JURNAL RESISTOR | ISSN 2598-7542 | E-ISSN 2598-9650
Vol. 5 No 2 – Oktober 2022 | https://s.id/jurnalresistor
DOI : https://doi.org/10.31598
Publishing : Prahasta Publisher | Manage by DRPM INSTIKI

Sentimen Analisis Inisiatif Vaksin Nasional Menggunakan Naïve Bayes dan Laplacian Smoothing Pada Komentar Video Youtube

I Putu Agus Eka Darma Udayana[1*], I Gede Iwan Sudipa[2], Risaldi[3]

[1,2,3]Teknik Informatika, Fakultas, Institut Bisnis dan Teknologi Indonesia
Jl Tukad Pakerisan No 97 Denpasar Selatan, Denpasar, Indonesia

e-mail: agus.ekadarma@gmail.com[1*], iwansudipa@instiki.ac.id[2], ichaldi2342@gmail.com[3]

Received : Mei, 2022	Accepted : Oktober, 2022	Published : Oktober, 2022

Abstract

COVID-19 pandemic that has been declared by who in march 2020 Has been Indonesia biggest health crisis end in the decade. WHO said one of the quickest way to end the pandemic is through immunity through vaccine thu's theory is a national vaccine program initiated by the government in the middle of 2021. YouTube is of de facto public space in Indonesia cyberspace for its netizen for various conversation. from gossiping to discuss in public policy YouTube has been a gold mine for social media data mining enthusiast since 2010. but has been not utilized much by Indonesia Academic. do lack of popularity compared to Twitter which has been a media darling what Indonesian Acdemic ever since This research is focused on sentiment analysis pantydeal YouTube about the national vaccine initiation on a news channel in YouTube. This research is primarily consist of naive bayes classifier a a popular algorithm Indonesian data mining enthusiast which has some limitation such as the problem known as zero probability problem and also the use of non-public data which can be fixed by the use of Laplacian smoothing algorithm which when tested Using 100 of random comments as a data testing has resulted in 71% percent of succes rate and when we do a statistical analysis the precision , recall rate and the F-meassurement score of the classifier all resulted in above 75% score which is satisfactory.

Keywords: *Covid 19, Social Media Mining, Naïve Bayes, Sentiment Analysis*

Abstrak

Pandemi COVID-19 yang dinyatakan sebagai pandemi dunia oleh WHO sejak Maret 2020 setelah menjadi pusat kesehatan terbesar di Indonesia pihak WHO menyatakan bahwa salah satu cara untuk keluar dari pandemi adalah dengan vaksin, maka dilakukanlah inisiasi vaksin Nasional oleh pemerintah di pertengahan tahun 2021 ini. YouTube pada faktanya telah menjadi ruang diskusi publik netizen Indonesia belakangan ini. Sebagai ranah data mining sosial media, YouTube masih sangat kurang eksplorasi dibandingkan dengan Twitter maupun Instagram. Penelitian ini berfokus pada penerapan sentimen analisis pada video-video YouTube tentang inisiasi vaksin nasional pada Kanal Berita di YouTube. Penulis menggunakan Naive Bayes yaitu sebuah algoritma populer yang digunakan untuk text mining satu dekade belakangan ini, penulis mengakui adanya keterbatasan metode ini dalam melakukan deteksi terhadap data baru diluar kamus yang ada. Ketika menemukan data baru maka penulis mencanangkan penggunaan laplacian smoothing, dimana setelah dilakukan training lebih dari 23.000 kata dan dilakukan testing terhadap 100 komentar acak yang didapatkan via YouTube. Improvisasi ini dapat memprediksi sentimen sebuah pernyataan secara akurat sampai dengan 71% dengan presisi tingkat recall dan skor F-meassure di atas 75%.

Kata Kunci: *Covid 19, Social Data Mining, Naïve Bayes, Sentimen Analisis*

1. PENDAHULUAN

Sentimen analisis adalah sebuah *reseach field* yang populer dari cabang ilmu *natural language processing* yang di dalamnya biasanya dibuat model untuk menilai sentimen seorang individu dalam menilai sebuah fenomena. Dalam dunia nyata, aplikasi dari sentimen analisis biasanya digunakan dalam bidang marketing, sosiologi, epidemiologi, bahkan mencakup ranah kebijakan publik dan surveilans. Meledaknya kepopuleran sosial media dalam dua dekade terakhir membuat banyaknya data yang dapat diolah bagi para praktisi teknologi informasi memunculkan demam *big data* dan mengakibatkan munculnya sebuah cabang ilmu baru yang disebut *social media mining* yang pada dasarnya adalah menggunakan sosial media sebagai tempat menggali data, mengolah data dan selanjutnya menulis data tersebut, sehingga didapatkanlah informasi yang akan berguna bagi stakeholder yang memerlukannya. Menggunakan sosial media sebagai sumber data yang mendasari kebijakan publik bukanlah sebuah hal yang baru di Amerika. Sosial media digunakan sebagai patokan atau pedoman untuk mengevaluasi kebijakan publik[1].

