التعليقات على لأن هذه الطريقة عبارة عن طريقة بسيطة لتصنيف النص Naive Bayes Classifier على مقطع فيديو .تم اختيار طريقة ، ثم لمعرفة مدى استخدام النموذج في التصنيف .YouTube يتم حسابها من الاحتمالات بناءً على ظهور الكلمات في تعليقات Youtube يتم إجراء قياسات النموذج للحصول على قيم الدقة والدقة واسترجاع البيانات .حوالي 1000 من بيانات التعليق على أضعاف ، وتم الحصول على قيمة الدقة في k-20 أظهرت النتائج أن قيمة الدقة التي تم الحصول عليها كانت 78٪ عند قيمة 1 الفئة الإيجابية بنسبة 69٪ ، وكانت الفئة السلبية 87٪ ، وكانت الفئة المحايدة 67٪ وقيمة الاسترجاع في الفئة الموجبة 86٪ والفئة يمكن استنتاج أن هذه ، k-fold السلبية 84٪ والفئة المحايدة 40٪ .بناءً على نتائج اختبار تقييم النموذج مع التحقق من صحة الطريقة تؤثر على قيمة الدقة وتحسن الأداء في النظام.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Interaksi sosial *online* saat ini sebagai salah satu karakteristik masyarakat informasi. Perubahan teknologi informasi telah mengubah masyarakat. Adanya teknologi ini memberikan peran penting dalam berbagai sector kehidupan dan membawa dunia menuju era globalisasi tanpa batas dan jarak. Kemajuan teknologi informasi ini salah satunya melalui penggunaan internet di media sosial seperti muncul situs yang menyediakan berbagai jenis informasi, maraknya industri global dan mulai beralihnya semua media dalam bentuk digital. Meningkatnya jumlah pengguna internet diiringi dengan perkembangan pengguna media sosial setiap tahunnya.

Indonesia menempati posisi sebagai salah satu negara yang memiliki jumlah pengguna internet terbesar di seluruh dunia. Hal ini terbukti bahwa negara Indonesia menempati peringkat keempat di dunia dalam hal waktu penggunaan internet per hari oleh penduduknya, sehingga sejak tahun 2004 tanpa kita sadar media sosial telah berkembang pesat sebagai platform untuk berdiskusi dan mengekspresikan pendapat. (Nuruzzaman, 2018).

Perkembangan pengguna media sosial diimbangi dengan antusiasme masyarakat yang berpengaruh terhadap budaya berpendapat yang kini dilakukan secara *virtual,* yakni melalui fitur komentar pada media sosial. Pada tahun 2020 menunjukkan bahwa platform yang ada di media sosial yang banyak digunakan yakni *youtube*. Data pada bulan Januari 2020 menunjukkan bahwa platform pada

social media yang aktif digunakan yakni media sosial *youtube* dengan presentase sebanyak 88%, seperti yang ditunjukkan dalam artikel (Digital, 2021) *"The Latest Insight Into The State of Digital"* bahwa masyarakat *online* yang mengakses bidang ini sekitar 61,8% dari total pengguna aktif media sosial yang mencapai sekitar 170 juta dari jumlah penduduk Indonesia sekitar 274,9 juta jiwa di tahun 2021. (Ni'matul Rohmah, 2020).

Media sosial menjadi alat komunikasi efektif yang dapat digunakan untuk berkomunikasi, membagikan berbagai informasi dan menyebarkan informasi dimanapun kapanpun. Kemudahan tersebut menjadikan pengguna media sosial sering sekali menyalahgunakan media sosial dengan menyebarkan opini atau komentar yang sifatnya menyakitkan. (Kusumawati et al., n.d.). Komentar tersebut seperti komentar yang tidak sopan, komentar tentang pencemaran nama baik, bentuk radikalisme atau komentar yang cenderung membuat pengguna lain merasa tidak nyaman.

Kata *toxic* sebagai nama yang ada didunia maya yang mulai berkembang ke berbagai media sosial lain yakni *whatsapp*, *twitter* serta media social *youtube* sebagai media yang penyebarannya yang paling cepat (Indah Amelia, n.d.). Pribadi atau individu yang memiliki perilaku *toxic* ialah individu yang memiliki perilaku negatif dan tidak menyenangkan. Hal ini dapat mengganggu hubungan sosial di masyarakat karena memiliki pengaruh yang buruk terhadap individu lain. Pada al-Qur'an surah Al-Imran ayat 159:

فَبِمَا رَحْمَةٍ مِنَ اللَّهِ لِنْتَ لَهُمْ ۖ وَلَوْ كُنْتَ فَظًّا غَلِيظَ الْقَلْبِ لَانْفَضُّوا مِنْ حَوْلِكَ ۖ فَاعْفُ عَنْهُمْ وَاسْتَغْفِرْ لَهُمْ وَشَاوِرْهُمْ فِي الْأَمْرِ ۖ فَإِذَا عَزَمْتَ فَتَوَكَّلْ عَلَى اللَّهِ ۚ إِنَّ اللَّهَ يُحِبُّ الْمُتَوَكِّلِينَ

"*Maka disebabkan rahmat dari Allah-lah kamu Berlaku lemah lembut terhadap mereka. Sekiranya kamu bersikap keras lagi berhati kasar, tentulah mereka menjauhkan diri dari sekelilingmu. Karena itu maafkanlah mereka, mohonkan lah ampun bagi mereka, dan bermusyawaratlah dengan mereka dalam urusan itu. Kemudian apabila kamu telah membulatkan tekad, maka bertawakkallah kepada Allah. Sesungguhnya Allah menyukai orang-orang yang bertawakkal kepada-Nya."*

surah Al -Humazah 1:

وَيْلٌ لِّكُلِّ هُمَزَةٍ لُّمَزَةٍ

"*Kecelakaan lah bagi setiap pengumpat lagi pencela*," (QS. Al-Humazah:1)

Pandangan tersebut memiliki relevansi yang sangat tinggi dengan situasi terkini di media social. Ayat di atas ditinjau dari "Tafsir Al-Azhar" karya Buya Hamka, Dalam hal ini Buya Hamka berpendapat bahwa pengumpat ialah orang yang merasa dirinya benar dan selalu suka membusuk-busuk kan orang lain, serta selalu suka membicarakan keburukan orang lain. Orang seperti ini selalu mencari cacat orang lain tanpa ia menyadari kecacatan pada dirinya sendiri.

Menurut Maria J. Jona (2019), menyebutkan bahwa seseorang yang memiliki pandangan negatif sewajarnya adalah normal, karena jika tidak memiliki pandangan negatif seseorang menjadi tidak waspada. Namun apabila pandangan negatif itu berlebihan maka orang tersebut tidak normal. Hal tersebut dapat dimanfaatkan untuk mencela orang lain. (Maria J, 2019). Hal ini terjadi jika interaksi antar individu tidak terkendali dan terus berlanjut, maka dampaknya akan berdampak pada individu dalam jangka waktu pendek hingga jangka waktu yang lebih lama.

Berdasarkan permasalahan pada penelitian upaya yang dapat dilakukan yakni dengan melakukan klasifikasi komentar *youtube*. Klasifikasi memiliki tujuan yakni memprediksi objek kelasnya dan ciri dari jenis data yang dimilikinya.

Algoritma klasifikasi teks yang seringkali digunakan yakni *Naïve Bayes Classifier* atau biasa disebut dengan NBC. Ini bagian dari pengembangan *machine learning* untuk klasifikasi teks dan sering dilakukan oleh beberapa peneliti dalam melakukan klasifikasi teks.

Terdapat penelitian terdahulu yang telah melakukan penelitian pada topik yang terkait. Penelitian yang dilakukan oleh Mustofa & Mahfudh (2019). Pada penelitiannya menyebutkan tentang klasifikasi *hoax* yang ada pada berita *online* dengan metode NBC, dimana dalam penelitiannya tersebut dilakukan pengujian dengan teknik cross validasi menggunakan nilai *k*-10. Hasil dalam penelitian tersebut yang memperoleh bahwa nilai *fold*-6 memiliki akurasi baik yakni sebesar 85,28%. Hasil akurasi tersebut menunjukkan bahwa sebanyak 307 terklasifikasi relevan dan yang tidak 53 sedangkan total *error rate* sebesar 14,72%. Nilai presisi yang diperoleh yakni sebesar 89% dan recall 85%. (Mustofa & Mahfudh, 2019).

Penelitian oleh Dwi Herlambang & Hadi Wijoyo (2019) yang berjudul tentang klasifikasi teks. Tentang sumber belajar teks pada mata pelajaran dengan metode NBC. Dari penelitian tersebut klasifikasi yang dilakukan menghasilkan 9 kelompok mata pelajaran produktif. Sedangkan pengujian yang berhasil diperoleh menghasilkan nilai akurasi terbaik yakni 81,48% dan akurasi terendah diperoleh dengan nilai 79,63%. (Dwi Herlambang & Hadi Wijoyo, 2019).

Berdasarkan tinjauan pustaka pada bab sebelum, dimana penulis menggunakan *Naïve Bayes.* Tujuannya untuk memprediksi proses klasifikasi komentar *youtube*. *Naïve Bayes Classifier* sebagai *supervised learning* memiliki performa baik untuk mendeteksi teks dan sebagai klasifikasi sederhana yang didasarkan dengan konsep probabilitas. Penelitian ini bertujuan untuk

mengevaluasi tingkat *accuracy*, *precision* dan *recall* yang dihasilkan dalam proses klasifikasi komentar media sosial *youtube*. Penelitian ini menggunakan beberapa pengujian salah satunya dengan teknik cross validasi.

1.2 Pernyataan Masalah

Bagaimana Performa Implementasi metode *Naïve Bayes Classifier* dalam pengklasifikasian komentar *youtube*?

1.3 Tujuan Penelitian

Mengetahui dan Menganalisis pengklasifikasian komentar *youtube* pada metode *Naïve Bayes Classifier* yang mengandung komentar positif, negatif dan netral.

1.4 Batasan Masalah

Penelitian yang dilakukan memiliki batasan masalah agar penelitian menjadi terarah yakni:

1. Data yang penulis ambil dari media sosial *youtube* berbahasa Indonesia yang telah dikumpulkan pada tanggal 16 september 2022 hingga 20 Maret 2023
2. Kelas klasifikasi yang digunakan menjadi komentar positif, negatif dan netral.
3. Keterbatasan penggunaan kata singkatan dalalm dokumen.
4. Mengetahui besar *accuracy*, *precision* dan *recall* metode *Naïve Bayes Classifier* dalam pengklasifikasian komentar *youtube*?

1.5 Manfaat Penelitian

Berdasarkan tujuan, sehingga harapan penulis yang dilakukan memberikan manfaat. Adapun manfaat yang diharapkan diantaranya:

1. Pengembang Penelitian

   Membantu mengetahui *text mining* tentang klasifikasi. Menambah wawasan dalam mengetahui kelebihan dan kekurangan metode *Multinomial Naïve Bayes Classifier* dalam pengklasifikasian teks.

2. Masyarakat Umum

   Memberikan informasi mengenai proses dan implementasi tentang pengolahan data tekstual dari data di media sosial kedalam bentuk pengklasifikasian.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Penelitian terkait digunakan untuk menganalisa serta memperkaya pembahasan, sebagai bahan referensi dan pembeda dengan penelitian yang sedang dilakukan. Dibawah ini beberapa penelitian terdahulu: Penelitian yang dilakukan oleh Hery Mustofa dan Adzhal Arwani Mahfudh (2019). Dalam penelitiannya terhadap berita *hoax* menggunakan metode NBC sebagai model untuk klasifikasi. Dalam jurnalnya, penelitian ini mengangkat masalah tentang berita *hoax* yang sangat banyak tersebar di internet. Tahap klasifikasi berita *hoax* yang dilakukan melalui beberapa tahap seperti proses *preprocessing* seperti *parsing*, *tokenezing*, *stopword removal* dan pembobotan kata dengan (*term-weighting*).

Kemudian setelah proses *preprocessing* selesai dilakukan klasifikasi terhadap teks *hoax* yang ada di berita media social dengan metode NBC. Pengujian yang dilakukan menggunakan fold 10 dengan cross validasi. Hasilnya diketahui bahwa pada *fold* ke-1 memberikan akurasi tertinggi dengan nilai yaitu sebesar 85,28% diketahui bahwa seanyak 307 dokumen relevan sedangkan yang tidak relevance jumlah 53. Total *error rate* sejumlah 14.72% dengan rata-rata yang diperoleh untuk untuk *precision* sebesar 0,896 dan *recall* sebesar 0,853. (Mustofa & Mahfudh, 2019).

Penelitian lain yang dilakukan oleh Aniq Noviciatie U. dkk., (2020) tentang analisis sentimen. Dalam penelitian tersebut dengan metode SVM terhadap berita online. Dalam jurnalnya, penelitian ini melakukan pendeteksian kata atau kalimat yang mengandung kata tidak baik yang terdapat pada berita *online* dengan

menggunakan metode *Neural Language Processing* (NLP) dengan metode SVM. Selanjutnya dilakukan penngukuran untuk mengetahui tingkat keakuratan yang dilakukan dalam penelitian tersebut. metode SVM digunakan untuk menganalisa komentar yakni terkait berita politik. Hasilnya menunjukkan bahwa nilai akurasi terbaik yakni sebesar 54%, *recall* sebesar 49,6%, *precision* 49% dan *f-measure* 49,23%. Penelitian ini memiliki *error rate* yakni sebesar 46,1%. (Kusumawati et al., 2020)

Penelitian oleh Robet Habibi dkk., (2016) dengan judul analisis sentimen. Penelitian ini menggunakan metode *Backpropagation* pada media sosial *twitter* pada tahap klasifikasi. Pada penelitian ini jaringan dalam metode *Backpropagation* dan hasil klasifikasi diuji menggunakan (WEKA) dengan *Multilayer Perceptron Classifier*. Hasil analisis sentimen dengan 30 responden mahasiswa terdapat kecenderungan emosi positif sebesar 33,3%, emosi netral sebesar 53,3% dan emosi negatif sebesar 13,3%. Hasil evaluasi menunjukkan bahwa performa nilai presisi yang diperoleh yakni 58%, *recall* 50% dan *f-measure* sebesar 47%. Hasilnya dijadikan acuan dalam memberikan perlakuan yang tepat kepada siswa selama proses pembelajaran. (Susanto, et. al., 2016)

Penelitian yang dilakukan oleh Berlian Kaida Palma dkk., tentang klasifikasi teks. Klasifikasi ini tentang berita *hoaks covid-19* menggunakan metode KNN. Dalam jurnalnya, berita tersebut menggunakan metode KNN dan cross validasi sebagai teknik untuk pengujian. Proses pengujian yang dilakukan yakni dengan perbandingan dataset 80:20 yakni 80 sebagai data latih dan 20 untuk data uji. Serta, parameter nilai *k* yang digunakan yakni *k*-3, *k*-5, *k*-7, *k*-9. Selanjunya pengujian cross validasi k-5 dan k-10. Hasilnya hasil akurasi *fI-Score* sebesar 48%

pada nilai *k*-5 dan hasil validasi *k*-5 sebesar 42% dan *k*-10 sebesar 45%. (Kaida Palma et al., n.d.)

Pada penelitian yang dilakukan oleh Aditya Quantano S., dkk., (2021) mengenai analisa pada twitter pada PSBB. Dalam jurnalnya, mereka menggunakan tiga metode klasifikasi yaitu NBC, KNN dan DT. Tujuannya untuk dibandingkan akurasi yang terbaik. Twitter mengenai efek PSBB yang diperoleh digunakan untuk sentiment public. Dimana data yang digunakan berjumlah sekitar 2439, kemudian diolah menggunakan *data mining*, seperti *text preprocessing* kemudian akan diklasifikasikan. Hasilnya menunjukkan bahwa akurasi tertinggi dalam penelitian ini diperoleh sebesar 84.78%, nilai presisi 84.78% dan *recall* 100% pada metode *Decision Tree*. (Aditya Quantano Surbakti et al., 2021).

Tabel 2.1 Penelitian Terkait

NO.	PENELITI	JUDUL	METODE	Hasil
1.	Mustofa dan Mahfudh, (2019)	Klasifikasi berita *hoax* dengan menggunakan metode *Naïve Bayes Classifier*	*Naïve Bayes Classifier* dan 10 *fold cross validation*	Penggunaan 10 *fold cross validation* dapat meningkatkan nilai akurasi sebesar 85,28% terklasifikasi 307 dokumen relevan dan 53 tidak relevan, nilai *precision* 0,896 dan *recall* sebesar 0,853
2.	Kusumawati et al., n.d. (2020)	Analisis sentiment *hate speech* portal berita *online* menggunakan *Support Vectore Machine*	*Support Vectore Machine*	Nilai akurasi yang dihasilkan sebesar 53.88%, *recall* 49,69%, *precision* 48,77% dengan *classification error rate* 46,12% dan *f-measure* sebesar 49,23%.
3.	Susanto, n.d. (2020)	Sentimen pada *twitter* mahasiswa menggunakan metode *Backpropagation*	*Backpropagation*	Hasil analisis dengan 30 responden mahasiswa terdapat kecenderungan emosi positif sebesar 33,33%,

				kecenderungan emosi netral sebesar 53,33% dan kecenderungan emosi negatif sebesar 13,33%, nilai *precision* sebesar 0,586, nilai *recall* sebesar 0,503 dan *f-measure* sebesar 0,474
4.	Kaida Palma et al., n.d. (2019)	Klasifikasi teks artikel berita *hoaks covid-19* menggunakan algoritma *K-Nearest Neighbor* (KNN)	*K-Nearest Neighbor* (KNN)	Akurasi terbaik dengan nilai *FI-Score* sebesar 48% dari nilai *k*=5, hasil validasi dari *k-fold cross validation k*=5 sebesar 42% dan *k*=10 sebesar 45%.
5.	Aditya Quantano, dkk., (2021)	Analisa tanggapan terhadap PSBB di Indonesia dengan decision tree pada twitter	*Decission Tree*	Akurasi terbaik dengan yang diperoleh sebesar 84,78% , nilai presisi sebesar 84,78% dan *recall* sebesar 100%

Berdasarkan beberapa penelitian terdahulu yang telah dipaparkan, penulis dalam penelitian ini mengklasifikasikan komentar *youtube* menjadi tiga kelas yakni komentar yang mengandung komentar positif, negatif dan netral. Penelitian ini metode yang digunakan adalah *Naïve Bayes Classifier* dengan menjadikan penelitian terkait sebagai bahan referensi. Proses yang pertama kali dilakukan yakni mengumpulkan *dataset*. Sumber yang digunakan media sosial *youtube* dengan cara *crawling* secara manual. Kemudian *dataset* dilakukan pelabelan dan diolah pada proses *preprocessing*. Setelah *dataset* diolah dalam proses *preprocessing* akan dihitung nilai dengan melakukan pembobotan kata menggunakan metode *term frequency*. Proses terakhir, dilakukan pengklasifikasian dengan metode NBC.

Pengujian yang digunakan yakni menggunakan cross validasi. Teknik *k-fold cross validation* atau cross validasi yakni *k-10, k-15* dan *k-20*. Tujuannya untuk

dibandingkan untuk mengetahui akurasi terbaik dari pemilihan *k* yang ditentukan. Banyak *dataset* yang digunakan yakni 1000 komentar *youtube*. Terdapat beberapa perbedaan dalam penelitian terkait yakni terletak pada studi kasus, metode yang digunakan, proses pengujian nilai *k* yang digunakan pada metode cross validasi, jumlah data dan hasil akurasi.