Sebagai salah satu faktor yang dipertimbangkan dalam meramalkan pergerakan pasar saham[2] bahkan digunakan untuk memenangkan sebuah agenda politik seperti kasus *Cambridge analytica* yang sempat mencoreng gambar demokrasi di Amerika Serikat beberapa tahun lalu[3]. Sentimen analisis dalam dunia *software engineering* juga merupakan sebuah masalah klasik bagi penggiat matematika Bayesian. Model yang digunakan kebanyakan menggunakan matematika *Bayesian* murni[4] maupun digabung dengan *classifier* lainnya seperti logika *fuzzy*[5] *support Vector machine*[6] bahkan sampai dengan CNN[7], namun bila kita melihat bahwa sentimen analisis sangat condong terhadap pola linguistik masyarakat yang menjadi subjek penelitiannya, alangkah baiknya kita menerapkan prinsip linguistik ke dalam model bayesian tersebut yang tentunya bukan merupakan hal baru. Analisis sentimen berbasis *lexicon* sudah banyak digunakan dan juga mencakup berbagai macam domain pemakaian seperti pada bisnis[8] keuangan ataupun seperti baru-baru ini juga digunakan epidemologi di India sebagai salah satu mitigasi penyebaran virus Covid 19[9].

YouTube.com sebagai salah satu media berbagi video terbesar di dunia juga menjadi salah satu platform sosial media terpopuler di Indonesia. Kontennya yang beragam dan mudah diakses oleh masyarakat menjadikan *market share* YouTube menjadi salah satu yang paling dominan di Indonesia. Persentase yang dicapai sebesar 65% dari seluruh penduduk produktif Indonesia. YouTube rata-rata menjadi konsumsi sekitar 170 juta pasang mata rakyat Indonesia setiap harinya[10]. Prinsip *open platform* yang sangat sejalan dengan prinsip demokrasi rakyat Indonesia menjadikan YouTube sebagai tempat bagi masyarakat Indonesia untuk menyuarakan pendapatnya di berbagai bidang populer seperti olahraga, hobi, kesadaran finansial, mata uang kripto, maupun menyuarakan pendapatnya mengenai kebijakan pemerintah yang terbaru.

Social media mining di Indonesia juga sudah sering dilakukan oleh para akademisi. Media yang paling sering digunakan adalah *Twitter*, namun karakteristik dari *Twitter* yang membatasi karakter yang bisa digunakan oleh penggunanya membuat analisa yang dihasilkan tentu tidak akan maksimal karena akan banyak ditemukan *internet slang, typo* maupun dialek internet yang tentunya bukan merupakan *lexicon* baku bahasa Indonesia. Topik yang dianalisa juga termasuk beragam mulai dari analisa *marketing campaign* sebuah brand terkenal[11] sampai dijadikan barometer untuk prediksi hasil pemilu[12]. Minimnya sosial media analisis yang dilakukan di platform YouTube membuat ranah ini menjadi menarik untuk dieksplorasi, terlebih sistem komentar di YouTube yang lebih *user friendly* membuat data yang didapatkan berpotensi menjadi lebih menarik daripada *social platform* yang lain.

2. TINJAUAN PUSTAKA

2.1. *State of The Art*

Penelitian yang terkait dengan *sentiment* analisis dan *social media mining* bukanlah sesuatu yang asing bagi kalangan *academia computer science* dalam satu dekade terakhir ini. Di Indonesia khususnya bidang ini cukup diminati karena melimpahnya data yang bisa didapat oleh peneliti, namun kebanyakan dari

penelitian yang ada sebagian besar menggunakan data yang diambil dari *platform twitter,* yang dimana memiliki keterbatasan dalam jumlah kata untuk postingannya dan merupakan platform yang memiliki idiom sendiri yang dimana residunya tentu akan menentukan bias pada informasi yang dihasilkan nantinya[13].