2.2 *Youtube*

Menurut Shirky, Social media *youtube* ini sebagai situs yang berbasis website yang memiliki manfaat untuk berbagi *video*. Dengan platform ini pengguna media social dapat mengunggah, melakukan tontonan berbagai klip video yang ada. Di Platform ini, terdapat berbagai jenis video, seperti TV, klipfilm, TV serta video buatannya sendiri. Dapat dilihat dari hal ini bahwa popularitas Youtube sangat tinggi, karena dapat dikatakan bahwa platform ini sebagai salah satu database video terbesar di internet. (hendra & Laugu, 2020)

2.3 Klasifikasi

Klasifikasi sebagai aktivitas untuk menilai objek data. Tujuannya untuk mengelompokkan dalam kelas tertentu. Dalam proses klasifikasi ini akan membangun model yang didasarkan pada data training atau biasa disebut dengan latih. Pada sistem yang melakukan proses klasifikasi ini diharapkan dapat melakukan klasifikasi semua data set dengan benar, tetapi tidak dapat dipungkiri bahwa kinerja sistem tidak bisa 100% benar sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya.

Dengan *confusion matrix*, dapat mengetahui kinerja pada sistem klasifikasi. Karena berdasarkan *matrix confussion* diketahui bahwa jumlah data dari masing-masing kelas yang diprediksi secara benar dan data yang diklasifikasikan secara

salah. Pada *matrix confussion* yang dihasilkan yakni tentang nilai akurasi dan nilai eror rate. Dengan mengetahui jumlah data yang diklasifikasi secara benar, dapat mengetahui akurasi hasil klasifikasi, dan dengan mengetahui jumlah data yang diklasifikasikan secara salah dapat mengetahui laju eror yang dilakukan. (Utomo & Mesran, 2020)

2.4 *Term Frequency*

Pembobotan kata yang digunakan berdasarkan banyaknya kata yang muncul dalam dokumen. Dalam penelitian klasifikasi perlu dilakukannya proses pembobotan. Proses pembobotan kata yang banyak digunakan yakni pembobotan kata dengan metode *term frequency* (TF). Pembobotan kata pada dokumen dilakukan dalam penelitian ini yakni menggunakan teknik perhitungan frekuensi kata (TF). Ini diidentifikasi sebagai metode dengan menghitung banyaknya kata yang muncul pada suatu dokumen. (Mustofa & Mahfudh, 2019).

2.5 *Preprocessing Text*

Dalam jurnal oleh Khairunnisa (2021) menyebutkan bahwa tahap *preprocessing* pada klasifikasi teks. Tahap ini digunakan untuk mempersiapkan data teks sebelum digunakan pada proses selanjutnya oleh *machine learning*. Tujuannya untuk memproses teks menjadi lebih efisien, perlu dilakukan tahap transformasi data ke dalam format tertentu. Dimana dalam penelitian ini, proses yang dilakukan melibatkan beberapa langkah yakni *case folding, cleansing, tokenizing, stopword removal,* serta terakhir *stemming* kata. (Khairunnisa et al., 2021).

2.6 *Naïve Bayes Classifier* (NBC)

Dalam jurnal Wibawa (2018) tentang Metode Klasifikasi menyebutkan bahwa metode NBC merupakan algortima klasifikasi yang didasarkan pada teorema Bayes. Ciri khas metode ini yakni asusmsi kuat akan independensi dari setiap variabel. Oleh karena itu, metode ini dinamakan "naïve" atau "naif". Dalam teorema Bayes dikalkulasikan:

$$P(A|B) = \frac{P(A)}{P(B)} P(B|A) \quad \text{.....(2.1)}$$

Kemudian Teorema Bayes dikembangkan atas hukum probabilitas, diperoleh:

$$P(A|B) = \frac{P(A)P(B|A)}{\sum_1^n=1 P(B)} \quad \text{.....(2.2)}$$

Untuk menggambarkan teorema NBC, diketahui bahwa dalam melakukan klasifikasi diperlukan beberapa petunjuk untuk menentukan kelas yang sesuai bagi sampel yang sedang dianalisis. Karena itu, teorema ini dikalkulasikan:

$$P(C|F_{1\ldots} F_n) = \frac{P(C)P(F1\ldots.Fn|C)}{P(F1\ldots Fn)} \quad \text{.....(2.3)}$$

Dimana variable C merepresentasikan kelas, sementara variable F1…Fn merepresentasikan karakteristik-karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Rumus tersebut menjelaskan bahwa peluang masuknya sampel dengan karakteristik tertentu dalam kelas C (*posterior*) peluang munculnya kelas C.

Dimana kelas C sebagai (*posterior/prior*) merupakan peluang munculnya kelas C (sebelum masuknya sampel tersebut), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (*likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (*evidence*). Sehingga:

$$Posterior = \frac{Prior\ x\ likelihood}{evidence} \quad \text{.....(2.4)}$$

Nilai *evidence* selalu tetap untuk setiap kelas pada satu sampel. Nilai dari *posterior* tersebut nantinya akan dibandingkan dengan nilai-nilai *posterior* kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. (Wibawa et al., 2018).

2.7 Model *Multinomial*

Metode Multinomial sebagai model dari Naïve Bayes akan memanfaatkan total term pada dokumen. Metode ini sebagai algoritma Naïve Bayes yang mana dalam model NBC untuk *multinomial* bahwa dokumen tersebut terdiri dari beberapa kejadian kata yang diasumsikan panjang dokumen tidak bergantung pada kelasnya. Pada model ini asumsi bayes yang sama bahwa kemungkinan tiap kejadian kata dalam sebuah dokumen adalah bebas tidak terpengaruh dengan konteks kata dan posisi kata dalam dokumen. (Destuardi & Sumpeno, 2009)

2.8 *Confusion Matrix*

Confusion matrix dilakukan pengukuran nilai akurasi, nilai presisi, nilai recall dan nilai f-measure. Nilai – nilai ini merujuk pada nilai TP, TN, FP dan FN. *Confusion matrix* membantu mengukur seberapa baik sebuah *classifier* yang digunakan mengenali *tuple* dari kelas yang berbeda. Dimana parameter yang dimaksudkan terdapat pada hasil *accuracy.* (Farah Zhafira et al., 2021). Dalam *confusion matrix* akurasi merupakan tingkat kedekatan dengan nilai prediksi dan nilai aktual.

a. Nilai *accuracy* merupakan nilai untuk mengetahui evaluasi banyaknya kelas prediksi yang sesuai dengan kelas actual

b. Presisi merupakan nilai untuk mengetahui tingkat ketepatan. Nilai ini akan memberikan informasi tentangapa yang dimau pengguna dengan jawaban yang diberikan oleh sistem
c. Nilai recal merupakan nilai untuk mengetahui tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. (Sriyano & Setiawan, n.d.)

2.9 Metode *k-fold Cross Validation*

Cross validasi untuk validasi model. Tujuannya digunakan untuk mengevaluasi seberapa baik model statistic dapat menggeneralisasi pada kumpulan data independent. Metode ini digunakan untuk memprediksi model dan memperkirakan akurasi model prediksi ketika dijalankan pada data yang tidak kenal sebelumnya. Dalam cross validasi, data training dibagi menjadi beberapa fold atau bagian, kemudian model dilatih pada beberapa fold dan diuji pada fold lainnya. Setiap fold nantinya dimanfaatkan sebagai data training dan data testing, sehingga model dapat diuji pada seluruh data training. Hal ini dilakukan untuk memperoleh estimasi yang lebih akurat tentang kemampuan model dalam menggeneralisasi data independent. (Rhomadhona & Permadi, 2019)

BAB III

METODE PENELITIAN

3.1 Pengumpulan Data

Pada sub bab ke 3, peneliti akan menjelaskan *dataset* yang digunakan objek penelitian. Sumber *dataset* diperoleh dari komentar media sosial *youtube.* Data yang digunakan merupakan data pada tahun 2020-2023. Pengambilan data dilakukan dengan *crawling* data secara manual yang diambil dari beberapa *channel youtube* yang berbeda-beda kemudian dipilih secara random untuk digunakan sebagai *dataset* penelitian.

Penelitian ini *dataset* yang digunakan sebagai penelitian terdiri dari dua jenis yakni data *training testing set*. Dimana *dataset training* akan digunakan untuk membentuk model klasifikasi yang digunakan dengan *Naïve Bayes Classifier* (NBC), sedangkan *dataset testing* untuk menguji performa sistem dari model tersebut. Selain itu dalam penelitian ini juga dibutuhkan beberapa data pendukung yakni *stopword* yang digunakan pada saat melakukan proses *preprocessing* data.

3.2 Pelabelan Data

Proses yang perlu dilakukan setelah pengumpulan *dataset* yakni proses pelabelan pada komentar. Proses pelabelan data atau *labelling* merupakan proses yang dilakukan untuk memberikan identitas pada setiap data komentar *youtube*. Pelabelan tersebut menjadi tiga kelas yakni positif, netral dan negatifs. Dalam pelabelan pada komentar divalidasi secara manual oleh ahli bahasa Dosen Bahasa Indonesia atas nama Ibu Siwi Tri Purnani,M.Pd. sebagai validator 1 dan Ibu Arianti Ningsih,S.Pd. sebagai validator 2 untuk menentukan komentar *youtube* berdasarkan kelas sebagai *dataset*. Data yang telah di crawling tersebut akan diubah ke dalam

format *.excel* agar dapat digunakan dan disimpan ke dalam *database*. Jumlah data yang digunakan berjumlah 1000 yang sudah terlabeli. Proses pelabelan data dilakukan dengan menggunakan beberapa indikator klasifikasi yang telah ditentukan oleh ahli bahasa. Tabel 3.1 menunjukkan Indikator Klasifikasi. Dalam penelitian ini terdapat indikator klasifikasi yang digunakan peneliti untuk mengetahui tingkatan kelas klasifikasi (Effendy, 1972):

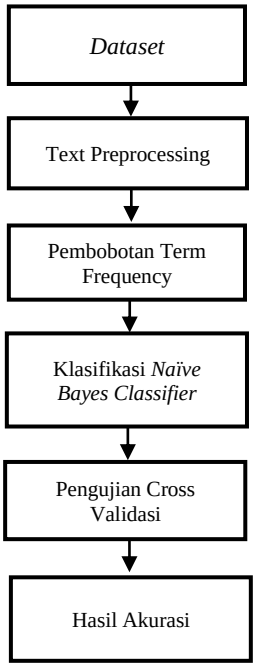
Tabel 3.1 Indikator klasifikasi (Effendy, 1972)

Positif	Jika opini yang ditampilkan secara eksplisit dan implisit mendukung objek opini (individu memberikan pernyataan setuju).
Negatif	Jika opini yang ditampilkan secara eksplisit dan implisit menolak atau mencela objek opini (individu memberikan pernyataan tidak setuju).
Netral	Kalimat netral apabila opini yang ditampilkan tidak memihak atau jika individu memberikan pernyataan ragu-ragu.

Berdasarkan indikator klasifikasi yang telah ditentukan oleh ahli bahasa, maka komentar *youtube* dapat dengan mudah di identifikasi berdasarkan kelasnya. Indikator klasifikasi digunakan untuk mengenali atau membedakan jenis komentar menjadi komentar yang mengandung kalimat positif, komentar yang mengandung kalimat negatif dan komentar yang mengandung kalimat netral.

3.3 Rancangan Sistem

Sistem klasifikasi komentar *youtube* dilakukan untuk mengidentifikasi jenis komentar tertentu dalam kelas positif, negatif dan netral. Serta, mengetahui nilai akurasi dari komentar *youtube* yang telah diklasifikasi menggunakan metode NBC. Sistem dalam penelitian ini memiliki rancangan yakni bagaimana alur sistem berjalan. Pada penelitian ini terdapat beberapa tahap yang diperlukan. Pada Gambar 3.1 ditunjukkan rancangan sistem klasifikasi yang menjelaskan tentang bagaimana alur sistem berjalan.



Gambar 3.1 Rancangan Sistem

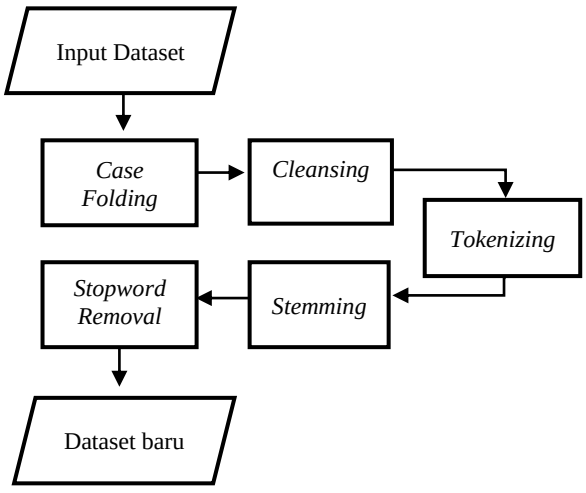
Gambar 3.1 sebagai rancangan sistem klasifikasi komentar *youtube.* Pada penelitian ini rancangan sistem yang dibuat, dimulai dari pengambilan data yang bersumber dari media sosial *youtube* tahun 2020-2023. Pengambilan data dilakukan dengan *crawling* secara manual. Setelah data terkumpul kemudian diproses agar dengan mudah diolah oleh *machine learning* dengan metode *Naïve Bayes Classifier* yakni melalui proses *preprocesing text.* Dalam proses *preprocessing text* terdapat beberapa tahapan yang perlu dilakukan.

Hasil *preprocessing text* tersebut kemudian akan dilanjutkan untuk menghitung bobot kata menggunakan metode TF atau menghitung jumlah frekuensi kata. Proses ini akan memberikan nilai bobot pada setiap kata yang didasarkan pada jumlah frekuensi kata yang muncul dalam dokumen yang dianalisis. Terakhir, dilakukan klasifikasi algoritma *Naïve Bayes Classifier* (NBC) model *multinomial*. Model *multinomial* sebagai metode klasifikasi yang menggunakan metode probabilitas dan statistic.

Selanjutnya dilakukan pengujian dengan metode cross validasi atau biasa disebut dengan teknik cross validasi dengan beberapa nilai k yang digunakan. Metode yang digunakan akan menentukan *dataset* berdasarkan nilai *k* pada *k-fold cross validation* yang telah ditentukan. Metode *cross validation* akan membagi semua dokumen dalam sekumpulan *k dataset*, dimana setiap dataset tersebut memiliki kesempatan untuk menjadi *testing set* dan *dataset* lainnya tersebut menjadi *training set.* Proses terakhir akan dilakukan evaluasi performa sistem dengan perhitungan dengan *confusion matrix.*

3.4 *Preprocessing Text*

Preprocessing merupakan tahap yang sangat krusial dalam klasifikasi data teks. Hal ini dilakukan untuk mengurangi jumlah atribut yang tidak digunakan pada proses klasifikasi. Dalam tahap ini data yang digunakan merupakan data mentah, kemudian diproses untuk memudahkan proses klasifikasi. Beberapa tahap yang diperlukan dalam proses *preprocessing* ditunjukkan Pada Gambar 3.2 berikut.



Gambar 3.2 Diagram *Preprocessing*

Gambar 3.2 sebagai diagram alir proses *preprocessing text.* Proses ini akan menghilangkan beberapa atribut dalam komentar yang tidak penting. Pada *preprocessing text* terdapat beberapa langkah-langkah yang perlu dilakukan yakni proses *case folding*. Proses *case folding* akan menyeragamkan karakter pada teks, mengubah huruf kapital pada semua komentar *youtube* menjadi huruf kecil. Tujuannya untuk menghilangkan redudansi data yang hanya berbeda pada huruf saja.

Tahap kedua proses *cleansing*. Proses ini akan membersihkan atribut atau kata yang tidak digunakan pada proses klasifikasi. Dalam komentar *youtube* terdapat beberapa atribut yang tidak berpengaruh dalam klasifikasi. Contoh atribut komentar yang tidak digunakan dalam klasifikasi seperti adanya *emoticon* (😊), atribut *hastag* (‘@’), atau karakter simbol seperti berikut ini (~!@#$%^&*()_-?.,{}[]/).

Tahap ketiga dalam *preprocessing* yakni proses *tokenizing*. Proses *tokenizing* akan mengubah teks menjadi ukuran token. Proses *tokenizing* akan dilakukan pemotongan kata. Pemotongan yang dilakukan ini didasarkan pada tiap kata yang menyusun menjadi potongan tunggal. Proses ini dilakukan dengan cara memisahkan kata berdarkan spasi yang terletak diantara dua kata.

Setelah tahap pemotongan tiap kata selesai dilakukan, maka proses selanjutnya yakni proses *stemming*. Dalam proses ini peneliti menggunakan teori Nazief dan Mirna Adriani. Algoritma ini dirancang oleh Bobby Nazief dan Mirna Adriani yakni: (Wahyudi et al., n.d.)

1. Kata yang akan di *stem* dalam kamus. Namun apabila kata ditemukan, maka kata tersebut dianggap sebagai kata dasar dan algoritma dihentikan.

2. Menghilangkan *inflection suffixes* dari kata, seperti *particles (“-lah”, “-kah”, “-tah”,* atau *“-pun”)* dan *inflectional possesive pronouns* (*“-ku”, “-mu”*, atau *“-nya”*). Kemudian, algoritma akan memeriksa apakah kata yang dihasilkan terdapat dalam kamus. Jika ditemukan, maka algoritma akan dihentikan, jika tidak lanjut ke tahap selanjunya.
3. Menghapus *derivation suffixes* (*“-i”, “-an”*, atau *“-kan”*). Apabila kata yang dihasilkan sudah ditemukan dalam kamus, maka algoritma akan berhenti. Namun, jika belum ditemukan, maka akan dilakukan tahap 3a:
    a. Pada tahap ini, jika akhiran *“-an”* sudah dihapus dan huruf terakhir dari kata tersebut adalah *“-k”*, maka *“-k”* juga akan dihapus. Jika kata tersebut sudah ditemukan dalam kamus, maka algoritma berhenti. Namun, jika kata tersebut belum ditemukan dalam kamus, maka selanjutnya adalah tahap 3b.
    b. Pada tahap ini, akhiran yang telah dihapus (*“-i”, “-an”* atau *“-kan”*) akan dikembalikan, dan algoritma akan menalnjutkan ke tahap 4.
4. Menghapus *Derivation Prefix* (*“be-”, “di-”, “ke-”, “me-”, “pe-”, “se-”* dan *“te-”*). Jika kata yang dihasilkan sudah terdapat dalam kamus kata dasar, maka proses pengolahan kata berhenti. Namun, jika belum ada dalam kamus, maka lakukan pengkodean ulang. Tahap ini akan dihentikan jika memenuhi kondisi berikut:
    a. Periksa table kombinasi awalan akhiran yang tidak diizinkan. Jika ditemukan, maka proses berhenti, jika tidak lanjut ke langkah 4b.

b. Untuk nilai I dari 1 hingga 3, tentukan tipe awalan kemudian hapus awalan. Jika kata dasar belum ditemukan, lakukan langkah 5, jika sudah maka algoritma berhenti.

5. Langkah telah dilakukan namun kata dasar tidak ditemukan dalam kamus, maka algoritma tersebut akan mengembalikan kata awal sebelum proses stemming dilakukan. Setelah itu, proses akan dianggap selesai.

6.

Proses terakhir *preprocessing* yakni *stopword removal*. Proses terakhir tentang cara membandingkan kata dalam dokumen teks dengan daftar kata *stoplist* yang telah disusun sebelumnya. *Stoplist* tersebut tentang kumpulan kata yang tidak relevan tetapi sering muncul dalam dokumen. Setiap kata dalam dokumen akan diperiksa pada stoplist yang ada. Biasanya, stoplist berisi kata-kata seperti kata ganti orang atau kata hubung. Tabel 3.2 menunjukkan langkah-langkah dalam *preprocessing text*.

Tabel 3.2 Proses *Preprocessing*

Preprocessing	***Input***	***Output***
Case Folding	Jangan jadi Egois. Semua orang berhak bahagia.	jangan jadi egois. semua orang berhak Bahagia
Cleansing	@Weny g tau malu!!!!	weny g tau malu
Tokenizing	Makanya jadi perempuan itu jangan murahan.	‘makanya’,‘jadi’,‘perempuan’,‘itu’,‘jangan’, ‘murahan’
Stemming	Balasan orang yg berbohong dan berbohong	‘balasan’, ‘orang’,’bohong’
Stopword Removal	Sukurin balasan orang yg berbohong dan berbohong	sukurin balasan orang berbohong

3.5 Pembobotan kata

Penelitian ini menggunakan skema pembobotan kata dengan metode *term frequency* (TF) untuk klasifikasi teks. Tujuan penggunaan skema TF untuk

menentukan bobot sebuah kata di dalam banyaknya dokumen. Data hasil preprocessing tersebut dalam bentuk numerik. Perhitungan statistik numerik untuk mencerminkan seberapa penting dan seberapa relevannya sebuah kata di dalam sebuah dokumen. Cara untuk mengubah data numerik yakni dengan menghitung bobot frekuensi kata. Metode frekuensi kata (TF) untuk mencari nilai dari kata tersebut yang ada pada dokumen.

3.5.1 Menghitung Nilai *Term Frequency* (TF)

Proses menghitung frekuensi kata dihitung nilai frekuensi kemunculan kata yang muncul dalam dokumen. Pada Table 3.3 ditunjukkan contoh data *training* yang digunakan dan Table 3.4 sebagai proses untuk mengetahui kata yang muncul pada dokumen dengan menghitung frekuensi kata yang terdapat pada komentar *youtube*.

Tabel 3.3 Contoh Dokumen Latih

Dokumen	Komentar	Kelas
D1	“makin tinggi ilmu bukan makin tinggi tingkah”	Negatif
D2	“ingat padi makin isi makin tunduk.”	Positif
D3	“dia rasa sedang atas, tau ilmu segala”	Netral

Tabel 3.4 ditunjukkan diatas merupakan contoh data penelitian yang digunakan. Ketiga data tersebut masing-masing sudah terlabeli dengan kelas positif, negatif dan netral. Contoh Data tersebut selanjutnya akan dihitung jumlah frekuensi kata (TF) dan nilai *document frequency* (DF) sebagai jumlah dari banyaknya dokumen sutau *term* muncul. Proses untuk menghitung banyaknya dokumen suatu *term* yang muncul (DF) ditunjukkan pada Tabel 3.6 seperti berikut.

Tabel 3.4 Hasil Perhitungan Frekuensi Kata (TF)

Term(t)	**kata muncul pada dokumen (kelas)**		
	D1	**D2**	**D3**
Tinggi	2	0	0
Ilmu	1	0	1
Tingkah	1	0	0
Ingat	0	1	0
Padi	0	1	0
Isi	0	1	0
Tunduk	0	1	0
Dia	0	0	1
Rasa	0	0	1
Sedang	0	0	1
Atas	0	0	1
Segala	0	0	1
Total	4	4	6

3.5.2 Menghitung Nilai *Document Frequency* (DF)

Proses yang dilakukan setelah nilai *term frequency* dilakukan selanjutnya menghitung nilai *document frequency* (DF). Tabel 3.6 berikut menunjukkan hasil dari perhitungan *document frequency* (DF) dari perhitungan proses *term frequency* yang sebelumnya sudah dilakukan.

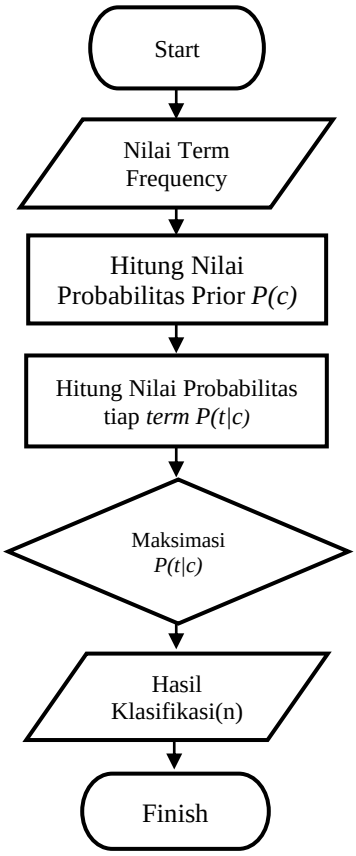
Tabel 3.6 Hasil Perhitungan DF

Term	**Nilai Df**
Tinggi	2
Ilmu	2
Tingkah	1
Ingat	1
Padi	1
Isi	1
Tunduk	1
Dia	1
Rasa	1

Sedang	1
Atas	1
Segala	1

3.6 Klasifikasi *Naïve Bayes Classifier* (NBC)

Pada metode *Naïve Bayes Classifier* (NBC) sebagai metode klasifikasi statik yang menghitung probabilitas kejadian dan digunakan untuk klasifikasi. Pada tahap awal, metode ini melalui tahap pelatihan atau training untuk membuat model klasifikasi. Setelah model dibuat, tahap testing dilakukan dengan mengacu pada dataset training yang ada pada klasifikasi komentar media social *Youtube* dengan menggunakan algoritma *Naïve Bayes Classifier* (NBC) dapat dijelaskan melalui diagram alur pada Gambar 3.3 berikut.



Gambar 3.3 Flowchart NBC

Pada algoritma *Naïve Bayes Classifier* (NBC) sebagai metode untuk teknik klasifikasi. Proses klasifikasi yang dapat memprediksi probabilitas keanggotaan *class*. Metode *Naïve bayes* mengasumsikan bahwa nilai atribut pada sebuah *class* adalah independent terhadap nilai pada atribut lain (Darujati & Bimo Gumelar, 2012).

Nilai atribut dicari nilai *probabilitas prior*, probabilitas atribut tiap kelas dan hasilnya nilai probabilitas yang paling tinggi sebagai hasil dari klasifikasi NBC. Dalam klasifikasi NBC memerlukan *probabilitas term* pada data yang sudah diberikan label sesuai dengan kelasnya. Pada penelitian ini kelas yang digunakan untuk data komentar *youtube* yakni positif, negatif dan netral. Proses untuk menghitung probabilitas *term* dilakukan pada tahap *preprocessing text*, dimana proses *term frekuency* mengubah data teks menjadi bentuk numeric, yang dapat digunakan oleh model *machine learning* yaitu *Naïve Bayes Classifier*.

Kemudian setelah menghitung kemunculan kata selesai ditentukan, selanjutnya menghitung *probabilitas prior* dapat dilakukan dengan menggunakan persamaan 3.1 berikut.

$$P\,(\mathrm{c}) = \frac{Nc}{N} \quad \text{..... (3.1)}$$

Di mana *probabilitas prior P*(c) dihitung dari *N*c yang merupakan jumlah dokumen pada kelas c dibagi dengan *N* sebagai jumlah keseluruhan dokumen. Perhitungan probabilitas dengan persamaan 3.1 ditunjukkan pada Tabel 3.7 berikut.

Tabel 3.7 Probabilitas Prior

Kelas	P (kelas)
Negatif	1/3
Positif	1/3
Netral	1/3

Setelah probabilitas *prior* ditemukan, langkah selanjutnya menghitung nilai probabilitas *term* pada kelas menggunakan Persamaan pada 3.2.

$$P\ (t|c) = \frac{freq(t,c)+1}{freq(c)+V} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.2)$$

Di mana *P*(t|c) adalah probabilitas *term* t pada kelas c, freq(t,c) jumlah frequensi *term* t pada kelas c, sedangkan freq (c) jumlah *term* yang ada pada kelas c dan V adalah jumlah keseluruhan *term* pada data latih. Hasil perhitungan probabilitas kelas dilakulasikan persamaan 3.2 ditunjukkan pada Table 3.8.

Tabel 3.8 Perhitungan probabilitas setiap kata

Term	D1	D2	D3
Tinggi	0,4285	0,1428	0,1111
Ilmu	0,2857	0,1428	0,2222
Tingkah	0,2857	0,1428	0,1111
Ingat	0,1428	0,2857	0,1111
Padi	0,1428	0,2857	0,1111
Isi	0,1428	0,2857	0,1111
Tunduk	0,1428	0,2857	0,1111
Dia	0,1428	0,1428	0,2222
Rasa	0,1428	0,1428	0,2222
Sedang	0,1428	0,1428	0,2222
Atas	0,1428	0,1428	0,2222
Segala	0,1428	0,1428	0,2222

Output dari perhitungan probabilitas setiap kata akan dijadikan dasar untuk tahap klasifikasi selanjutnya. Sebagai contoh, Tabel 3.9 menunjukkan dokumen uji yang akan digunakan.

Tabel 3.9 Dokumen Uji

Komentar	Kelas
tinggi ilmu dia sedang berada diatas	?

Selanjutnya menghitung probabilitas kelas tiap *term* menggunakan Persamaan 3.2 berikut. Perhitungan pada persamaan 3.3 terdapat pada Tabel 3.10.

Tabel 3.10 Probabilitas kata pada Dokumen Uji

Term	D1	D2	D3
Tinggi	0,4285	0,1428	0,1111
Ilmu	0,2857	0,1428	0,2222
Dia	0,1428	0,1428	0,2222
Sedang	0,1428	0,1428	0,2222
Atas	0,1428	0,1428	0,2222

Di mana probabilitas kelas c untuk *term* t, diketahui bahwa $P(t|c)$ sebagai probabilitas *term* t pada kelas c, $P(c)$ sebagai probabilitas prior dan $P(t)$ sebagai probabilitas jumlah dokumen yang terdapat pada *term* t terhadap keseluruhan dokumen. Hasil perhitungan $P(t)$ digunakan untuk mencari nilai kelas tertinggi yang akan digunakan untuk proses klasifikasi.

Setelah keseluruhan nilai probabilitas tiap *term* ditemukan, perhitungan selanjutnya adalah menghitung nilai probabilitas kelas tiap dokumen dengan Persamaan 3.3 berikut.

$$P(c|u) = \prod^{n}_{i=1} P(t_i|c).\ P(c) \text{…...……} (3.3)$$

Dari nilai probabilitas kelas tiap dokumen yang telah diperoleh, selanjutnya menentukan klasifikasi dokumen menggunakan nilai tertinggi dari nilai probabilitas

setiap kelas. Untuk itu perhitungan nilai probabilitas klasikasi dikalkulasikan berdasarkan persamaan 3.4. seperti berikut

$$P(c|u)=P(t_i|c)*P(c1|u)*P(c2|u)*P(c3|u)*P(c4|u)*P(c5|u) \quad \text{.....(3.4)}$$

Sehingga perhitungannya dikalkulasikan sebagai berikut.

$$P(c|u) = P(1/3) * P(\text{Tinggi}|\text{Neg})*P(\text{Ilmu}|\text{Neg})*P(\text{Dia}|\text{Neg})*P(\text{Sedang}|\text{Neg})* P(\text{atas}|\text{Neg})$$
$$= \frac{1}{3} * 0{,}4285 * 0{,}2857 * 0{,}1428 * 0{,}1428 * 0{,}1428$$
$$= 0{,}00011883$$

Sebagai contoh berikut ini table 3.11 hasil dari perhitungan untuk kelas beracun pada probabilitas klasifikasi berikut:

Tabel 3.11 Hasil Probabilitas Klasifikasi

Term	**Probabilitas**
Negatif	0,00011883
Positif	0,00001979
Netral	0,00009028

Kelas terbaik dalam klasifikasi *Naïve Bayes Classifier* ditentukan dengan mencari *maximum a posterior* (MAP) kelas C_{MAP} sehingga diperoleh persamaan sebagai berikut.

$$CMAP = \arg\max (P(c|u)) \quad \text{.....(3.5)}$$

3.7 Skenario Uji Coba

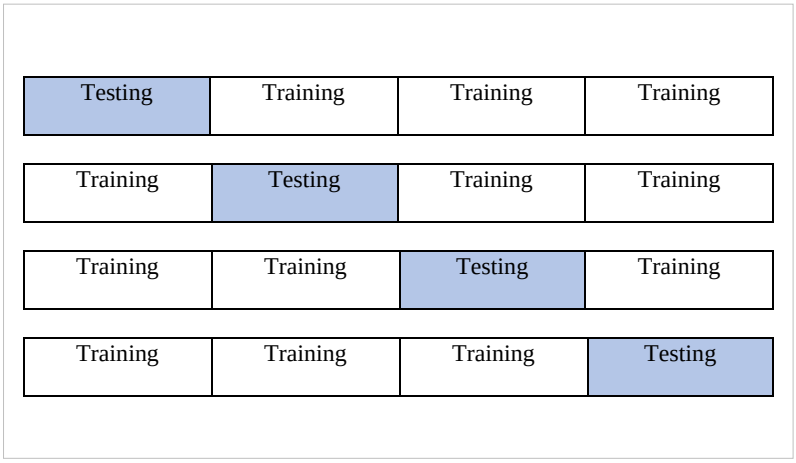
Untuk menguji kinerja sistem, kemudian dilakukan uji coba menggunakan data latih model klasifikasi menggunakan Naïve Bayes Classifier. Selanjutnya, data testing digunakan untk menguji akurasi dan performa sistem dari model yang telah dibuat. Pengujian dilakukan beberapa kali untuk memperoleh akurasi terbaik pada proses klasifikasi komentar. Pengujian pertama dilakukan dengan pembagian rasio data.

Pengujian ini dilakukan untuk menemukan rasio pembagian data *training* dan *testing* yang paling optimal untuk menghasilkan model yang akurat dari berbagai pembagian data yang telah ditentukan yakni sebesar 9:1 maka *dataset* akan dibagi menjadi dua bagian menjadi 90% untuk data *training* dan 10% untuk data *testing*, 8:2, 7:3, 6:4 dan 5:5. Dengan adanya pembagian rasio pembagian data akan dilihat rasio mana yang memiliki akurasi tertinggi.

Kemudian setelah hasil akurasi tertinggi dari pengujian pertama diperoleh selanjutnya akan dilakukan pengujian terhadap *dataset* menggunakan metode *k-fold cross validation.* Pemilihan nilai *k* yang digunakan pada pengujian *k-fold cross validation* menggunakan sebanyak tiga kali yakni *k-10, k-15* dan *k-20*. Tujuan pemilihan nilai *k* yang digunakan pada penelitian ini untuk membandingkan diantara ketiga nilai *k* untuk mengetahui nilai akurasi yang terbaik.

Pengujian dengan menggunakan metode *k-fold cross validation*, teknik ini digunakan untuk mengetahui seberapa akurat sebuah model. Dengan metode ini, sebanyak 1000 dokumen digunakan sebagai dataset untuk pelatihan dan pengujian klasifikasi komentar *youtube* menggunakan *Naïve Bayes Classifier*. Pelatihan dan pengujian dilakukan sebanyak *k* yang digunakan. Diketahui bahwa saat dataset ke-

i menjadi *testing* set, maka dataset yang lainnya digunakan menjadi *training set*. Alur kerja dari metode *k-fold cross validation* diilustrasikan pada Gambar 3.4.



Gambar 3.4 Visualisasi *4-fold cross validation*

Gambar 3.4 sebagai visualisasi 4-*fold cross validation,* data *training* dan *testing* pada teknik validasi *k-fold cross validation*. Setelah pengujian kedua diperoleh, proses selanjutnya yakni proses evaluasi akurasi klasifikasi. Proses ini dilakukan dengan memperhatikan *confusion matrix* (Said et al., n.d.). Performa sistem akan dievaluasi menggunakan akurasi, presisi dan recall yang dihitung menggunakan *confusion matrix*. Tabel *confusion matrix* dapat dilihat pada Tabel 3.12.

Tabel 3.12 Pengujian *Confusion Matrix*

Aktual	**Prediksi**		
	Positive	Netral	Negative
Positive	*True Positive* (TP)	*False Positive* (FP)	*False Positive* (FP)
Netral	*False Negative* (FN)	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Negative* (FN)	*False Positive* (FP)	*True Positive* (TP)

Tabel 3.12 sebagai pengujian menggunakan *confusion matrix*. Evaluasi sistem dilakukan dengan model *Confusion Matrix* untuk membandingkan kelas data

sebenarnya dengan kelas hasil prediksi. Perhitungan untuk akurasi dikalkulasi dengan persamaan 4.1. 4.2. 4.3. terhadap sistem. Dimana nilai True Positive (TP) terjadi ketika prediksi dan kenyataan sama sama positif, nilai False Positive (FP) terjadi ketika prediksi positif tetapi kenyataannya negatif, nilai True Negative (TN) terjadi ketika prediksi dan kenyataan sama sama negatif, dan terakhir nilai False Negative (FN) ketika prediksi negative tetapi kenyataanya positif.

Nilai akurasi menunjukkan seberapa akurat sistem dapat mengklasifikasikan hasil akurasi data secara benar. Perhitungan untuk nilai *accuracy*, *precision* dan *recall* dikalkulasi menggunakan persamaan berikut.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ x } 100\% \text{ ..……..…….. (4.1)}$$

$$Precision = \frac{TP}{(TP+FP)} \text{ x } 100\% \text{ ..……..…….(4.2)}$$

$$Recall = \frac{TP}{(TP+FN)} \text{ x } 100\% \text{ ..…..……..…….(4.3)}$$

3.8 Implementasi Sistem

Pada penelitian klasifikasi komentar *youtube* sistem diimplementasikan menggunakan bahasa *python jupyter notebook.* Sistem ini dapat mengklasifikasikan komentar secara otomatis. Implementasi *Naïve Bayes Classifier* dimulai dengan tahap pengumpulan data yang sudah memiliki label atau kelas masing-masing yang disimpan dalam *excel* yang akan diinput untuk proses *preprocessing*. Kemudian menghitung bobot kata hingga mendapatkan klasifikasi secara otomatis.