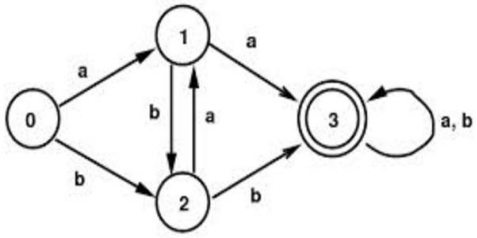
Kami mencoba mengambil *approach* yang berbeda. Dalam hal ini kami menggunakan *platform* YouTube sebagai sumber data yang dimana limitasi karakter dari twitter tidak ditemukan dan juga merupakan salah satu social media paling populer di negeri ini[10]. Kemudian untuk data yang kami pilih dibanding memakai sentimen analisis untuk mendeteksi *sentiment* terhadap sebuah materi promosi[14] atau brand[11] kami menggunakan netizen social media data untuk menganalisa *sentiment* terhadap kebijakan publik yang tentunya jarang dilakukan.

Penelitian yang umum dilakukan pun tidak murni menyasar domain *sentiment,* ada yang menggunkan *opinion mining* baik yang berbahasa Inggris maupun berbahasa Indonesia. Al Halah pada penelitiannya di tahun ini bahkan hanya memakai *feature smiley* untuk menentukan *sentiment* dari sebuah pernyataan yang tentunya sangat kecil jika dilihat dari aspek linguistik dan kebahasaannya[15], meskipun dalam penelitiannya menggunakan data yang berbahasa Indonesia dan tidak menggunakan data publik, begitu juga dengan penelitian oleh Destuardi[16]. Jika kita beralih ke metode yang digunakan, pada umumnya yang beredar adalah berbasis Naïve Bayes classifier murni[11] atau polynomial Naïve Bayes[10], namun kebanyakan tidak adaptif dengan data baru yang dimana penemuan data baru adalah sesuatu yang lumrah ditemukan di kasus-kasus nyata[16]. Pada umumnya banyak menggunakan Naïve Bayes tanpa mengindahkan potensi bias[17] dan zero probability problem[18] yang dimana, nantinya akan sangat berpengaruh pada hasil penelitian. Perbaikan terhadap kasus zero probability tentunya bukan merupakan sesuatu yang asing dilakukan pada ranah text processing, namun alih - alih melakukannya pada text yang statis, kami memutuskan untuk menerapkannya pada ranah social media mining, khususnya pada analisis sentiment kebijakan publik.

2.2. Sentimen Analisis

Sentimen analisis adalah sebuah cabang dari ilmu *natural language processing* yang sering digunakan untuk analisis emosi sebuah populasi tertentu di masyarakat pada umumnya. Dengan perkembangan teknologi informasi yang semakin maju dan kompleks yang berdampak pada pola komunikasi manusia satu sama lain. Masyarakat pada masa kini banyak menggunakan jejaring sosial untuk mengemukakan pendapatnya. Pada bidang ilmu *computer science* sebenarnya sudah banyak peneliti yang mempelajari masalah ini. Beberapa yang menarik contohnya oleh Liu yang dimana mendefinisikan sentimen analisis sebagai tahapan evaluasi emosi seseorang dalam bahasa tekstual[19].

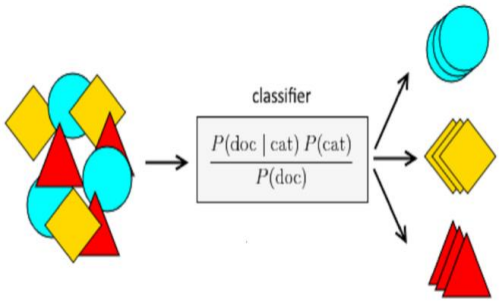
Analisis sentimen kebetulan adalah salah satu cabang ilmu yang paling aktif dalam pengolahan data dengan bidang ilmu linguistik dan kebetulan juga merupakan irisan dalam ilmu *data mining* dan juga sering menjadi *sandbox* dalam bidang ilmu *machine learning* dan pengolahan informasi[19] selain ilmu komputer. Sentimen analisis juga popular di ilmu manajemen dan ilmu sosial lainnya seperti hukum dan antropologi. Bahkan dalam bukunya yang membahas tentang data mining Prabowo dan Thewall menggaris bawahi tentang pentingnya nilai emosi dari komentar, tanggapan atau kritik yang tentu bisa diolah menjadi informasi sesuai tujuan tertentu[19]. Atensi ini dapat digolongkan menjadi dua jenis yaitu positif atau negatif ataupun menjadi beberapa point tertentu sesuai dengan penskalaan yang disukai peneliti dalam kasus-kasus tertentu, analisa sentimen ini juga dapat dipandang sebagai kasus klasifikasi dimana setiap kategori mewakili sesuatu intensi tertentu[19] analisis sentimen ini secara teknis menggunakan berbagai macam algoritma untuk mengklasifikasikan data. Salah satu metode yang paling umum digunakan adalah menggunakan *regular expression* sebagai salah satu *tools* pengolah kata. *Regular expression* yang merupakan operasi text berbasis rules / aturan kebanyakan menggunakan operasi aritmatis didalamnya yang berbasis himpunan.