3.8.1 Implementasi *Preprocessing Text*

Dalam proses *preprocessing* terdapat beberapa tahap yang diperlukan. Berikut pada Gambar 3.5 ditunjukkan beberapa library pada *python* yang digunakan pada proses *preprocessing* sebagai berikut.

```
!pip install Sastrawi
!pip install emoji

import nltk
nltk.download('punkt')
import seaborn as sns
from nltk.corpus import stopwords
import pandas as pd
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
import re
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
nltk.download('stopwords')
from nltk.corpus import stopwords
```

Gambar 3.5 *Library Python Preprocessing*

Kemudian dalam proses klasifikasi beberapa tahapan dalam penelitian ini diuraikan sebagai berikut. Pertama dimulai dari proses *preprocessing*, dalam *preprocessing* terdapat beberapa tahap yang diperlukan yakni salah satunya kamus pendukung stopword. Pada Gambar 3.6. dapat dilihat implementasi dari proses *preprocessing*.

```
#Cleaning
def remove(text):
    text = re.sub(r'[^\w\s]', '', text)
    text = re.sub('[0-9]+', '', text)
    text = re.sub(r'\$\w*', '', text)
    text = re.sub(r'((www\.[^\s]+)|(https?://[^\s]+))','URL',text)
    text = re.sub(r'#', '', text)
    text = re.sub(r',','',text)

    return text
df['clean'] = df['Komentar'].apply(lambda x: remove(x))

# Casefolding
df['casefolding']= df['clean'].apply(lambda x: x.lower())

# Tokenisasi
df["tokenized"] = df["casefolding"].apply(word_tokenize)

# Stemming
factory = StemmerFactory()
stemmer = factory.create_stemmer()
def stemming(text):
    stem_word = [stemmer.stem(word) for word in text]# stemming word
    return stem_word
df["stemmed"] = df["tokenized"].apply(lambda x: [stemmer.stem(word) for word in x])

# Stopword Removal
stopwords_indonesia = stopwords.words('indonesian')

def remove_stopwords(text):
    output= " ".join([i for i in text if i not in stopwords_indonesia])
    return output

df['stopword']= df['stemmed'].apply(lambda x:remove_stopwords(x))
df.head(10)
```

Gambar 3.6 Proses *Preprocessing*

Pada *preprocessing* terdapat beberapa tahap diantaranya *cleaning*. Proses *cleaning* dilakukan untuk membersihkan data yang tidak berpengaruh seperti adanya karakter dalam teks, kemudian *case folding* untuk menyamakan semua karakter teks, hasil dari *case folding* akan diproses untuk mengubah ke proses *tokenizing* yang diubah denga memberikan spasi. Kemudian terdapat proses

stemming untuk menghilangkan kata imbuhan dan terakhir proses proses *preprocessing* yakni *stopword removal.*

3.8.2 Implementasi Pembobotan Kata (TF)

Data komentar yang sudah melewati proses *preprocessing* selanjutnya akan dihitung frekuensi kemunculan kata dengan metode (*term frequency*) dan probabilitas. Proses menghitung frekuensi kemunculan kata ini menggunakan fungsi pemanggilan fungsi "*get_word_frequency*". Pada Gambar 3.7 menunjukkan implementasi untuk pembobotan kata dengan *term frequency.*

```
def get_word_frequency(word_list):
    frequency_dict = {}
    for word in word_list:
        if word in frequency_dict:
            frequency_dict[word] += 1
        else:
            frequency_dict[word] = 1
    return frequency_dict

#TF POSITIVE
positive_word_frequency = get_word_frequency(positive_comment_words)
print("Kata-kata yang muncul pada komentar positif:")
for word in positive_word_frequency:
    print(word, ":", positive_word_frequency[word])
negative_word_frequency = get_word_frequency(negative_comment_words)
print("Kata-kata yang muncul pada komentar negatif:")
for word in negative_word_frequency:
    print(word, ":", negative_word_frequency[word])
netral_word_frequency = get_word_frequency(netral_comment_words)
print("Kata-kata yang muncul pada komentar netral:")
for word in netral_word_frequency:
    print(word, ":", netral_word_frequency[word])
```

Gambar 3.7 Pembobotan Kata TF

Pada Gambar 3.7 diatas, pembobotan kata dilakukan dengan metode *term frequency*. Dimana fungsi "*get_word_frequency*" akan memanggil kata yang

muncul pada *dataset* berdasarkan kelas masing-masing. Kemudian nilai TF pada masing-masing kelas akan dicetak meggunakan fungsi “print” yang digunakan. Hasil yang diperoleh pada perhitungan frekuensi kata pada komentar positif, negatif dan netral ditunjukkan pada Gambar 3.8., 3.9. dan 3.10.

```
Kata-kata yang muncul pada komentar positif:
rakyat : 5
attitude : 1
kawal : 1
tuntas : 3
usut : 3
tuntaskita : 1
warga : 1
negara : 5
manja : 1
jabat : 18
kerabat : 1
```

Gambar 3.8 Frekuensi kata muncul pada komentar positif

Gambar 3.8 ditampilkan 13 kemunculan kata pada komentar positif. Kemudian fungsi “*get_word_frequency*” akan memanggil kata yang muncul pada kelas negatif, sehingga Gambar 3.9 ditampilkan hasil kemunculannya.

```
Kata-kata yang muncul pada komentar negatif:
hah : 2
hoh : 2
rakyat : 5
wajib : 3
bayar : 4
pajak : 18
oknum : 2
jabat : 21
bencana : 1
copot : 1
tindak : 2
```

Gambar 3.9 Frekuensi kata muncul pada komentar negatif

```
Kata-kata yang muncul pada komentar netral:
denny : 2
tampar : 1
definisi : 2
kalo : 4
blom : 1
puas : 1
main : 2
mending : 1
nikah : 14
dlu : 1
podcast : 10
```

Gambar 3.10 Frekuensi kata muncul pada komentar netral

3.8.3 Implementasi Kata Unik

Setelah mengetahui probabilitas frekuensi kata yang dihitung dengan membagi jumlah kemunculan kata dengan jumlah total kata dalam kelas komentar. Selanjutnya menampilkan kata unik yang ada dalam *dataset*, proses ini sebagai implementasi kata unik yang muncul dalam masing-masing kelas akan ditampilkan menggunakan fungsi dalam python "*get_unique_word_frequency*". Implementasi untuk memanggil kata unik ditunjukkan pada Gambar 3.11.

```
def get_word_frequency(word_list):
    frequency_dict = {}
    for word in word_list:
        if word in frequency_dict:
            frequency_dict[word] += 1
        else:
            frequency_dict[word] = 1
    return frequency_dict

#TF POSITIVE
positive_word_frequency = get_word_frequency(positive_comment_words)
print("Kata-kata yang muncul pada komentar positif:")
for word in positive_word_frequency:
    print(word, ":", positive_word_frequency[word])
negative_word_frequency = get_word_frequency(negative_comment_words)
print("Kata-kata yang muncul pada komentar negatif:")
for word in negative_word_frequency:
    print(word, ":", negative_word_frequency[word])
netral_word_frequency = get_word_frequency(netral_comment_words)
print("Kata-kata yang muncul pada komentar netral:")
for word in netral_word_frequency:
    print(word, ":", netral_word_frequency[word])
```

Gambar 3.11 Probabilitas Kata Unik

3.8.4 Implementasi Probabilitas Prior

Proses selanjutnya adalah implementasi klasifikasi metode Naïve Bayes Clssifier (NBC). Untuk dapat mengklasifikasikan suatu data teks dengan metode NBC perlu menghitung probabilitas prior. Nilai dalam probabilitas prior dihitung dengan menghitung seluruh kata dalam kelas yang ingin dicari dibagi dengan total seluruh kata dalam dokumen. Implementasi nilai probabilitas prior pada masing-masing kelas ditunjukkan pada Gambar 3.12.

```
N = len(positive_comment_words) + len(negative_comment_words) +
len(netral_comment_words)

P_positive = len(positive_comment_words)
P_negative = len(negative_comment_words)
P_netral = len(netral_comment_words)

P_prior = {
    'positif': P_positive/N,
    'negative': P_negative/N,
    'netral': P_netral/N
}

print("Probabilitas Prior:")
for kelas, prob in P_prior.items():
    print(kelas, ":", prob)
```

Gambar 3.12 Implementasi Probabilitas Prior

Total keseluruhan kata yang ada dalam dokumen diperoleh sejumlah 7447, maka perhitungan yang diperoleh untuk probabilitas prior ditunjukkan pada Tabel 3.13.

Tabel 3.13 Perhitungan Probabilitas Prior

No	Probabilitas Prior		
	Positive	Negative	Netral
1	$\frac{3308}{7447}$ = 0.4442,	$\frac{2295}{7447}$ = 0.3081,	$\frac{1844}{7447}$ = 0.2476,

3.8.5 Implementasi Nilai *Likelihood*

Setelah perhitungan probabilitas prior diketahui, selanjutnya menghitung nilai *likelihood*. Proses perhitungan *likelihood* dilakukan dengan menghitung jumlah kemunculan kata di dalam kategori klasifikasi. *Naïve Bayes* sebagai salah satu algoritma klasifikasi yang digunakan untuk membedakan *instance* dataset

berdasarkan atribut yang telah ditentukan. (Yanti Liliana et al., 2021). Berikut ini ditunjukkan implementasi dari *likelihood* pada Gambar 3.13.

```
positive_likelihood = {}
for word in new_comment_tokens:
    if word in vocabulary:
        count = positive_word_count.get(word, 0)
        positive_likelihood[word] = (count + 1) / (len(positive_comment_words) + len(voca
bulary))
negative_likelihood = {}
for word in new_comment_tokens:
    if word in vocabulary:
        count = negative_word_count.get(word, 0)
        negative_likelihood[word] = (count + 1) / (len(negative_comment_words) + len(voca
bulary))
netral_likelihood = {}
for word in new_comment_tokens:
    if word in vocabulary:
        count = netral_word_count.get(word, 0)
        netral_likelihood[word] = (count + 1) / (len(netral_comment_words) + len(vocabula
ry))
```

Gambar 3.13 Implementasi *likelihood*

Perhitungan nilai *likelihood* dilakukan dengan menghitung jumlah kemunculan kata di dalam kategori klasifikasi, ditambahkan dengan satu karena untuk menghindari pembagian dengan nol kemudian dibagi dengan jumlah kata di dalam kategori klasifikasi tersebut, ditambah dengan jumlah kata dalam kamus *"vocabulary"*. Fungsi ini sebagai kamus kata yang berisi semua kata yang ditemukan di dalam kumpulan komentar positif, negatif dan netral.

3.8.6 Implementasi Probabilitas Posterior

Proses selanjutnya menghitung nilai probabilitas posterior dokumen. Probabilitas prior dilakukan untuk memprediksi peluang suatu data. Prediksi *Naïve Bayes* dilakukan dengan memilih probabilitas posterior maksimum dari ketiga

kelas. (Yanti Liliana et al., 2021). Pada Gambar 3.14 menunjukkan implementasi dari probabilitas posterior.

```
positive posterior = P positive
for word in new_comment_tokens:
    if word in positive_likelihood:
        positive_posterior *= positive_likelihood[word]
positive_posterior = positive_posterior * P_positive

negative_posterior = P_negative
for word in new_comment_tokens:
    if word in negative_likelihood:
        negative_posterior *= negative_likelihood[word]
negative_posterior = negative_posterior * P_negative

netral_posterior = P_netral
for word in new_comment_tokens:
    if word in netral_likelihood:
        netral_posterior *= netral_likelihood[word]
netral_posterior = netral_posterior * P_netral
```

Gambar 3.14 Implementasi Probabilitas Posterior

Setelah nilai probabilitas kelas tiap dokumen diketahui, selanjutnya yakni menentukan klasifikasi dokumen dengan *Naïve Bayes Classifier*.

```
#posterior probability tertinggi (MAP)
posterior_probabilities = {'positive': positive_posterior, 'negative': negative_posteri
or, 'netral': netral_posterior}
predicted_class = max(posterior_probabilities, key=posterior_probabilities.get)
```

Gambar 3.15 Implementasi klasifikasi NBC

Gambar 3.5 digunakan untuk memprediksi klasifikasi dari sebuah komentar berdasarkan kemungkinan kata-kata di dalam komentar tersebut yang ada pada kelas positive, negative dan netral. Dengan menggunakan nilai *likelihood* yang telah dihitung sebelumnya, code ini digunakan untuk memilih klasifikasi dengan

kemungkinan tertinggi sebagai prediksi untuk kelas komentar tersebut pada *Naïve Bayes Classifier*.

3.8.7 Implementasi *Naïve Bayes Classifier* data *testing*

Dalam memprediksi data testing, beberapa tahap yang sebelumnya dijelaskan akan digunakan untuk proses ini. Berikut pada Gambar 3.16 menunjukkan contoh data testing yang berhasil diklasifikasikan secara otomatis dengan sistem.

```
new_comment = "tangis nya pc tangis sandiwara"
"
```

Gambar 3.16 Data testing

Gambar 3.16 sebagai data testing yang belum memiliki kelas. Kemudian akan diprediksi dengan klasifikasi dengan metode *Naïve Bayes Classifier* model *multinomial*. Pengklasifikasian ini didasarkan nilai probabilitas kata perhitungan nilai *term frequency* pada proses *training* yang telah dilakukan sebelumnya. Pada Gambar 3.17 menunjukkan hasil klasifikasi yang diperoleh dari data *testing.*

```
Probabilitas positif: 1.1409657903790601e-10
Probabilitas negative: 4.722039668449006e-08
Probabilitas netral: 3.857751303872347e-11
--------------------------------------------
Komentar:  tangis nya pc tangis sandiwara
Kelas prediksi:  negative
```

Gambar 3.17 Hasil Klasifikasi NBC data testing

Hasil yang diperoleh dari klasifikasi testing yang telah diinputkan tersebut ditunjukkan Pada Gambar 3.17 sebagai hasil dari klasifikasi *Naïve Bayes Classifier*. Data tersebut termasuk dalam kelas negatif. Hal ini terjadi karena nilai probabilitas kata yang diperoleh pada kelas negatif sebagai kelas dengan nilai yang paling tinggi.

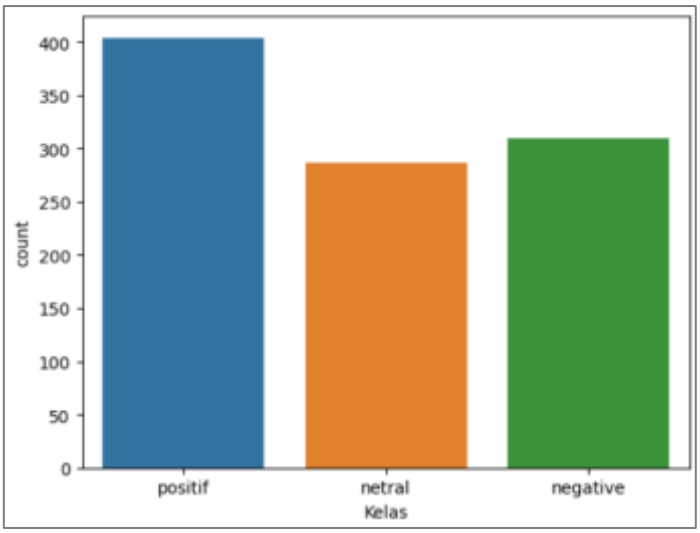
BAB IV

UJI COBA DAN PEMBAHASAN

Bab 4 membahas tentang pengujian sistem dan evaluasi hasilnya. Pada bab keempat terdiri dari hasil pengujian, pembahasan tentang hasil pengujian, serta integrasi sistem dengan islam.

4.1 Data Penelitian

Penelitian ini menggunakan dataset dengan total data komentar youtube sebanyak 1000 yang diperoleh secara manual pada beberapa *channel youtube*. Data hasil dari *crawling* tersebut dilakukan pelabelan dengan bantuan ahli bahasa Indonesia. Data terkumpul sebanyak 404 komentar dengan kelas positif, sebanyak 309 komentar negatif dan sebanyak 287 komentar netral. Pada Gambar 4.1 berikut merupakan visualisasi persebaran data yang akan digunakan untuk klasifikasi.



Gambar 4.1 Grafik Hasil Klasifikasi

Gambar 4.1 sebagai grafik batang hasil klasifikasi. Perolehan jumlah klasifikasi tersebut merupakan hasil dari pelabelan manual yang telah divalidasi oleh ahli bahasa. Dalam grafik batang tersebut menunjukkan bahwa klasifikasi pada

kelas negatif memiliki jumlah yang paling banyak. Hal ini ditunjukkan pada grafik batang warna hijau, sedangkan jumlah pada kelas positif yang ditunjukkan pada warna biru berada diurutan kedua dan jumlah yang paling sedikit yakni kelas netral yang ditunjukkan dalam grafik batang berwarna orange. Total data yang digunakan sebanyak 1000 data komentar *youtube* digunakan sebagai data penelitian, berikut Tabel 4.1 ditunjukkan sampel data pada penelitian yang digunakan.

Tabel 4.1 Sampel Data Penelitian

No.	**Komentar *youtube***	**Kelas**
1.	Bu Ala keliatan orangnya penyabar & pekerja keras	Positif
2.	Selalu suka sama mama Ala. Kaya raya tapi humble dan bersahaja	Positif
3.	memang ibu ini luar biasa aslinya	Positif
4.	Wanita hebat. Semoga ada anak sy yg sukses seperti Bu Ala	Positif
5.	MasyaAllah semoga sayaa bisa menjadi seperti ibu Ala	Netral
6.	Masya Allah luar biasa sangat menginspirasi 👍	Positif
7.	Bu Ala masya Allah ibu sultan panutan😇	Positif
8.	Ramah kepada setiap org & ga sombong	Positif
9.	Seneng lihat bu Alah orang nya rendah hati	Positif
10.	Bu Ala, emang beda ya yg kaya "beneran" dengan yang "sok" kaya	Positif
11.	Wanita karir, tangguh dan mandiri 👍👍👍	Positif
12.	Semoga anak-anakku pekerja cerdas seperti bu Ala.	Positif
13.	Beliau adalah single parent dan sudah terbiasa hidup dg kerja keras	Positif
14.	Bu ala keren sy suka wanita hebat. Cerdas	Positif
15.	Masya allah tabarakallah. kereeen uar biasaaaah..👍👍👍😍😍	Positif
16.	Masyaallah. barokah bu Ala	Netral

17.	Bismillah. pingin bisa ketemu dengan Bu Ala	Netral
18.	Bener2 Super Wonder woman 💪💪👍👍♡♡	Positif
19.	beliau humble banget.	Positif
20.	Keren ya single mom mandiri dan suksess	Positif
21	Keliatan ibunya ryan orang nya baik dan sopan	Positif
22	Keliatan banget ibunya Ryan wanita pintar dan berwibawa.	Positif
23	semoga lekas dapat jodoh yang terbaik bagi keduanya.	Netral
24	Semoga Yessy dapat pasangan yg Sholeh..In SyaaAllah. Aamiin👏	Netral
25	Saya n suami bersyukur nikah secara sangat sederhana, yg pnting bahagia ♡	Positif

4.2 Menampilkan Hasil *Training*

Pada tahap ini penulis akan menampilkan hasil training sebagai langkah awal. Pada penelitian yang dilakukan menggunakan proses *training* sebagai awal pengujian. Proses *training* yang dilakukan bertujuan untuk pembentukan kelas dan sebagai acuan bagaimana suatu dokumen testing akan diklasifikasikan. (Wijaya & Santoso, 2016).

Hasil dari proses *training* yang berhasil dibangun mendapatkan nilai akurasi sebesar 90.3%. Kemudian dari proses *training* yang telah dilakukan langkah selanjutnya yakni melakukan pengujian pada data *testing.* Tujuannya untuk mengukur kinerja model klasifikasi yang telah dibangun. Pada pengujian ini, data *testing* digunakan untuk memeriksa apakah model dapat secara akurat memprediksi kelas dari data baru yang belum pernah dilihat sebelumnya. Selama proses *training*, jumlah distribusi kata yang digunakan pada data *training* ditampilkan dengan

proses vektorisasi yakni fungsi “CountVectorizer”. Fungsi tersebut digunakan untuk menghitung jumlah kata yang muncul pada setiap dokumen dalam *corpus*.

Hasil dari vektorisasi merupakan representasi numerik dari dokumen dalam bentuk matriks, di mana setiap matriks baris mewakili dokumen dan setiap kolom mewakili kata yang muncul dalam seluruh dokumen *corpus*. Pada data *training* yang digunakan, setiap kata akan memiliki frekuensi kemunculan yang berbeda dalam seluruh dokumen dalam *corpus*. Dimana frekuensi kemunculan kata tersebut diurutkan secara *descending* yang didasarkan pada jumlah kemunculan kata. Hasil perhitungan ditampilkan dalam bentuk table. Berikut Tabel 4.2 ditampilkan sampel kata dengan frekuensi teratas pada data *training*.

Tabel 4.2 distribusi jumlah kata pada grafik distribusi jumlah kata

No	Kata	Frekuensi
1	Moga	117
2	Orang	112
3	Anak	108
4	Yg	94
5	Nya	90
6	Ya	85
7	Banget	82
8	Ryan	79
9	Allah	63
10	Keluarga	63
11	Hokum	50
12	Gak	48
13	Jabat	43
14	Kaya	37
15	Ga	35
16	Jodoh	34
17	Mas	33
18	Ajar	33
19	Sehat	33
20	Aja	33

Tabel distribusi kata pada data *training* dibuat dengan memperhitungkan presentase kemunculan kata. Hal ini dilakukan dengan menggunakan iterasi pada list frekuensi kata yang dihasilkan dari proses vektorisasi pada “train_word_freq”.

Dengan begitu, kata yang paling penting dan sering muncul pada data training dapat diketahui.

4.3 Hasil Uji Coba

Tahap skenario uji coba yang telah dijelaskan pada bab ketiga. Pengujian yang awal akan dilakukan pembagian dataset penelitian. Pembagian yang dilakukan menggunakan beberapa rasio pembagian dataset *training* dan *testing*. Pada Tabel 4.3 menunjukkan perbandingan hasil akurasi dari uji coba pertama yang dilakukan pada data *testing* dengan rasio pembagian *dataset* yang digunakan.

Tabel 4.3 Hasil Uji Coba Pembagian dataset *testing*

Rasio Pembagian *dataset*	Hasil Akurasi
5:5	64%
6:4	63%
7:3	65%
8:2	69%
9:1	74%

Tabel 4.3 Hasil uji coba pembagian *datatse*t diperoleh akurasi tertinggi sebesar 74% pada rasio pembagian data 9:1. Dalam melakukan pembagian dataset *training* dan *testing* dalam pengujian dilakukan secara acak (random) yakni menggunakan nilai random state. Tujuan dari penggunaan nilai random state untuk menghindari bias pada pembagian dataset dan memastikan bahwa model yang dihasilkan dapat menggeneralisasi dengan baik pada data yang belum pernah dilihat sebelumnya. Dengan menggunakan pengujian random state dilakukan untuk memastikan pembagian dataset yang acak dapat diulang dengan hasil yang sama.

Pengujian selanjutnya dengan mengevaluasi performa sistem menggunakan model *confusion matrix*. Hasil pengujian dengan *confusion matrix* ditunjukkan pada Tabel 4.4. Untuk klasifikasi tiga kelas yakni positif, negatif dan netral maka model

confusion matrix yang digunakan terdiri dari matriks 3x3. Hal ini berarti kolom mewakili kelas prediksi dan baris mewakili kelas sebenarnya.

Tabel 4.4 Hasil Pengujian *confusion matrix*

Aktual	**Prediksi**		
	Positif	Negatif	Netral
Positif	26	0	5
Negatif	3	10	12
Netral	1	5	38

Tabel 4.4 hasil *confusion matrix*, dapat dilihat bahwa Nilai *True Positive* (TP) diperoleh sebesar 26, 10 dan 38, kemudian nilai *False Positive* (FP) diperoleh nilai sebesar 0, 5 dan 5 dan nilai *False Negative* (FN) yang diperoleh sebesar 3, 1 dan 12. Pada hasil uji coba akan dilakukan evaluasi model dengan *confusion matrix*, evaluasi ini dilakukan pada rasio pembagian data 9:1. Hasil dari evaluasi *confusion matrix* dapat dilihat pada Gambar 4.4. dan hasil perhitungan nilai akurasi diperoleh dengan perhitungan berikut.

$$Accuracy = \frac{\text{TP}}{(Keseluruhan\ Data)} \text{ x } 100\%$$

$$= \frac{(26+10+38)}{(26+10+38+5+12+5+1+3)} \text{ x } 100\% = 74\%$$

Perhitunngan nilai *precission* dan *recall* dihitung dengan rumus pada persamaan 4.2 dan 4.3, dengan persamaan tersebut hasilnya ditunjukkan pada Tabel 4.5 menggunakan *confusion matrix.* Hasilnya dihitung dari masing-masing kelas yakni kelas postif, negatif dan netral karena termasuk sebagai *multiclass*. Untuk kasus klasifikasi tiga kelas model *confusion matrix* terdiri dari matriks 3x3 berikut. Pada Tabel 4.5 sebagai hasil dari perhitungan dengan *confusion matrix* dari skenario uji coba yang sebelumnya dijelaskan, hasilnya seperti berikut.

Tabel 4.5 Hasil Perhitungan *Confussion Matrix*

Aktual	**Prediksi**	
	Precision	***Recall***
Positive	$\frac{38}{38+12+5}$ x 100% = 69%	$\frac{38}{38+1+5}$ x 100% = 86%
Negative	$\frac{26}{26+3+1}$ x 100% = 87%	$\frac{26}{26+0+5}$ x 100% = 84%
Netral	$\frac{10}{10+0+5}$ x 100% = 67%	$\frac{10}{10+3+12}$ x 100% = 40%
Rata-Rata	74.33%	70%

Tabel 4.5 hasil perhitungan dengan *confusion matrix*, dapat dilihat bahwa perhitungan dengan c*onfusion matrix* untuk *multiclass* dilakukan dengan menghitung semua kelas. Hasil dari pengujian *confusion matrix* menunjukkan kinerja model dalam mengklasifikasikan data ke dalam tiga kelas yang berbeda yakni nilai *precision* untuk kelas positif, negatif, netral dan nilai *recall* pada kelas positif, negatif dan netral.

Nilai *precision* merupakan rasio antara jumlah benar positif (*True Positive*) dengan jumlah positif yang diprediksi (*True Positive + False Positive*) untuk mengukur seberapa akurat model dalam memprediksi kelas positif. Dalam penelitian ini, rata-rata nilai *precision* yang diperoleh yakni sebesar 74.33% sedangkan rata-rata nilai *recall* sebesar 70%, artinya nilai *precision* memiliki nilai lebih tinggi bahwa model yang digunakan berhasil memprediksi 74.33% kelas positif dengan benar. Hal ini menunjukkan bahwa model lebih baik dalam membatasi jumlah false positif (*False Positive*) atau kesalahan dalam memprediksi kelas positif.

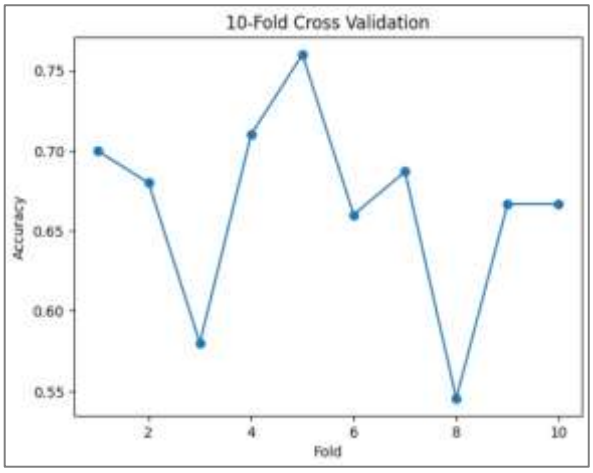
Hasil dari masing-masing nilai yang telah diperoleh yakni nilai presisi pada kelas positif diperoleh sebesar 69%. Dari peroleh nilai tersebut menunjukkan bahwa dari semua data yang diklasifikasi sebagai kelas positif oleh model hanya 69%

diantaranya yang benar-benar termasuk dalam kelas positif. Kemudian untuk kelas negatif yang diperoleh sebesar 87%, hal ini berarti sebesar 87% dari semua data diklasifikasikan sebagai kelas negatif oleh model. Sedangkan kelas netral diperoleh nilai sebesar 67% berarti sekitar 67% dari semua data diklasifikasikan sebagai kelas netral oleh model.

Kemudian untuk nilai *recall* kelas positif diperoleh sebesar 86% sehingga nilai *recall* kelas positif sebesar 86% yang menunjukkan bahwa model dapat mengenali sebagian besar data yang seharusnya termasuk dalam kelas positif. Sedangkan untuk kelas negatif diperoleh nilai *recall* sebesar 84% yang berarti nilai *recall* kelas negatif sebesar 84% yang menunjukkan bahwa model dapat mengenali sebagian besar data yang seharusnya termasuk dalam kelas negatif dan terakhir, nilai *recall* pada kelas netral diperoleh nilai sebesar 40%. Hal ini menunjukkan bahwa model kurang efektif dala mengenali data yang seharusnya termasuk ke dalam kelas netral. Hal ini berarti secara keseluruhan hasil dari *confusion matrix* membantu memahami kinerja model dalam mengklasifikasikan data ke dalam tiga yang berbeda.

Berdasarkan hasil dari pengujian yang sebelumnya telah dilakukan, diperoleh hasil terbaik yakni pada rasio perbandingan data 9:1 terhadap data *training* dan data *testing*. Kemudian pengujian dilanjutkan dengan metode *k-fold cross validation*. Pengujian tersebut divalidasi dengan teknik *k-fold cross validation*, dengan beberapa pemilihan nilai *k* yang digunakan yakni *k*-20, *k*-15 dan *k*-10. Tujuannya untuk membandingkan nilai akurasi terbaik dari nilai *k* yang ditentukan. Pengujian ini untuk mengetahui nilai maksimal, minimal dan rata-rata dari masing-masing pemilihan *k* yang digunakan terhadap *dataset*. Pengujian

dengan *k-20* maka akan terjadi percobaan sebanyak 20 iterasi. Pada *k*-15 maka akan terjadi percobaan sebanyak 15 iterasi dan nilai *k*-10 akan dilakukan percobaan sebanyak 10 iterasi terhadap *dataset*. Berikut ditampilkan grafik dari hasil pengujian terhadap metode *k-fold cross validation*. Pada Gambar 4.2, Gambar 4.3 dan Gambar 4.4 menunjukkan hasil perbandingan dari ketiga nilai *k* yang digunakan. Hasil pertama validasi dengan *k-fold cross validation* ditunjukkan pada Gambar 4.2 dengan nilai *k-10*.



Gambar 4.2 Grafik Model 10-*fold cross validation*

Gambar 4.2 grafik hasil pengujian dengan model *k-fold cross validation* yakni pemilihan nilai *k*-10. Dalam hal ini, dipilih nilai *k*-10 sehingga data akan dibagi menjadi 10 *fold*. Dimana, setiap *fold* digunakan sebagai data validasi satu kali, sedangkan 9 *fold* lainnya digunakan sebagai data *training.* Hasil grafik dari pemilihan nilai *k-10* menunjukkan terdapat satu titik dimana akurasi tertinggi diperoleh pada rentang nilai *fold* ke 4 sampai 6. Sedangkan akurasi rendah diperoleh pada rentang ke 8. Dari Gambar gafik hasil pengujian menggunakan *k-fold cross validation* juga ditampilkan kedalam bentuk tabel. Pada Tabel 4.6 ditunjukkan hasil

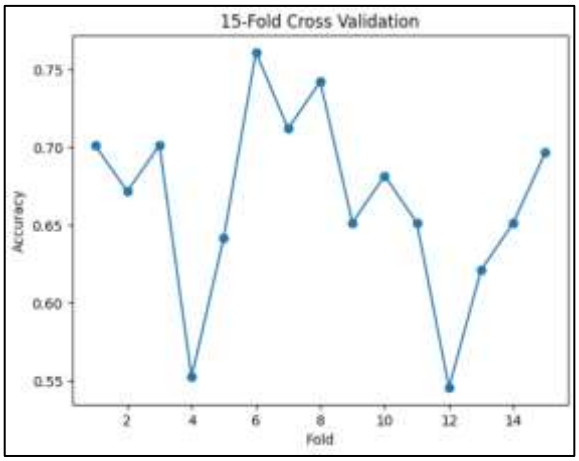
dari penguiian menggunakan metode *k-fold cross validation* dengan nilai *k* sebesar 10.

Tabel 4.6 Hasil Pengujian Menggunakan 10-*fold cross validation*

Fold ke-	Hasil Akurasi
1	72%
2	68%
3	63%
4	67%
5	72%
6	64%
7	69%
8	57%
9	66%
10	67%
rata-rata	66,5%

Tabel 4.6 menunjukkan hasil pengujian menggunakan *k-fold cross validation* dengan nilai *k* yang digunakan bernilai 10 sehingga pengujian dilakukan sebanyak 10 iterasi. Pada Tabel yang ditunjukkan merepresentasikan hasil *k-fold cross validation* dalam tabel yang artinya bahwa dalam tabel itu nilai akurasi pada setiap iterasi berbeda-beda. Nilai rata-rata akurasi yang diperoleh pada pengujian ini yakni sebesar 66,5%.

Dari perolehan nilai rata-rata akurasi tersebut digunakan untuk indikator performa model dalam mengklasifikasikan data. Dengan membagi data menjadi beberapa *fold* dan melakukan pengujian pada setiap *fold* artinya model akan mengevaluasi lebih baik lagi. Selanjutnya pengujian kedua dilakukan dengan pemilihan nilai *k-15*. Tujuannya untuk mengetahui nilai akurasi yang lebih baik dari pengujian yang sebelumnya telah dilakukan yakni dengan *k-10*. Hasil pengujian kedua dengan nilai *k-15* ditunjukkan pada Gambar 4.3 berikut.



Gambar 4.3 Grafik Model 15-*fold cross validation*

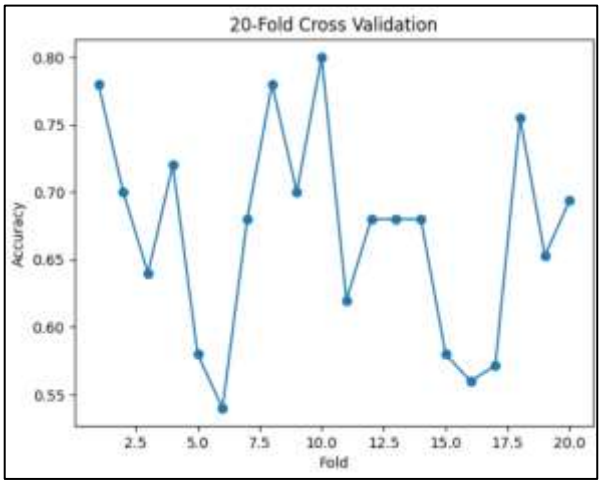
Gambar 4.3 dilakukan pemilihan nilai *k*-15. Dalam hal ini, dipilih nilai *k*-15 sehingga data akan membagi menjadi 15 *fold*. Dimana, bahwa setiap *fold* digunakan sebagai data validasi satu kali, sedangkan 14 *fold* lainnya digunakan sebagai data *training.* Hasil grafik dari pemilihan nilai *k-15* menunjukkan terdapat satu titik dimana akurasi tertinggi diperoleh yakni pada nilai *fold* 6. Sedangkan akurasi rendah diperoleh pada *fold* 4 dan12. Pada Tabel 4.7 ditunjukkan hasil dari penguiian kedalam bentuk tabel menggunakan metode *k-fold cross validation.*

Tabel 4.7 Hasil Pengujian Menggunakan 15-*fold cross validation*

Fold ke-	Hasil Akurasi
1	73.13%
2	74.62%
3	62.68%
4	64.17%
5	73.13%
6	62.68%
7	70.14%
8	67.16%
9	59.7%
10	71.21%
11	68.18%
12	46.96%
13	63%
14	63.63%
15	69.69%
rata-rata	65,5%

Dapat dilihat pada Tabel 4.7 menunjukkan pengujian menggunakan metode cross validasi dengan nilai *k* yang digunakan sebesar 15. Berdasarkan hasil pengujian ini nilai maksimal akurasi yang diperoleh sebesar 73.13% sedangkan nilai terendah yang diperoleh sebesar 46.96% yang diperoleh pada iterasi ke 12. Tabel 4.7 merepresentasikan nilai akurasi pada setiap iterasi yang berbeda-beda. Hasil dari pengujian tersebut nilai rata-rata akurasi yang diperoleh yakni sebesar 65,5%.

Dari perolehan nilai rata-rata akurasi tersebut digunakan untuk indikator performa model dalam mengklasifikasikan data. Selanjutnya pengujian ketiga dilakukan dengan pemilihan nilai *k-20.* Tujuannya untuk mengetahui nilai akurasi yang lebih baik dari pengujian yang sebelumnya dilakukan yakni *k-15*. Hasil pengujian ketiga ini dilakukan dengan metode *k-fold cross validation* dengan nilai *k-20* ditunjukkan pada Gambar 4.4 berikut. Hasil pengujian tersebut ditunjukkan dengan grafik pada Gambar 4.4 berikut.



Gambar 4.4 Grafik model 10-*fold cross validation*

Gambar 4.2, Gambar 4.3 dan Gambar 4.4 merupakan grafik untuk pengujian dengan model *k-fold cross validation.* Dalam grafik tersebut menunjukkan bahwa

dataset dibagi menjadi beberapa kali yaitu nilai *k*-20, *k*-15 dan *k*-10, dimana subset yang saling *overlap* dibagi dengan ukuran yang sama. Hasil grafik dari pemilihan nilai *k-20* menunjukkan terdapat satu titik dimana akurasi tertinggi diperoleh pada rentang nilai *fold* ke 10.0. Sedangkan akurasi rendah diperoleh pada rentang ke 5.0 sampai 7.5. Pada Tabel 4.8 ditunjukkan hasil dari penguiian menggunakan metode *k-fold cross validation.*

Tabel 4.8 Hasil Pengujian Menggunakan 10-*fold cross validation*

Fold ke-	Hasil Akurasi
1	78%
2	68%
3	76%
4	62%
5	64%
6	62%
7	78%
8	62%
9	76%
10	70%
11	62%
12	64%
13	70%
14	70%
15	64%
16	48%
17	58%
18	74%
19	66%
20	69%
rata-rata	67,5%

Dapat dilihat pada Tabel 4.8 menunjukkan pengujian menggunakan *k-fold cross validation* dengan nilai *k* yang digunakan sebesar 20. Hasil pengujian ini nilai maksimal akurasi yang diperoleh sebesar 78%, namun untuk nilai terendah yang diperoleh sebesar 48% dengan rata-rata akurasi 67,5%. Dari Tabel tersebut merepresentasikan dalam tabel yang artinya dalam tabel tersebut menunjukkan nilai akurasi pada setiap iterasi yang diperoleh berbeda-beda. Dari Tabel 4.8 bahwa rata-

rata akurasi yang diperoleh sebesar 67,5%. Dari perolehan nilai rata-rata akurasi tersebut digunakan untuk indikator performa model dalam mengklasifikasikan data.

Hasil dari pengujian ketiga yang telah dilakukan yakni dengan nilai *k-20* menunjukkan bahwa pada pemilihan nilai *k* yang digunakan memperoleh nilai akurasi yang baik yakni sebesar 78% dan perolehan rata-rata sebesar 67,5%. Hasilnya menunjukkan bahwa akurasi yang diperoleh merupakan nilai akurasi yang paling baik diantara pengujian lain yang sebelumnya dilakukan. Untuk mengetahui hasil dari perbandingan nilai akurasi pada masing-masing pemilihan nilai *k* pada metode *k-fold cross validation* dapat dilihat pada Tabel 4.9.