Gambar 1. Ilustrasi Regular Expression

2.3. Naïve Bayes Classifier

Naive Bayes Classifier adalah sebuah metode klasifikasi dalam dunia machine learning era Bayesian, yang dimana teknik ini lebih condong ke daerah ilmu data mining yang dimana kebanyakan digunakan untuk menambang sebuah informasi dari sebuah kumpulan data dan merupakan salah satu algoritma yang memiliki banyak kegunaan dalam ranah text processing dan ilmu komputer.



Gambar 2. Ilustrasi Naïve Bayes Classifier

Naive Bayes Classifier mengambil teorema dasarnya dari teorema *Bayes* yang dimana merupakan penggabungan antara probabilitas dan statistik yang dipopulerkan oleh seorang pemikir dan matematikawan Inggris yang bernama Thomas Bayes dimana Thomas Bayes ini memprediksi peluang di masa yang akan datang dengan kemungkinan perulangan kejadian yang sebelumnya dialami olehnya. Ciri utama dari teorema ini adalah asumsi yang sangat condong dan bias serta cenderung kuat sehingga terkesan naif terhadap independensi sebuah kondisi / kejadian.

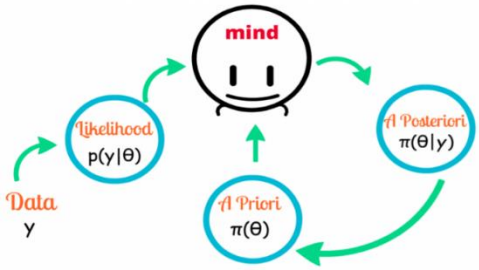
Ada sebuah *textbook* yang mengatakan bahwa *Naïve Bayesian calculation* atau kalkulasi algoritma ini uniknya ada pada asumsi yang diterapkan, diketahui disini bahwa kita dapat mengasumsikan atribut objek yang ada adalah tidak terhubung satu sama lainnya atau singkatnya tidak mempengaruhi antara satu dan lainnya. Dimana nilai probabilitas atau kemungkinan suatu kejadian yang akan terjadi di kemudian hari dapat diramalkan sesuai frekuensi. *Naive Bayes* bekerja sangat baik dalam kasus-kasus tertentu yang dibuktikan dalam berbagai penelitian terdahulu.

Algoritma umum dari sebuah *Naive Bayes classifier* yang sederhana adalah sebagai berikut:

Adalah sebuah himpunan data training yang memiliki panjang N dimensi yang dilambangkan dengan X1,X2,X3….Xn, ini menggambarkan bahwa data dapat direpresentasikan sebagai tuple, dan jika tiap tuple dilambangkan dengan Ai, maka akan ada tuple A1,A2,A3….An.

Katakanlah data tersebut digolongkan menjadi berbagai macam kelas berjumlah m kelas maka kita memiliki kelas C_1, C_2, ..., C_m. Dan jika kita diberi sebuah matrix X, dan jika ada sebuah data X maka *classifier* akan mendeteksi kelas dengan *posterior probability* tertinggi dan hanya akan mendeteksi C_i jika dan hanya jika berbagai kondisi terpenuhi.

Ketika Probabilitas (X) adalah sebuah konstanta yang ditujukan untuk semua kelas, dan kemungkinan itu hanya berlaku jika P (X | C_i) maka kemungkinan sebuah data masuk kelas tertentu atau dilambangkan dengan P (C_i) butuh dimaksimalkan. Dan diingat bahwa ada asumsi bahwa tiap kelas yang ada adalah independen satu sama lainnya.



Gambar 3. Ilustrasi Naïve Bayes Classifier

Namun perlu kita ingat bahwa *Bayesian mathematics* memiliki sifat sifat berikut

Jika A_i dan data adalah sebuah data kategorikal, maka P (X_k | C_i) adalah jumlah *tuple* kelas C_i. Jika A_i adalah sebuah data berdata kontinyu maka data itu diasumsikan memiliki distribusi *Gaussian.*