Tabel 4.9 Hasil perbandingan dengan *k-fold cross validation*

Ket.	***k*-10**	***k*-15**	***k*-20**
Min	63%	47%	48%
Max	72%	75%	78%
Rata-rata	66,5%	65,5%	67,5%

Tabel 4.9 hasil perbandingan yang telah dilakukan menunjukkan bahwa akurasi terbaik diperoleh sebesar 78% terjadi pada itarasi ke 1. Hal ini terjadi karena dalam pembagian *dataset* nilai akurasi yang diperoleh mengalami peningkatan. Sedangkan akurasi terendah terjadi pada iterasi ke 16 yakni sebesar 48% dengan hasil rata-rata yang diperoleh dengan pemilihan nilai *k*-20 yakni sebesar 67,5%. Hasil dari perbandingan uji coba dengan metode *k-fold cross validation* dapat dilihat Pada Tabel 4.9 yang menunjukkan bahwa semakin sering dilakukan iterasi pada *fold* maka nilai akurasi yang diperoleh semakin tinggi. Terbukti bahwa pengujian yang dilakukan dengan *k-10, k-15* dan *k-20* nilai akurasi yang paling tinggi dihasilkan dari pengujian *k-fold cross validation* dengan nilai *k-20.*

Tabel 4.9 hasil perbandingan dengan metode *k-fold cross validation* dapat dilihat nilai akurasi yang diperoleh menunjukkan bahwa setiap *fold* pada masing-masing nilai *k* memiliki hasil akurasi yang berbeda. Hal ini terjadi karena pada dasarnya metode *k-fold cross validation* tidak menunjukkan bahwa semakin sering dilakukan iterasi pada *fold* maka semakin tinggi nilai akurasi yang diperoleh. Hal ini terjadi karena model yang digunakan pada penelitian ini tidak stabil. Pada Tabel 4.9 menunjukkan bahwa pada *k-fold cross validation* bahwa penggunaan nilai *k*-10 menunjukkan bahwa nilai max yang diperoleh sebar 72%, nilai min sebesar 63% dengan rata-rata yang diperoleh sebesar 66,5% menunjukkan bahwa model yang diuji memiliki kinerja yang bervariasi pada setiap percobaan namun secara keseluruhan memiliki kinerja yang baik dengan rata-rata yang diperoleh sebesar 66,5% dibandingkan dengan nilai min, max dan rata-rata yang diperoleh dari percobaan *k-15* dan *k-20* yang menunjukkan berbedaan yang signifikan terhadap nilai akurasi min dan max yang diperoleh.

Terjadinya perbedaan nilai akurasi pada masing-masing nilai *k* yang dipilih terjadi karena masing-masing *fold* mempelajari pola lebih baik dari data sehingga dapat mempengaruhi hasil akurasi, sedangkan *fold* yang lain tidak mempelajari pola dengan efektif yang mana akan menghasilkan akurasi yang rendah. Oleh karena itu, dengan teknik cross validasi digunakan untuk mendapatkan estimasi yang lebih akurat dari kinerja model pada *dataset* tertentu dengan mempertimbangkan variasi dalam pembagian data *training* dan data *testing*.

4.4 Pembahasan

Hasil dari skenario uji coba yang dilakukan pada 1000 teks komentar *youtube*, setelah dilakukan pelabelan secara manual dan divalidasi oleh ahli bahasa diketahui bahwa data berhasil diklasifikasikan pada kelas positif, negatif dan netral. Hasil dari pelabelan yang telah dilakukan telah divalidasi oleh ahli bahasa dengan indikator klasifikasi yang ditentukan. Hasilnya diperoleh bahwa klasifikasi pada kelas positif memiliki jumlah paling banyak dibandingkan data dengan pada kelas negatif yang berada pada diurutan kedua dan data dengan kelas netral memiliki jumlah yang paling sedikit. Hal ini terjadi karena pelabelan yang dilakukan didasarkan pada indikator klasifikasi yang telah ditentukan oleh validator berdasarkan referensi yang digunakan.

Data hasil crawling manual yang dilakukan diperoleh sebanyak 1000 komentar *youtube* digunakan untuk dilakukan pengujian. Hasil uji coba pertama yang dilakukan dengan pembagian rasio terhadap *dataset training* dan *testing* dengan nilai pembagian yakni sebesar 9:10, 8:2, 7:3, 6:4 dan 5:5. Hasil dari percobaan pertama menunjukkan bahwa diperoleh akurasi tertinggi sebesar 74% pada rasio pembagian data 9:1. Hal ini menunjukkan bahwa dari seluruh data yang digunakan untuk pengujian data *testing* yang digunakan sebanyak 10% dari total data dan 90% data digunakan untuk proses *training*. Pada rasio pembagian data 9:1 diperoleh hasil nilai akurasi diperoleh sebesar 90.3%. Dalam penelitian ini pengujian yang dilakukan terhadap *dataset training* dan *testing* dilakukan secara acak yakni dengan menggunakan nilai random state sebesar 2002. Tujuan pengujian secara acak dilakukan untuk mendapatkan akurasi terbaik dengan melihat data yang sebelumnya belum pernah dilihat. Nilai akurasi terbaik diperoleh sebesar 74%

sehingga hal ini menunjukkan bahwa sebanyak 74% diantaranya diprediksi dengan benar oleh model.

Hasil uji coba awal yang dilakukan, akan dilanjutkan dengan pengujian kedua yakni dengan teknik cross validasi. metode ini digunakan untuk memvalidasi model *machine learning* dengan membagi *dataset*, dengan pemilihan nilai yang ditentukan yakni *k*-20, *k*-15 dan *k*-10 untuk melakukan perbandingan dari ketiga nilai yang digunakan dalam mendapatkan akurasi terbaik. Pemilihan nilai *k*-20 artinya dengan ini dilakukan iterasi sebanyak 20 kali. Dimana setiap subset pernah digunakan sebagai data pengujian dan pelatihan. Setiap iterasi menghasilkan akurasi yang berbeda karena model dilatih pada subset yang berbeda. Dengan pengujian ini diperoleh tiga perbandingan dengan masing-masing nilai yang berbeda setiap *k*, hasil yang menunjukkan akurasi tertinggi terdapat pada pemilihan nilai *k*-20 dengan nilai maksimal akurasi yang diperoleh sebesar 78% terjadi pada iterasi ke 1 dan nilai minimal yang diperoleh sebesar 48% pada iteras yang ke 16 dengan peroleh nilai rata-rata akurasi sebesar 67.5%. Dari perolehan nilai rata-rata akurasi tersebut yakni sebesar 67.5% akan digunakan sebagai indikator performa model dalam mengklasifikasikan data

Pengujian ini menunjukkan bahwa penggunaan teknik *k-fold cross validation* dapat membantu meningkatkan jumlah akurasi sehingga mempengaruhi performa sistem. Selain itu, penggunaan teknik ini dapat membantu mengurangi terjadinya model yang terlalu kompleks karena terlalu menyesuaikan diri dengan data *training* yang spesifik karena dilakukan beberapa kali pengujian *dataset*.

Dari hasil evaluasi model yang diukur selama uji coba, dapat disimpulkan bahwa sistem berhasil secara efektif melakukan klasifikasi terhadap komentar

media sosial *youtube* menjadi tiga kelas yakni kelas positif, negatif dan netral dengan menggunakan metode *Naïve Bayes Classifier* dengan rasio pembagian data set 9:1. Kemudian dilanjutkan dengan pengujian *k-fold cross validation*. Metode ini membandingkan pemilihan nilai *k* sebanyak *k*-20, *k-15* dan *k-10* untuk dicari nilai akurasi terbaik pada *dataset.* Dari beberapa pengujian yang telah dilakukan akurasi yang berhasil diperoleh sebesar 78% dengan rata-rata yang diperoleh sebesar 67.5% sehingga sebesar 67.5% digunakan sebagai indikator performa model dalam mengklasifikasikan data

Dalam penelitian ini, hasil akurasi yang diperoleh dengan metode *Naïve Bayes Classifier* sangat dipengaruhi pada beberapa hal diantaranya penggunaan fitur pada *dataset* karena dengan fitur yang tepat membantu meningkatkan kualitas model, rasio pembagian data pada klasifikasi, proses *preprocessing* dengan tepat karena proses *preprocessing* data sangat penting untuk memperbaiki kualitas data sebelum dilakukan klasifikasi. Penggunaan metode *k-fold cross validation* memberikan dampak yang signifikan terhadap akurasi model.

Ketepatan akurasi bergantung pada langkah *preprocessing* yang akurat dan data yang digunakan, yang dihasilkan dari proses pelabelan data *training.* Hal ini disebabkan oleh fakta bahwa metode *Naïve Bayes Classifier* (NBC) sebagai bagian teknik pembelajaran terawasi (*supervised learning)* (Oktasari et al., n.d.), dimana hasil prediksi kelas yang dilakukan akan didasarkan pada pola yang dihasilkan dari pemrosesan data latih (*training)* yang diimplementasikan pada model. Model mempelajari pola dan fitur dari data yang sudah diberi label sehingga dapat mengklasifikasikan data baru ke dalam kategori yang tepat untuk data yang belum diberi label.

Hasil sistem klasifikasi yang dilakukan, harapannya dapat meningkatkan kinerja klasifikasi komentar *youtube* dan menghasilkan prediksi yang lebih akurat. Sistem dapat membantu memprediksi komentar *youtube* menjadi komentar yang mengandung komentar positif, komentar negatif dan komentar netral. Serta performa sistem yang telah dibuat untuk model yang digunakan akan diukur dengan nilai akurasi, presisi dan *recall* yang diperoleh. Melalui pengukuran kinerja sistem yang telah dilakukan, dapat membantu mengurangi risiko kesalahan dalam pengambilan keputusan dalam klasifikasi, karena adanya potensi kesalahan dalam informasi yang digunakan dalam analisis data yang dilakukan secara manual.

4.5 Integrasi Islam

Hasil penelitian yang dilakukan penulis, sistem dapat diimplementasikan untuk keperluan opinion mining pad text mining. Konsep Muamalah diterapkan dalam penelitian ini, yang mencakup *Mumalah Ma'Allah* atau hubungan manusia dengan Allah SWT, hubungan manusia dengan sesama manusia *Muamalah Ma'a an-nas* dan *Muamalah Ma'a alam* yakni hubungan manusia dengan alam. Beberapa konsep Muamalah yang diterapkan dalam penelitian ini antara lain yakni *Mumalah Ma'Allah* dan *Muamalah Ma'a an-nas.*

4.5.1 *Muamalah Ma'a Allah*

Dalam surah Az-Zumar Ayat 18 Al Qur'an yang berbunyi:

الَّذِينَ يَسْتَمِعُونَ الْقَوْلَ فَيَتَّبِعُونَ أَحْسَنَهُ ۚ أُولَـٰئِكَ الَّذِينَ هَدَاهُمُ اللَّهُ ۖ وَأُولَـٰئِكَ هُمْ أُولُو الْأَلْبَابِ

"(yaitu) mereka yang mendengarkan perkataan lalu mengikuti apa yang paling baik di antaranya. Mereka itulah orang-orang yang telah diberi Allah petunjuk dan mereka itulah orang-orang yang mempunyai akal" (QS. Az-Zumar:18)

Berdasarkan penjelasan dalam "tafsir *Fi Zhilalil Al-Quran*", *Ulul Albab* itu tidak hanya berpikir tentang alam fisik, botani dan sejarah. Merekapun ternyata mempunyai karakteristik yang berkaitan tidak hanya dengan aktivitas pikirannya, melainkan dengan amal konkritnya. Dalam hal ini, bersikap kritis dalam menerima pengetahuan atau mendengar pembicaraan orang lain. *Ulul Albab* memiliki kemampuan menimbang ucapan, teori, proporsi dan atau dalil yang dikemukakan orang lain (QS. Al-Zumar: 18)

Ulul Albab, mendengar perkataan yang telah mereka dengar. Lalu *qalbu* mereka memungut bagian tuturan yang baik dan membuang sisanya. Sesungguhnya Allah Swt mengetahui kebaikan yang ada pada jiwa mereka. Maka, Dia menunjukkan mereka untuk menyimak dan merespon perkataan yang baik. Petunjuk itu adalah petunjuk Allah Swt. Allah Swt memberikan sifat kepada mereka tiga hal: bertahudi kepada Allah Swt atau menjauhi *thagbut,* kembali kepada Allah Swt, dan mengikuti perkataan yang paling benar (wahyu). Yaitu bahwa perkataan-perkataan yang mereka dengarkan, mereka memperhatikan baik-baik, pasang telinga menyalakan mata dan sambut dengan penuh kesadaran, lalu mengikuti mana yang terbaik.

Hamka mengutip satu tafsir dari Ibnu Abbas: "didengarkannya ada kata-kata yang baik dan ada yang tidak baik untuk didengar. Maka yang dipegangnya ialah yang baik, sedang yang tidak baik didengar itu tidak dipercakapkannya. Hasbi As-Shidieqy mengutip ayat di atas dengan menguraikan,; Ya Muhammad, gembirakanlah hamba-hamba-Ku yang menjauhi diri dari penyembahan selain Allah Swt dan kembali kepada Tuhan, serta mau mendengarkan perkataan yang benar, lalu mengikuti mana yang lebih utama untuk diterima dan mana yang dapat menunjuk kepada kebenaran, bahwa mereka akan diberikan oleh Allah Swt nikmat

yang kekal di dalam surge (*jannatun na'im)*. Merekalah orang-orang yang ditaufiqkan oleh Allah Swt kepada kebenaran, bukan orang-orang yang berpaling dari kebenaran dan menyembah berhala. Orang itulah yang mempunyai akal yang sejahtera dan fitrah yang sehat yang tidak dapat ditundukkan oleh hawa nafsu. Karena itu senantiasa mereka memilih mana yang lebih baik untuk agama dan dunianya. Allah Swt memuji mereka bahwa mereka adalah orang-orang yang kritis dalam beragama, mereka dapat membedakan antara yang baik dan yang lebih baik, dan antara utama dengan yang lebih utama. Orang-orang yang mendengarkan perkataan yang baik dan mengerjakan yang baik dari perkataan itu adalah orang mendapat taufiq dari Allah Swt dan selalu menggunakan akal pikirannya.

4.5.2 *Muamalah Ma'a an-Nas*

Dari Abu Hurairah RA bahwa Rasulullah SAW bersabda, "*Barang siapa yang beriman kepada Allah dan hari akhir, maka hendaklah ia tidak menyakiti tetangganya. Dan barang siapa yang beriman kepada Allah dan hari akhir,maka hendaklah ia memuliakan tamunya. Dan barang siapa yang beriman kepada Allah dan hari akhir, maka hendaklah ia berkata-kata yang baik atau hendaklah ia diam."* (HR Bukhari & Muslim).

Terdapat di dalam Al-Qur'an yang membahas tentang ahklak dalam surat Al-Qalam ayat 4: Sebagaimana firman Allah *subhanahu wa ta'ala* dalam surah Al-Qalam Ayat 4 dalam Al Qur'an yang berbunyi:

وَاِنَّكَ لَعَلٰى خُلُقٍ عَظِيْمٍ

"Dan sesungguhnya kamu benar-benar berbudi pekerti yang agung"

Ahklak yang baik menempati kedudukan yang sangat tinggi dibandingkan dengan ilmu-ilmu lainnya, mengingat pentingnya keberadaan perilaku dengan

ahklak yang baik, hakikat nilai-nilai moral sudah seharusnya mendapat tempat dalam kehidupan agar manusia dapat memilih dan menentukan sesuatu benar dan yang salah, khususnya dalam berbicara.

Mengucapkan ucapan buruk telah dijelaskan oleh Allah dalam firman-Nya. Sebagaimana dalam Firman Allah dalam surah An-Nisa' (04: [148]:

لَّا يُحِبُّ ٱللَّهُ ٱلْجَهْرَ بِٱلسُّوٓءِ مِنَ ٱلْقَوْلِ إِلَّا مَن ظُلِمَ ۚ وَكَانَ ٱللَّهُ سَمِيعًا عَلِيمًا

"Allah tidak menyukai ucapan buruk, (yang diucapkan) dengan terus terang kecuali oleh orang yang dianiaya. Allah adalah Maha Mendengar lagi Maha Mengetahui".

Allah memerintahkan kita untuk berkata-kata yang baik, karena inilah aturan dasar dalam berbicara dengan orang lain. Allah berfirman dalam Q.S Al-Isra' ayat 53 yang berbunyi:

وَقُلْ لِّعِبَادِيْ يَقُوْلُوا الَّتِيْ هِيَ اَحْسَنُۗ اِنَّ الشَّيْطٰنَ يَنْزَغُ بَيْنَهُمْۗ اِنَّ الشَّيْطٰنَ كَانَ لِلْاِنْسَانِ عَدُوًّا مُّبِيْنًا

"Dan katakanlah kepada hamba-hamba-Ku: "Hendaklah mereka mengucapkan perkataan yang lebih baik (benar). Sesungguhnya syaitan itu menimbulkan perselisihan di antara mereka. Sesungguhnya syaitan itu adalah musuh yang nyata bagi manusia..". (QS. Al-Isra': 53).

Ajaran Islam amat sangat serius memperhatikan soal menjaga lisan sehingga Rasulullah Shallallaahu alaihi wa Salam bersabda:

"Barang siapa yang memberi jaminan kepadaku (untuk menjaga) apa yang ada antara dua janggutnya (lisan) dan apa yang ada antara dua kakinya (kemaluannya) maka aku menjamin surga untuknya." (HR. Al-Bukhari).

Ibnu Abbas (dalam Tafsir Katsir), ia mengatakan: "Katakanlah hal ini adalah hak sebagaimana yang dikatakan kepada kalian." Sedang Ikrimah mengatakan: "Katakanlah "tiada ilah yang hak selain Allah". Dan Qatadah mengatakan: "Hal itu berarti, "Hapuskanlah kesalahan-kesalahan kami. *Niscaya*

kami ampuni kesalahan-kesalahanmu. Dan kelak Kami akan menambah (pemberian Kami) kepada orang-orang yang berbuat baik." Ini merupakan jawaban atas perintah sebelumnya. Artinya, jika kalian mengerjakan apa yang Kami perintahkan, maka Kami akan mengampuni kesalahan-kesalahan kalian dan kami lipatgandakan kebaikan atas kalian."

Mereka diperintahkkan untuk tunduk kepada Allah ketika memperoleh kemenangan, baik dalam perbuatan maupun ucapan. Selaint itu hendaklah mereka mengakui dosa-dosa yang telah diperbuatnya, memohon ampunan atasnya, mensyukuri nikmat, serta bersegera melakukan semua perbuatan yang disukai Allah Swt sebagimana firman-Nya:

إِذَا جَاءَ نَصْرُ اللَّهِ وَالْفَتْحُ وَرَأَيْتَ النَّاسَ يَدْخُلُونَ فِي دِينِ اللَّهِ أَفْوَاجًا فَسَبِّحْ بِحَمْدِ رَبِّكَ وَاسْتَغْفِرْهُ ۚ إِنَّهُ كَانَ تَوَّابًا

"Apabila telah datang pertolongan Allah dan kemenangan. Dan kamu menyaksikan menusia masuk agama islam secara berbondong-bondong. Maka bertasbihlah dengan memuji Rabb-mu serta memohonlah ampunan kepada-Nya. Sesungguhnya Dia adalah Mahapenerima Taubat." (QS. An-Nasr: 1-3)

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini diimplementasikan metode *Naïve Bayes Classifier* (NBC) untuk model *Multinomial* dengan bahasa pemrograman *python* pada *jupyter notebook* untuk proses klasifikasi. Hasil *crawling dataset* penelitian sebanyak 1000 data komentar *youtube* diklasifikasi sebanyak 404 komentar kedalam kelas positif, 309 komentar kedalam kelas negatif dan sebanyak 287 komentar sebagai kelas netral. Data *training* yang diimplementasikan dalam sistem digunakan untuk membuat model klasifikasi sehingga hasil yang diperoleh menunjukkan bahwa sistem mengklasifikasikan secara otomatis komentar *youtube* yang mengandung komentar yang bersifat positif, komentar yang bersifat negatif dan komentar yang bersifat netral berdasarkan nilai probabilitas yang diperoleh. Data *testing* yang digunakan pada sistem digunakan untuk menguji kinjerja sistem dan mengetahui performa model yang diprediksi dengan parameter nilai *accuracy* yang diperoleh sebesar 78%, sebagai *multiclass* masing-masing kelas memiliki nilai *precision* pada kelas positif sebesar 69%, kelas negatif sebesar 87% dan kelas netral sebesar 67%. Kemudian untuk nilai *recall* pada kelas positif diperoleh sebesar 86%, pada kelas negatif sebesar 84% dan pada kelas netral sebesar 40%.