2.4. Laplace Smoothing Untuk Mitigasi Zero Probability Case

Pada proses pengklasifikasian algoritma *Naive Bayes* sering ditemukan kelemahan apabila ada data yang tidak tersimpan di *bank* datanya atau telah diketahui sebelum di kamus (*dictionary*), namun hal tersebut tentu akan terus ditemukan

semasih *classifier* tersebut diaplikasikan di dunia nyata. Maka banyak pemikir memunculkan metode yang akan digunakan untuk memperkecil bias yang terjadi pada kasus-kasus ini yang kebanyakan ditemukan pada *unknown data* atau data yang tidak diketahui. Salah satu metode yang populer adalah *laplace smoothing* yang dimana singkatnya bekerja dengan menyisipkan satu konstanta baru sehingga hasil klasifikasi tidak akan bias terhadap nilai 0. Pada praktiknya banyak digunakan konstanta 0.5 yang disini kami memilih memakai konstanta alpha =0.3. Meskipun ada berbagai macam teknik *smoothing* lainnya yang banyak diuji oleh Nadia M pada penelitiannya[20]. Kami menemukan bahwa ini adalah metode yang paling sederhana. Perhitungan peluang setiap dokumen menggunakan empat metode *smoothing* ini masih tetap menggunakan kaidah *Naïve Bayes* namun berbeda pada rumus pendugaan parameter P($tk|c$), sesuai rumus 1 berikut ini :

$$P\mu(t|c) = \frac{T_{ct} + \mu \,.P(t|C)}{\Sigma_{t\epsilon V} T_{ct^1} + \mu} \quad (1)$$

Dimana :

T_{ct} : Banyaknya t dalam data latih dari kelas c

μ : Koefisien control

T_{ct^1} : Banyaknya t yang cocok dengan data latih

C : Kombinasi yang memilik status spam/ham

t : Kombinasi yang terbentuk dari data uji

2.5. Pengujian Sistem

2.5.1. Sentiment Polarity

Dalam studi ini kami mendemonstrasikan sebuah grafik yang menyajikan data berupa polaritas sentimen terhadap berita - berita vaksin di platform YouTube yang kami kumpulkan dalam waktu 2 minggu penelitian dari tanggal 1 Juni sampai dengan 14 Juni 2020. Sentimen disajikan dalam bentuk persentase, dan secara garis besar ternyata lebih banyak orang yang memberikan sentimen positif atau optimis terhadap kebijakan vaksin ini. Namun, ternyata masyarakat yang memandang negatif kebijakan vaksin ini juga banyak yaitu mencapai 40%.

2.5.2. Subjektivitas

Dalam aktivitas dari pendapat disini kita bisa lihat adanya sudut pandang yang diambil oleh netizen Indonesia, disini kami menyajikan data dengan menggunakan chat yang berisi penjabaran dari pendapat - pendapat netizen tersebut. Ternyata kebanyakan dari pendapat tersebut mencapai 64% lebih tepatnya ada sentiment - sentiment subjektif yang dilakukan oleh netizen dengan 27% adalah sentiment - sentiment objektif yang memakai data dan sisanya adalah dengan sudut pandang yang tidak jelas. Data tersebut dapat disajikan dalam bentuk grafik dan bagan. Secara garis besar, 60% dari netizen Indonesia menyambut baik kebijakan vaksin sebagai cara pemerintah Indonesia memitigasi pandemi dan mencapai imunitas kelompok, namun sebesar 23% menyampaikan ketakutannya terhadap kebijakan vaksin ini di media social.

Dirangkum oleh sistem, kebanyakan lebih condong kepada percaya diri ke arah kiblat sentimen positif atau kebanyakan takut, ragu-ragu dan khawatir pada konspirasi pada sentimen negatif. Dari studi singkat ini dapat dilihat bahwa kebanyakan masyarakat Indonesia sangatlah mistik pada kebijakan yang diambil pemerintah dan menyambutnya dengan baik dibandingkan dengan data netizen yang tidak menyambutnya dengan baik.

3. METODE PENELITIAN

3.1. Tahapan Penelitian

Terdapat beberapa tahapan yang dilakukan untuk menyelesaikan penelitian ini, dimana setiap tahapan akan sangat mempengaruhi kinerja dari model yang dihasilkan. Gambar 4 merupakan ilustrasi dari alur penelitian yang dilakukan. Seperti yang terlihat pada gambar 4, proses pengembangan sistem dimulai dari mendefinisikan masalah dan dilanjutkan dengan pengumpulan data serta studi literatur.