5.2 Saran

Dalam penelitian ini penulis menyadari bahwa penelitian yang dilakukan tentu belum sempurna. Terdapat kekurangan yang diperlukan untuk memperbaiki performa sistem. Berikut beberapa saran penulis yang perlu dilakukan untuk penelitian selanjutnya:

1. Mengoptimalkan proses *preprocessing* untuk menangani kata singkatan, memperbaiki proses *stemming* yang masih ada kata imbuhan dan kata tidak baku didalamnya yang berpengaruh terhadap sistem.
2. Menambahkan jumlah *dataset* yang digunakan, karena dalam *Naïve Bayes Classifier* membutuhkan banyak data *training* untuk dapat menghasilkan model yang akurat. Penggunaan jumlah *dataset* yang sedikit dapat menyebabkan model yang digunakan teralu kompleks karena terlalu menyesuaikan dengan data *training.* Selain itu, jika juga dapat membuat model terlalu sederhana sehingga mengurangi performa metode yang digunakan.
3. Menggunakan Kamus kata unik seperti (Lexicon) sebagai alternatif proses pelabelan yang dilakukan manual. Dalam hal ini kamus akan menghitung nilai sentiment dari kata yang terdapat pada suatu teks yakni dengan memberikan bobot sentiment yang diberikan untuk masing-masing kata.
4. Menggunakan teknik pengelompokan kata seperti metode *Clustering*. Tujuannya untuk jumlah kata unik yang berlebihan yakni dengan menggabungkan kata-kata yang mirip untuk mengurangi jumlah kata unik yang terdapat pada pengujian menggunakan data training yang sebelumnya telah dilakukan.

DAFTAR PUSTAKA

hendra, J., & Laugu, N. (2020). Eksistensi Media Sosial,Youtube, Instagram dan Whatsapp Ditengah Pandemi Covid-19 Dikalangan Masyarakat Virtual Indonesia. *Baitul Ulum: Jurnal Ilmu Perpustakaan Dan Informasi*, *4*(1). https://databooks.com

Aditya Quantano Surbakti, Regiolina Hayami, & Januar Al Amien. (2021). Analisa Tanggapan Terhadap Psbb Di Indonesia Dengan Algoritma Decision Tree Pada Twitter. *Jurnal CoSciTech (Computer Science and Information Technology)*, *2*(2), 91–97. https://doi.org/10.37859/coscitech.v2i2.2851

Darujati, C., & Bimo Gumelar, A. (2012). *PEMANFAATAN TEKNIK SUPERVISED UNTUK KLASIFIKASI TEKS BAHASA INDONESIA* (Vol. 16, Issue 1).

Destuardi, I., & Sumpeno, S. (2009). Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes. In *Seminar Nasional Pascasarjana IX-ITS*.

Dwi Herlambang, A., & Hadi Wijoyo, S. (2019). *ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI SUMBER BELAJAR BERBASIS TEKS PADA MATA PELAJARAN PRODUKTIF DI SMK RUMPUN TEKNOLOGI INFORMASI DAN KOMUNIKASI*. *6*(4), 431–436. https://doi.org/10.25126/jtiik.201961323

Effendy, Onong Uchjana. 1993. Ilmu, Teori dan Filsafat Komunikasi. Bandung: PT Mandar Maju

Farah Zhafira, D., Rahayudi, B., & Korespondensi, P. (2021). *ANALISIS SENTIMEN KEBIJAKAN KAMPUS MERDEKA MENGGUNAKAN NAIVE BAYES DAN PEMBOBOTAN TF-IDF BERDASARKAN KOMENTAR PADA YOUTUBE* (Vol. 2, Issue 1).

Kaida Palma, B., Triantoro Murdiansyah, D., & Astuti, W. (n.d.). *Klasifikasi Teks Artikel Berita Hoaks Covid-19 dengan Menggunakan Algotrima K-Nearest Neighbor*.

Khairunnisa, S., Adiwijaya, A., & Faraby, S. al. (2021). Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19). *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 406. https://doi.org/10.30865/mib.v5i2.2835

Kusumawati, N. D., al Faraby, S., & Dwifebri, M. (n.d.). *Analisis Sentimen Komentar Beracun pad Media Sosial Menggunakan Word2Vec dan Support Vectore Machine (SVM)*. http://j.mp/KteNOK

Mustofa, H., & Mahfudh, A. A. (2019a). Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes. *Walisongo Journal of Information Technology*, *1*(1), 1. https://doi.org/10.21580/wjit.2019.1.1.3915

Mustofa, H., & Mahfudh, A. A. (2019b). Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes. *Walisongo Journal of Information Technology*, *1*(1), 1. https://doi.org/10.21580/wjit.2019.1.1.3915

Ni'matul Rohmah, N. (2020). *Media Sosial Sebagai Media Alternatif Manfaat dan Pemuas Kebutuhan Informasi Masa Pandemik Global Covid 19 (Kajian Analisis Teori Uses And Gratification)*. *4*(1), 1–16. https://www.kompas.com/tren/read/2020/03/29/092500765/update-virus-corona-di-dunia-29-maret--662.073-kasus-di-200-

Oktasari, L., Chrisnanto, Y. H., Program, R. Y., Informatika, S., Matematika, F., Pengetahuan, I., Universitas, A., Yani, J. A., Terusan, J., & Sudirman, J. (n.d.). *TEXT MINING DALAM ANALISIS SENTIMEN ASURANSI MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER*.

Rhomadhona, H., & Permadi, J. (2019). Klasifikasi Berita Kriminal Menggunakan NaÃ¯ve Bayes Classifier (NBC) dengan Pengujian K-Fold Cross Validation. *Jurnal Sains Dan Informatika*, *5*(2), 108–117. https://doi.org/10.34128/jsi.v5i2.177

Said, B., Yuliana, D., & Pranoto, M. (n.d.). *KLASIFIKASI DATA SMS CENTER BUPATI PAMEKASAN MENGGUNAKAN NAÏVE BAYES DENGAN MAD SMOOTHING*. *SKRIPSI INDAH AMELIA*. (n.d.).

Sriyano, C. S., & Setiawan, E. B. (n.d.). *Pendeteksian Berita Hoax Menggunakan Naive Bayes Multinomial Pada Twitter dengan Fitur Pembobotan TF-IDF*.

Susanto, B. (n.d.). *Editorial Team Editors Layout Editor*. https://ti.ukdw.ac.id/ojs/index.php/informatika/about/editorialTeam

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *4*(2), 437. https://doi.org/10.30865/mib.v4i2.2080

Wahyudi, D., Susyanto, T., Nugroho, D., Studi Teknik Informatika, P., Sinar Nusantara Surakarta, S., & Studi Sistem Informasi, P. (n.d.). *IMPLEMENTASI DAN ANALISIS ALGORITMA STEMMING NAZIEF & ADRIANI DAN PORTER PADA DOKUMEN BERBAHASA INDONESIA*.

Wibawa, A. P., Guntur, M., Purnama, A., Fathony Akbar, M., & Dwiyanto, F. A. (2018). Metode-metode Klasifikasi. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1).

Wijaya, A. P., & Santoso, H. A. (2016). Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government Naïve Bayes Classification on Document Classification to Identify E-Government Content. *Journal of Applied Intelligent System*, *1*(1), 48–55.

Yanti Liliana, D., Maulana, H., & Setiawan, A. (2021). *Data Mining untuk Prediksi Status Pasien Covid-19 dengan Pengklasifikasi Naïve Bayes* (Vol. 7, Issue MEI). www.kaggle.com

LAMPIRAN

LAMPIRAN

Lampiran 1

Sampel Data Klasifikasi

No	Komentar	Validator			Hasil
		Validator I	Validator II	Validator III	
		Kelas			
1	Si bilar itu ga punya harta buang si ngapain dih	negative	negative	negative	negative
2	Mamanya calon Mama mantu idaman. Baik sekali..	positif	Positif	positif	positif
3	Air mata kejahatan	negative	negative	netral	negative
4	Hanya Alloh yang tau isi hati manusia	netral	Netral	netral	netral
5	Yaampun, ibu mas ryan ituu baik bangettt....	positif	Positif	positif	positif
6	Saya salut dgn Lesty	positif	Positif	positif	positif
7	bikin quiz dong, brp kali mamanya yessy blg 'istilahnya"😈	netral	Netral	netral	netral
8	Masya Allah mamahnya aa Ryan baik dan bijak...	positif	Positif	positif	positif
9	Itu suminya tidak tau diri…	negative	negative	negative	negative
10	Keliatan banget ibunya Ryan wanita pintar dan berwibawa.	positif	Positif	positif	positif
11	Angkuh banget keluarga gen ni	negative	negative	negative	negative
12	amar zoni ciri manusia tidak bersyukur.	negative	negative	negative	negative
13	Livy tidak tau diri main peluk2 orang aja	negative	negative	negative	negative
14	Usia tidak selalu mencerminkan kedewasaan	netral	Netral	netral	netral
15	untuk fuji, tetep semangat anak baik!!	positif	Positif	positif	positif
16	Kapokmu kapan zon zoni	negative	negative	negative	negative
17	Alhamdulillah	netral	Positif	netral	netral
18	Sehat dan sukses	positif	Positif	positif	positif
19	Saya dari malang....	netral	Netral	netral	netral
20	Anggap aja bukan jodoh.. karena segala sesuatu jodoh.rejeki.umur.	netral	Netral	netral	netral
21	Begini lah cinta penderitaan tiada akhir	Netral	Netral	negative	netral
Total		21	20	19	-

LAMPIRAN 2

Hasil Perbandingan sistem dengan validator 1

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil Klasifikasi (Pakar)	Hasil
1	Selalu suka sama mama Ala.. Kaya raya tapi humble dan bersahaja	Positif	Positif	1
2	Wanita hebat..... Semoga ada anak saya sukses seperti Bu Ala	Positif	Positif	1
3	Ga usah lebay bu, km seharusnya sadar	Negative	Negative	1
4	Bu Pc Kebanyakan makan uang haram	Negative	Negative	1
5	Tangis nya PC tangis sandiwara	Negative	Negative	1
6	ini orang purah2 sedih dasar tukang bohong	Negative	Negative	1
7	Beliau adalah single parent dan sudah terbiasa hidup dg kerja keras dari kecil	Netral	Netral	1
8	Bu ala keren saya suka wanita hebat ..cerdas	Positif	Positif	1
9	Masya allah tabarakallah..keren luar biasaaaah.	Positif	Positif	1
10	masyaallah ...barokah bu Ala	Netral	Netral	1
11	beliau humble banget	Positif	Positif	1
12	Keren ya single mom mandiri dan suksess	Positif	Positif	1
13	Sehat ya bu ala sekeluarga	Positif	Positif	1
14	Semoga masalahnya cepat selesai. Semoga jd pembelajaran buat kedepannya	Netral	Netral	1
15	chika penjara aja deh bikin ulah aja ngomong nya beda dih gila	negative	Negative	1
16	selamat jalan ibadah puasa ramadhan ya om ded moga sehat sukses	positive	Positive	1
17	netizen indo emang lucu	positive	Netral	0
18	gak ded podcast beda gw emak emak jd emosi liat celoteh hujat karna	negative	Negative	1
19	mas ded orang pinter ya	positive	Positive	1
20	undang prestasi undang nyinyirin mantan	negative	Negative	1
21	terimakasih om ded	positive	Positive	1

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil Klasifikasi (Pakar)	Hasil
22	jgn salah netizen cikhacoba introspeksi	positive	Positive	1
23	sumpah gua ketawa liat om dedi	Positive	Netral	0
24	hadeh chika songong neng gua liat hujat	Negative	Negative	1
25	moga umur mas deddy chika	Positive	positive	1
26	podcastnya om dedy seru enak dengar tamu seru	positive	positive	1
27	stop bela ranah hukum	negative	negative	1
28	ga habis pikirnda harga banget wanita	negative	Negative	1
29	heran mbak wenny ya ga malu aib umbar umbar	negative	Negative	1
30	wanita tdk tau malu hancur rumah tangga orang dr nuntut	negative	Negative	1
31	perempuan murah barang obral mbak wen	Negative	negative	1
32	perempuan ken rusak rumah tangga cik	Negative	Negative	1
33	hukum karma laku ya wen	Negative	Negative	1
34	hadeh mbk weny tua keriput	Negative	Negative	1
35	ta rasain malu lu wen	Negative	Negative	1
36	moga cepat selesai moga jd ajar	Netral	netral	1
37	moga buna pribadi	Positive	Positive	0
38	kena mental bangett bunaa semangatttttt	Positive	Positive	1
39	semangat karya peduli orang maju pantang mundur	Positive	Positive	1
40	betapa rapuh batas manusia	Netral	Netral	1
41	moga hukum indonesia jalan mesti	Positive	Netral	0
42	only in indonesia orang habis scandal langsung kasih panggung view	Netral	Netral	1
43	okin gaul orang kaya gin dewasa biar blajar	Negative	Netral	0

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil Klasifikasi (Pakar)	Hasil
44	muak gw denger nya si okin jati mulu	Negative	Negative	1
45	si okin kayak emang ga ngenali deh gaul orangorang kayak bang den bocah banget pikir	Negative	Negative	1
46	mmbuktikan bungkus arah bad boy	Negative	Negative	1
47	gw sadar okin sadar	Negative	Negative	1
48	point arti inti okin ga	Negative	Negative	1
49	lihat beda pola pikir orang meni muda dg orang meni dewasa	Netral	Netral	1
50	alhamdullilah chel cerai orang kayak	Negative	Positif	0
51	si okin gw ga nangkep blass diomongin densu otak cuman fun single being bad etc	Netral	Netral	1
52	gambar fuckboy meni ya gin labil ujung cerai	Negative	Negative	1
53	liat okin hilang arah lepas rachel	Netral	Netral	1
54	lo liat tolol konten njing	Negative	Negative	1
55	sampe detik gua bom temu sisi ganteng ni orang	Negative	Negative	1
56	okin emg liat bgt bad boy nya	negative	Negative	1
57	okin gak bener ya gak nangkep mulu gemes	negative	Negative	1
58	okin gagal jdi suami gagal ayah	negative	Negative	1
59	moga okin sadar bang denny tampar	netral	Netral	1
60	udah liat okin egois sih	negative	Negative	1
61	susah jatuh cinta orang gampang cipok bobo ranjang	negative	Negative	1
62	blm matang nih niko anak anak bangeeeeet	negative	Negative	1
63	anjiirgemes bgt sm okinditerangi bolak gak masuk otak	negative	Negative	1
64	cakep ngga jamin kualitas	negative	Negative	1
65	okin definisi kalo blom puas main mending nikah dlu	netral	Netral	1

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil Klasifikasi (Pakar)	Hasil
66	okin inickckcksayang banget idupnyakasian	negative	Negative	1
67	denger okin bikin mulesss	negative	Negative	1
68	okin liat ga ngomong bgt anjir kaku beud skakmat	negative	Negative	1
69	pikir okin gitu ya malu dengernya	negative	Negative	1
70	labil sumpah bicara dr dewasa	positive	Negative	0
71	letak ganteng okin sih dlu nyari nggk dpt	negative	Negative	1
72	ngomong lo bom anak orang anak toxic njir	negative	Negative	1
73	liat bego sumpah ngakak	negative	Negative	1
74	muka si okin aneh kalo ketawa	negative	Negative	1
75	masya allah gemesin banget queen lalacerdas	positif	positif	1
76	lala nih asli lucu kocak bangetsehat trus ya lala	positif	Positif	1
77	kali podcast om ded jam rokok extream	netral	Netral	1
78	om deddy emang dewa nya podcast	positif	Positif	1
79	lala apresiasi om deddy pertamakalinya nyanyi dipodcast lagu potong bebek angsa	positif	Positif	1
80	ya allahngakak om deddy ngimbangin lala	positif	Netral	0
81	ga syukur bgt dpt sambung	negative	Negative	1
82	selamat sukses putri hasil hancur keluarga	positif	Negative	1
83	moga putri dina jodoh duda	negative	Negative	1
84	poko parah putri ga dewasa	negative	Negative	1
85	putri dina hrus instrospeksi	negative	Negative	1
86	suka bgt lala sabar pikir positif	positif	Positif	1
87	ade lala gemesinnpintar banget ngomongnya	positif	Positif	1

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil Klasifikasi (Pakar)	Hasil
88	pintar lala sehat dek lala	positif	Positif	1
89	bangun tidur siang disuguhin oceh lala jam mood banget	positif	Positif	1
90	salut ortunya lala gak egoismemaksakan didik anak	positif	Positif	1
91	anak pintergemesin lucu banget	positif	Positif	1
92	hallo lala bikin gemes pinter ngomong	positif	Positif	1
93	hebat om ded orang teman lala pusing	positif	Positif	1
94	om dedy sabar banget ngadapin lala bawel cerdas	positif	Positif	1
95	mama anak cerewet sebelas duabelas lala	netral	Netral	1
96	si billar pansos genjot pamor lesti	negative	Negative	1
97	bilar gatel orang soksokan suami lesty	negative	negative	1
98	bilar modus doang numpang popular harta	negative	Negative	1
99	laki jual tampang numpang tenar cerai aja	negative	Negative	1
100	ga nang kau lupa ehhhh ngabisin nyawa orang eh	negative	Negative	1

Lampiran 3

Hasil Klasifikasi Sistem berdasarkan nilai Probabilitas

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil		
			Positif	Negatif	Netral
1	Selalu suka sama mama Ala.. Kaya raya tapi humble dan bersahaja	Positif	3.18013495758422e-05	2.0976691195679952e-07	3.106261349878014e-07
2	Wanita hebat..... Semoga ada anak saya sukses seperti Bu Ala	Positif	0.8904059867740058	0.020426000610065366	0.0020843013657681473
3	Ga usah lebay bu, km seharusnya sadar	Negative	1.987584348490138e-06	5.899694398784987e-06	3.494544018612766e-07
4	Bu Pc Kebanyakan makan uang haram	Negative	0.0009258529274239513	0.008452138183475323	0.0010421506828840737
5	Tangis nya PC tangis sandiwara	Negative	0.003438882301860391	0.037858535613483214	0.0006252904097304443
6	ini orang purah2 sedih dasar tukang bohong	Negative	1.1790139700217976e-13	3.673414302283626e-12	4.862913032079525e-13
7	Beliau adalah single parent dan sudah terbiasa hidup dg kerja keras dari kecil	Netral	1.4387967091791427e-12	2.878031148512813e-13	1.0348662216051005e-11
8	Bu ala keren saya suka wanita hebat ..cerdas	Positif	3.0637737125260663e-09	2.1977692406825113e-13	5.27015205447042e-14
9	Masya allah tabarakallah..keren luar biasaaaah.	Positif	0.010845705721252003	0.0005282586364672077	0.0012505808194608885
10	masyaallah ...barokah bu Ala	Netral	0.005092191100831732	0.0002934770202595598	0.005627613687573998
11	beliau humble banget	Positif	0.10316646905581173	0.0011152126769863276	0.0010421506828840737
12	Keren ya single mom mandiri dan suksess	Positif	1.4744788675667851e-09	7.613831403390645e-11	2.796869695307452e-10
13	Sehat ya bu ala sekeluarga	Positif	0.2749783194449136	0.015260805053497114	0.1496007305280088
14	Semoga masalahnya cepat selesai. Semoga jd pembelajaran buat kedepannya	Netral	0.0007935882235062441	0.0001173908081038239 4	0.005002323277843553
15	chika penjara aja deh bikin ulah aja ngomong nya beda dih gila	negative	6.593306457054747e-31	2.3539856649365753e-27	1.6249685481385398e-32

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil		
			Positif	Negatif	Netral
16	selamat jalan ibadah puasa ramadhan ya om ded moga sehat sukses	positive	6.765734044890827e-24	8.189579700158992e-30	2.863610065965758e-26
17	netizen indo emang lucu	positive	2.1221619600568434e-06	6.29863302142919e-07	1.865609680975486e-06
18	gak ded podcast beda gw emak emak jd emosi liat celoteh hujat karna	negative	1.141542032841627e-35	1.512897724312093e-33	6.528134515130481e-35
19	mas ded orang pinter ya	positive	1.8819406941970363e-05	3.567255296989053e-08	3.248401121532885e-07
20	undang prestasi undang nyinyirin mantan	negative	1.0548048909274035e-10	1.1726677504894977e-09	1.0300122462252512e-10
21	terimakasih om ded	positive	0.05398722149240245	0.0001174695058496544	0.006883322385432469
22	jgn salah netizen cikhacoba introspeksi	positive	2.109609781854807e-10	1.6417348506852967e-10	1.5450183693378766e-10
23	sumpah gua ketawa liat om dedi	Positive	2.2059205619449524e-12	3.5368665396801207e-13	1.4778436576275344e-12
24	hadeh chika songong neng gua liat hujat	Negative	6.306301345268718e-17	2.9868980160837486e-15	1.1126246246019458e-17
25	moga umur mas deddy chika	Positive	3.0853043059626556e-07	2.216342048425151e-09	3.1415373509870156e-08
26	podcastnya om dedy seru enak dengar tamu seru	positive	2.7136646940639534e-17	1.30768983606331e-22	7.81897695372441e-20
27	stop bela ranah hukum	Negative	2.411547681882777e-07	5.66876971928627e-06	1.2955622784551986e-07
28	ga habis pikirnda harga banget wanita	Negative	8.798851624170414e-12	4.529809071931059e-11	1.4394581080787672e-12
29	heran mbak wenny ya ga malu aib umbar umbar	Negative	2.6473642208838e-23	3.1674631821647812e-18	7.113066430077417e-24
30	wanita tdk tau malu hancur rumah tangga orang dr nuntut	Negative	4.7823353667578315e-24	1.053753875185834e-21	2.120665767973391e-24
31	perempuan murah barang obral mbak wen	Negative	6.007408937758587e-15	3.065284334389438e-12	1.599397897865297e-15
32	perempuan ken rusak rumah tangga cik	Negative	1.441778145062061e-14	8.174091558371834e-12	9.596387387191782e-14
33	hukum karma laku ya wen	Negative	2.0304994150352517e-08	3.9514212520494117e-07	6.534140186991437e-09

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil		
			Positif	Negative	Netral
34	hadeh mbk weny tua keriput	Negative	1.977759170488882e-11	8.443207803524385e-10	2.253151788617737e-11
35	ta rasain malu lu wen	Negative	8.790040757728367e-12	1.0319476204307579e-08	1.287515307781564e-11
36	moga cepat selesai moga jd ajar	Netral	2.2335546430586428e-10	9.549539657136326e-13	5.1419746753546475e-09
37	moga buna pribadi	Positive	0.017201810769638035	0.0005286127763234448	0.006361858568354253
38	kena mental bangett bunaa semangatttttt	Positive	1.5822073363911055e-10	5.863338752447488e-12	6.43757653890782e-12
39	semangat karya peduli orang maju pantang mundur	Positive	9.921914116556115e-15	5.706889213563892e-18	6.914167310026376e-17
40	betapa rapuh batas manusia	Netral	4.823095363765554e-08	1.1809936915179731e-07	1.088272313902367e-06
41	moga hukum indonesia jalan mesti	Positive	1.8283284776074997e-07	4.2743739505342206e-09	1.4136918079441573e-07
42	only in indonesia orang habis scandal langsung kasih panggung view	Netral	5.005060561755973e-27	1.2727971783997097e-27	7.634396764704207e-25
43	okin gaul orang kaya gin dewasa biar blajar	Negative	4.944456699919471e-19	1.6156769462529415e-16	4.328869968016513e-17
44	muak gw denger nya si okin jati mulu	Negative	5.187888385710502e-20	1.366405109702553e-16	8.529793040426628e-20
45	si okin kayak emang ga ngenali deh gaul orangorang kayak bang den bocah banget pikir	Negative	1.1652503912497489e-40	2.6050986651589196e-37	7.974386547629382e-41
46	mmbuktikan bungkus arah bad boy	Negative	2.411547681882777e-08	3.149316510714595e-07	5.182249113820795e-08
47	gw sadar okin sadar	Negative	1.5072173011767352e-06	4.4287263431924005e-05	4.1976217821948446e-06
48	point arti inti okin ga	Negative	5.713526492523438e-11	9.146808453818082e-09	1.931272961672346e-10
49	lihat beda pola pikir orang meni muda dg orang meni dewasa	Netral	1.6347173608961029e-25	4.990525129784386e-26	2.3294868041171358e-24
50	alhamdullilah chel cerai orang kayak	Negative	7.260573665883628e-09	9.604148876508988e-09	1.8668971962832673e-10
51	si okin gw ga nangkep blass diomongin densu otak cuman fun single being bad etc	Netral	3.325960871270869e-45	2.295375938521135e-42	1.2759018476207012e-41

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil		
			Positif	Negatif	Netral
52	gambar fuckboy meni ya gin labil ujung cerai	Negative	1.0056522101531125e-20	5.71198920392454e-18	4.7239652428751646e-20
53	liat okin hilang arah lepas rachel	Netral	1.922370860082748e-14	7.073733079360242e-13	1.0747953873654795e-12
54	lo liat tolol konten njing	Negative	2.6370122273185087e-11	1.055400975440548e-09	1.351891073170642e-10
55	sampe detik gua bom temu sisi ganteng ni orang	Negative	8.650771700241829e-22	2.461121353663388e-21	2.9874879006325155e-23
56	okin emg liat bgt bad boy nya	negative	4.190186893856327e-15	2.2935109766357354e-13	4.0054486485670037e-16
57	okin gak bener ya gak nangkep mulu gemes	negative	9.681078609740627e-18	8.330834254116726e-16	6.73379772802569e-18
58	okin gagal jdi suami gagal ayah	negative	3.203951433471247e-14	1.6210638306867223e-11	4.030482702620548e-13
59	moga okin sadar bang denny tampar	netral	5.7270631873298535e-12	1.7684332698400601e-12	7.902625013352431e-11
60	udah liat okin egois sih	negative	1.898648803669326e-09	4.1793878627445703e-07	1.6222692878047706e-09
61	susah jatuh cinta orang gampang cipok bobo ranjang	negative	2.8254038285254115e-20	3.263993830814022e-19	3.435611085727392e-20
62	blm matang nih niko anak anak bangeeeeet	negative	1.130462907818541e-15	7.088483193019916e-14	1.5894637494313512e-16
63	anjiirgemes bgt sm okinditerangi bolak gak masuk otak	negative	9.300952716125086e-20	1.8642426302918545e-18	5.5285695632394825e-21
64	cakep ngga jamin kualitas	negative	3.617321522824165e-08	4.199088680952793e-07	5.182249113820795e-08
65	okin definisi kalo blom puas main mending nikah dlu	netral	1.1636765806082634e-24	2.1035225244986224e-23	7.947012149465804e-22
66	okin inickckcksayang banget idupnyakasian	negative	7.837529966119026e-07	1.495925342589433e-05	7.773373670731191e-07
67	denger okin bikin mulesss	negative	4.823095363765554e-07	9.447949532143786e-06	3.109349468292476e-07
68	okin liat ga ngomong bgt anjir kaku beud skakmat	negative	4.349241220023384e-23	2.2150092182970496e-20	1.545252362396128e-23
69	pikir okin gitu ya malu dengernya	negative	3.3641490051448087e-13	2.809843973190318e-10	2.226361873828493e-12
70	labil sumpah bicara dr dewasa	positive	3.922555688136282e-09	3.166202926321644e-09	6.759455365853211e-10

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil		
			Positif	Negatif	Netral
71	letak ganteng okin sih dlu nyari nggk dpt	negative	8.790040757728367e-12	1.0319476204307579e-08	1.287515307781564e-11
72	ngomong lo bom anak orang anak toxic njir	negative	7.597196961146109e-18	1.5440730815894592e-15	3.4356110857273916e-18
73	liat bego sumpah ngakak	negative	4.340785827388998e-07	2.361987383035946e-06	5.441361569511834e-07
74	muka si okin aneh kalo ketawa	negative	8.650668870372362e-14	2.1221199238080726e-11	1.439458108078767e-13
75	masya allah gemesin banget queen lalacerdas	positif	2.3425458619346093e-07	1.5039463900027812e-09	1.5450183693378767e-09
76	lala nih asli lucu kocak bangetsehat trus ya lala	positif	4.3009486419281414e-20	3.418224102310261e-24	3.6418900121996374e-22
77	kali podcast om ded jam rokok extream	netral	2.5809122172303468e-15	8.779829559329064e-18	8.654630115653705e-15
78	om deddy emang dewa nya podcast	positif	5.89116957975225e-10	1.3518689884999572e-12	1.8578605981603285e-11
79	lala apresiasi om deddy pertamakalinya nyanyi dipodcast lagu potong bebek angsa	positif	8.410489340275007e-29	7.29129246809027e-34	5.329294997691396e-31
80	ya allahngakak om deddy ngimbangin lala	positif	4.575242646996939e-11	1.7029357413274655e-14	3.265330748281789e-11
81	ga syukur bgt dpt sambung	negative	4.13944994383323e-08	2.1609334972145218e-07	4.828182404180865e-10
82	selamat sukses putri hasil hancur keluarga	positif	2.745145588198163e-11	4.841577307651009e-12	9.596387387191782e-14
83	moga putri dina jodoh duda	negative	8.570289738785155e-09	2.0896939313722854e-08	5.890382533100655e-09
84	poko parah putri ga dewasa	negative	1.714057947757031e-09	3.018446789759968e-08	1.1265758943088685e-10
85	putri dina hrus instrospeksi	negative	7.23464304564833e-08	4.618997549048074e-06	1.2955622784551988e-08
86	suka bgt lala sabar pikir positif	positif	2.837098994338788e-09	2.3579110264534135e-13	6.141687927802739e-13
87	ade lala gemesinnpintar banget ngomongnya	Positif	3.428115895514062e-08	5.5701718148251143e-11	2.575030615563128e-10
88	pintar lala sehat dek lala	Positif	3.1995748358131245e-07	4.6906710019579905e-11	1.8540220432054516e-09
89	bangun tidur siang disuguhin oceh lala jam mood banget	Positif	2.2691693321861126e-21	1.3877405543567298e-25	1.1773351332541932e-23

No	Komentar	Hasil Klasifikasi (Sistem)	Hasil		
			Positif	Negatif	Netral
90	salut ortunya lala gak egoismemaksakan didik anak	positif	1.867190723311646e-12	1.6804593776555831e-15	8.900996996815564e-16
91	anak pintergemesin lucu banget	positif	0.0007586729007203216	4.338183493509355e-05	1.036449822764159e-05
92	hallo lala bikin gemes pinter ngomong	positif	3.0757933761323964e-11	9.431644105813652e-14	1.535421981950685e-13
93	hebat om ded orang teman lala pusing	positif	1.2650440498609048e-12	3.4241335281383353e-17	3.042233616411605e-15
94	om dedy sabar banget ngadapin lala bawel cerdas	positif	2.017881066505956e-15	6.211526721300724e-22	1.042530260496588e-18
95	mama anak cerewet sebelas duabelas lala	netral	1.9338089667002398e-09	1.7003682382097717e-10	4.120048984901005e-09
96	si billar pansos genjot pamor lesti	negative	3.123852647634466e-14	3.772657642325462e-12	2.2391570570114157e-14
97	bilar gatel orang soksokan suami lesty	negative	5.670994037244106e-13	3.034631491045545e-11	1.623388866333276e-13
98	bilar modus doang numpang popular harta	negative	1.441778145062061e-14	3.7726576423254616e-12	3.998494744663242e-15
99	laki jual tampang numpang tenar cerai aja	negative	1.1678335824571705e-17	1.2390095474125176e-14	9.934148433945945e-18
100	ga nang kau lupa ehhhh ngabisin nyawa orang eh	negative	4.462699686632691e-23	2.8795119837861643e-20	3.200879893534837e-23

Lampiran 4

Hasil Perhitungan manual pada data *training*

3.1 Perhitungan Probabilitas Pada Data Training

Tabel 3.1 Contoh Dokumen Latih

Dokumen	Komentar	Kelas
D1	Saya dari malang	Netral
D2	Itu suami tidak tau diri	Negatif
D3	Saya salut dengan lesty	Positif
D4	Untuk fuji, tetap semangat anak baik!	Positif
D5	Fuji ciri manusia tidak bersyukur	Negatif
D6	Alhamdulillah	Netral
D7	Ciri manusia yang tidak bersyukur	Negatif
D8	Usia tidak mencerminkan kedewasaan	Netral
D9	Sehat dan sukses	Positif

Kata	Kata muncul pada dokumen (kelas)		
	Positif	Negatif	Netral
Saya	1	0	1
Suami	0	1	0
Salut	1	0	0
Semangat	1	0	0
Manusia	1	1	0
Syukur	0	2	0
Dewasa	0	0	1
Dengan	1	0	0
Lesty	1	0	0
Total	**6**	**4**	**2**

3.2 Menghitung probabilitas prior

a. Jumlah dokumen training kelas Positif (Npos) = 3

b. Jumlah dokumen training kelas Positif (Npos) = 3

c. Jumlah dokumen training kelas Positif (Npos) = 3

- Total jumlah dokumen training (Ntotal) = Npos + Nneg + Nnet = 9

  a. Probabilitas prior

Positif (P(pos)) = Npos / Ntotal

$= \frac{3}{9}$

$= \frac{1}{3}$

Negatif (P(neg)) = Nneg/Ntotal

$= \frac{3}{9}$

$= \frac{1}{3}$

Netral (P(net) = Nnet/Ntotal

$= \frac{3}{9}$

$= \frac{1}{3}$

3.3 Menghitung Probabilitas Likelihood

Untuk menghitung probabilitas likelihood dapat menggunakan persamaan 3.1 yakni:

$$P(t|c) = \frac{freq(t,c)+1}{freq(c)+V} \quad \text{.....3.1}$$

Dimana,

a. Freq t|c = frekuensi *term* t pada kelas c
b. Freq c = jumlah *term* pada kelas c
c. V = jumlah keseluruhan *term*

Sehingga perhitungan sebagai berikut,

$$P(Saya|Pos) = \frac{freq(Saya,Pos)+1}{freq(Pos)+9} = \frac{1+1}{6+9} = 0{,}1333$$

$$P(Saya|Net) = \frac{freq(Saya,Net)+1}{freq(Net)+9} = \frac{1+1}{2+9} = 0{,}1818$$

$$P(Saya|Neg) = \frac{freq(Saya,Neg)+1}{freq(Neg)+9} = \frac{0+1}{4+9} = 0{,}0769$$

Kata	Kata muncul pada dokumen (kelas)		
	Positif	**Negatif**	**Netral**
Saya	0,1333	0,0769	0,1818
Suami	0,0666	0,1538	0,0909
Salut	0,1333	0,0769	0,0909
Semangat	0,1333	0,0769	0,0909
Manusia	0,1333	0,1538	0,0909
Syukur	0,0666	0,2307	0,0909
Dewasa	0,0666	0,0769	0,1818
Dengan	0,1333	0,0769	0,0909
Lesty	0,1333	0,0769	0,0909

4.2 Menghitung Probabilitas Pada data *testing*

Berikut contoh data testing yang digunakan mengacu pada proses training yang sebelumnya dilakukan.

Teks	Kelas
Saya salut dengan Lesty	?

Menghitung probabilitas likelihood menggunakan persamaan 3.1, Sehingga hasil perhitungan yang diperoleh sebagai berikut.

$$P\ (Salut|Pos) = \frac{freq(Salut,Pos)+1}{freq(Pos)+9} = \frac{1+1}{6+9} = 0{,}1333$$

$$P\ (Saya|Net) = \frac{freq(Saya,Net)+1}{freq(Net)+9} = \frac{1+1}{2+9} = 0{,}1818$$

$$P\ (Saya|Neg) = \frac{freq(Saya,Neg)+1}{freq(Neg)+9} = \frac{0+1}{4+9} = 0{,}0769$$

Kata	Positif	Negatif	Netral
Salut	0,1333	0,0769	0,0909
dengan	0,1333	0,0769	0,0909
Lesty	0,1333	0,0769	0,0909
Saya	0,0666	0,1538	0,0909
TOTAL	0,00005258	0,00002331	0,00002276

Kemudian dilakukan perhitungan nilai probabilitas klsifikasi yang dirumuskan menggunakan persamaan 3.2 yakni:

$V_j(c) = P(\text{Probabilitas prior } c) * P(t1|c) * P(t2|c) * P(t3|c) * P(t4|c)$ ……3.2

Sehingga,

V_j (Pos) = $P(1/3)$ * P(Salut|Pos) * P(dengan|Pos) * P(suami|Pos) * P (saya|Pos)

$= \frac{1}{3}$ * 0,1333 * 0,1333 * 0,0666 * 0,0666

= 0,00005258 (Probabilitas tertinggi)

V_j (Pneg) = $P(1/3) * P$(Salut|Neg)* P(dengan|Neg) * P(suami|Neg) * P (saya|Neg)

$= \frac{1}{3}$ * 0,0769 * 0,0769 * 0,1538 * 0,1538

= 0,00002331

V_j (Pnet) = $P(1/3)$ * P(Salut|Net) * P(dengan|Net) * P(suami|Net) * P (saya|Net)

$= \frac{1}{3}$ * 0,0909 * 0,0909 * 0,0909 * 0,0909

= 0,00002276

Teks	Kelas
Saya salut dengan Lesty	Positif

Hasil data testing yang diperoleh nilai 0,00005258 memiliki nilai probabilitas yang paling tinggi di antara ketiga nilai yang diperoleh. Oleh karena itu, data testing pada kalimat “Saya salut dengan Lesty” diprediksi sebagai kelas positif. Hal ini menunjukkan bahwa sesuai dengan proses latih yang sebelumnya telah dilakukan