Dataset yang dikumpulkan dalam penelitian ini adalah data text berupa komentar netizen yang didapatkan dari berita - berita mengenai usaha pemberian vaksin. Komentar tersebut adalah sebuah representasi pendapat masyarakat mengenai kebijakan pemerintah tentang vaksin sebagai cara mendapatkan kekebalan kelompok / *herd imunity* sebagai langkah awal guna penanganan Covid-19. Komentar ini diambil dari kanal stasiun TV Nasional yang terdaftar di dewan pers sehingga konten dari berita dapat dipertanggungjawabkan. Proses *dataset* untuk *training* didapatkan dari public *Dataset of*

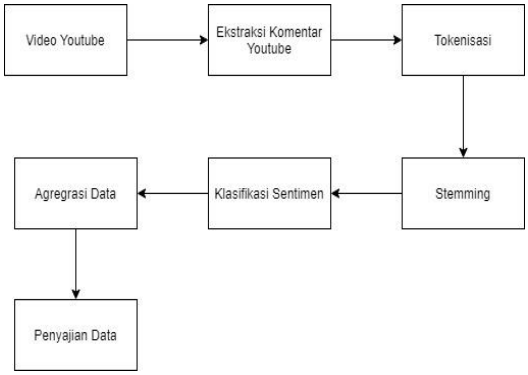
Indonesian Word Sentiment yang disebarkan secara bebas dan gratis pada github. Penggunaan *dataset* yang dikumpulkan secara kolektif diharapkan dapat membuat proses *training* data mendapatkan hasil yang bagus atau sesuai dengan kebutuhan.

Setelah mengumpulkan dataset berupa *lexicon* kata berbahasa Indonesia dan mencari sumber data berupa berita pada kanal YouTube. Sesuai dengan alur penelitian berikutnya, dilaksanakan diskusi untuk menentukan video mana saja yang dijadikan *dataset* dan *metric* apa yang dicari oleh sistem sebagai dasar pembentukan *classifier*nya. Tidak hanya sampai disana saja, pada penelitian ini penulis juga membangun sistem kecil berupa *web scraping* berbasis *python* dan *google script* untuk mempermudah penelitian, dan selanjutnya membangun *word list* tambahan agar *classifier* dapat mencapai akurasi maksimal.

Tahapan akhir dari penelitian ini adalah dilakukan pengujian dan penyajian data. Pengujian dilakukan dengan menguji sentimen dari beberapa data uji yang penulis ambil secara acak untuk menentukan apakah *classifier* cukup akurat untuk dijadikan *sentiment evaluator* yang nantinya akan menjadi dasar dalam penyajian data.

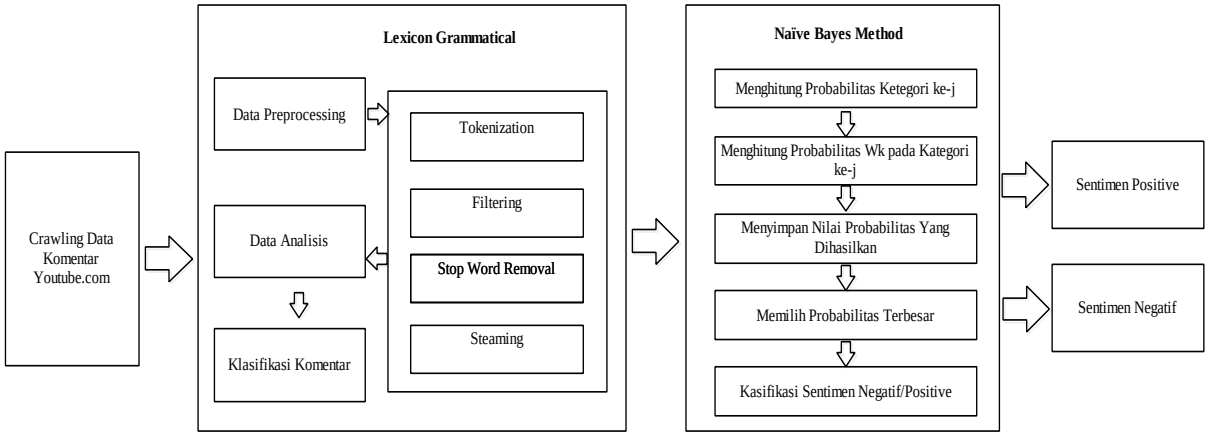
3.2. Desain Sistem

Berdasarkan dari penelitian yang dilakukan sebelumnya, penelitian ini dimulai dengan memperoleh video YouTube kemudian dilakukan ekstraksi komentar pada video YouTue tersebut, kemudian dilakukan tokenisasi, stemming, klasifikasi sentimen, agregasi data dan kemudian baru dilakukan penyajian data. Untuk gambaran umum sistem dapat dilihat pada gambar 4 di bawah ini.

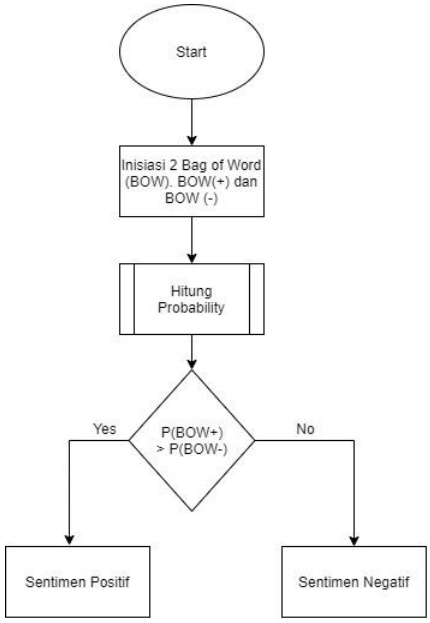


Gambar 4. Gambaran Umum Sistem

Penelitian ini dimulai dengan menyeleksi berita - berita mengenai usaha - usaha vaksin di Indonesia dikanal YouTube berita TV nasional yang tercatat dengan dewan pers. Melalui video tersebut maka *scraper* yang dibangun oleh tim peneliti akan mengekstraksi komentar YouTube dari netizen Indonesia dan menyimpannya dalam sebuah media penyimpanan *cloud* yang kemudian dari *dataset* komentar yang dikumpulkan oleh sistem akan diproses oleh sistem. Pertama kali akan dilakukan proses sosialisasi atau pemisahan struktur kalimat menjadi potongan - potongan kata yang nantinya akan dinormalisasi lagi bentuknya ke dalam proses cermin sehingga didapatkan *lexicon* bahasa Indonesia yang dapat dimengerti oleh *classifier*. Kemudian kata tersebut akan masuk ke dalam sistem yang berbasis *Naive Bayes*. Setelah sistem menilai sentimen dari kata tersebut, maka hasil klasifikasinya akan dikumpulkan kembali ke dalam sebuah modul aplikasi data untuk disajikan menjadi informasi yang berguna bagi segala stakeholder yang berperan dalam penelitian ini.



Gambar 5. Youtube scraping + Naïve bayes



Gambar 6. Proses *Training* Data

3.3. Pengujian Sistem

Untuk melakukan proses pengujian sistem, penulis menggunakan pengujian akurasi sistem. Pengujian akurasi ini digunakan untuk menilai komparasi penggunaan metode *Naïve Bayes* yang telah diperkuat dengan *laplace smoothing*. Adapun banyak data yang digunakan dalam setiap pengujian adalah 107 data komentar YouTube diluar *data training* dan formula yang digunakan untuk menentukan akurasi sistem tertuang pada formula 1.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Training Naïve Bayes

Proses *training* yang dilakukan pada penelitian ini telah menghasilkan model terbaik yang dapat dihasilkan dengan model matematika ini dalam perancangan sistem. Untuk melakukan *training,* pada penelitian ini menggunakan sebanyak 182.000 data kalimat yang didapatkan dari internet.

Tabel 1. Tabel Trainig Sentiment

No	Lexicon	Sentimen
1	Bergizi	Positif
2	Berguna	Positif
3	Berharga	Positif
4	Berhasil	Positif
5	Berjaya	Positif
6	Berjenis	Positif
7	Berjuang	Positif
8	Berkelakuan baik	Positif
9	Anarkis	Negatif
10	Ancaman	Negatif
11	Aneh lagi	Negatif
12	Antek	Negatif
13	Babi	Negatif
14	Awas	Negatif
15	Barbar	Negatif
16	Basi	Negatif
17	Bawahan	Negatif
18	Terbaik	Positif
19	Terkuat	Positif
20	Terpercaya	Positif
21	Terbaik	Positif
22	Ahlinya	Positif
23	Ambisius	Positif
24	Aneh	Positif
25	Berikut	Positif
26	Berimbang	Positif
27	Berhutang	Positif

Seperti yang kita lihat dalam **Tabel 1** bahwa masing-masing reaksinya sudah digolongkan berdasarkan kelasnya masing-masing. Data tersebut didapat dari rutan komen-komen di Kanal Berita sosial media yang terkait dengan inisiasi aksi vaksin Covid-19 nasional yang dimana merupakan mitigasi pemerintah terhadap pandemi Covid-19. Data yang ditampilkan dalam tabel 1 hanyalah sebagian kecil data yang sudah di filter dari data olahan awal yang masih mentah.

4.2. Hasil Perhitungan Laplace Smoothing

Pada proses penghitungan *Naïve Bayes* terdapat sedikit keraguan apabila ada data yang tidak tersimpan di kamus / *bank* datanya, namun mengingat bahwa di dalam dunia nyata merupakan skenario yang tidak bisa dihindari maka muncullah metode - metode mitigasi dalam kondisi ini. Kebanyakan metode - metode ini digunakan untuk memperkecil bias dari classifier terhadap unknown data. Salah satu dari banyaknya metode yang ada adalah laplace smoothing. Dalam penentuan alpha pada laplace smoothing disarankan untuk mendekati nilai 0.5 untuk menghindari bias dan dalam praktiknya kami memakai nilai alpha berupa 0.3.

4.3. Hasil dan Analisa dari Uji *Classifier*

Latar belakang untuk meningkatkan akurasi sistem maka kami menggabungkan *Naive Bayes*

classifier dengan penambahan konstanta dari *laplacian smoothing* di mana kami menguji keakuratan *classifier* menggunakan data komentar netizen pada berita-berita usaha vaksin nasional yang berkaitan dengan mitigasi pandemi Covid-19 yang dilakukan pemerintah Indonesia data ini kami sajikan dalam tabel sebagai berikut :

Tabel 2. Pengujian Metode Naïve Bayes

ID	Sumber	Jumlah Komentar	Word Count
VIDSTIKI001	Tribun News	11646	98384
VIDSTIKI002	Tribun News	1355	10735
VIDSTIKI003	Detik	980	775
VIDSTIKI004	Detik	1050	785
VIDSTIKI005	Liputan 6	946	5833
VIDSTIKI006	Liputan 6	13	71
VIDSTIKI007	Kompas	68	495
VIDSTIKI008	Kompas	68	749
VIDSTIKI009	OKEZONE	4	24
VIDSTIKI010	OKEZONE	4	22
VIDSTIKI011	CNBC Indonesia	29	141
VIDSTIKI012	CNBC Indonesia	92	946
VIDSTIKI013	CNBC Indonesia	29	277
VIDSTIKI014	Vivanews	3013	30551
VIDSTIKI015	Vivanews	464	3781
VIDSTIKI016	Vivanews	705	4135
VIDSTIKI017	CNN	107	1555
VIDSTIKI018	CNN	1417	10749
VIDSTIKI019	CNN	149	686
VIDSTIKI020	CNN	682	3723
VIDSTIKI021	CNN	281	2257
VIDSTIKI022	Metro TV	363	4686
VIDSTIKI023	Metro TV	15	616
VIDSTIKI024	Metro TV	20	107
VIDSTIKI025	TV ONE	16	99
VIDSTIKI026	TV ONE	148	204
VIDSTIKI027	TV ONE	24	490

Hasil akurasi pengujian yang didapatkan dari proses *classifier* kemudian dilanjutkan dengan proses evaluasi pengujian. Proses pengujian menggunakan 107 data uji. Hasil evaluasi pengujian dapat dilihat pada Tabel 3 sebagai berikut :

Tabel 3.Tabel hasil evaluasi pengujian

No	Pengujian	Hasil
1	Accuracy	71 %
2	Recall	84 %
3	Precision	78 %
4	F-score	81 %

4. KESIMPULAN

Ketika kita lihat lebih detail hasil dari penelitian ini, *classifier* dapat bekerja diatas 50% dalam mendeteksi sentimen sebuah pernyataan hal tersebut juga diperkuat jika kita melakukan uji statistika kecil pada hasil uji sentimen analisis di penelitian ini, dimana posisi dari *classifier* juga memiliki presisi dan tingkat *recall* yang lumayan bagus dan mendekati 1 atau sempurna *F-meassure* yang dihasilkan juga dapat dikatakan cukup optimal.

DAFTAR PUSTAKA

[1] N. G. Bacaksizlar, S. Shaikh, and M. Hadzikadic, "Anger in protest networks on twitter," *Multi Conf. Comput. Sci. Inf. Syst. MCCSIS 2019 - Proc. Int. Conf. ICT, Soc. Hum. Beings 2019, Connect. Smart Cities 2019 Web Based Communities Soc. Media 2019*, no. July, pp. 415–419, 2019, doi: 10.33965/wbc2019_201908c054.

[2] A. E. Khedr, S. E. Salama, and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," *Int. J. Intell. Syst. Appl.*, vol. 9, no. 7, pp. 22–30, 2017, doi: 10.5815/ijisa.2017.07.03.

[3] L. Oikonomou and C. Tjortjis, "A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter," in *South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference, SEEDA_CECNSM 2018*, 2018, no. July. doi: 10.23919/SEEDA-CECNSM.2018.8544919.

[4] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes," *Proc. 2019 Int. Conf. Inf. Commun. Technol. Syst. ICTS 2019*, no. July, pp. 49–54, 2019, doi: 10.1109/ICTS.2019.8850982.

[5] R. Mehra, M. K. Bedi, G. Singh, R. Arora, T. Bala, and S. Saxena, "Sentimental analysis using fuzzy and naive bayes," in *Proceedings of the International Conference on Computing Methodologies and Communication, ICCMC 2017*, 2018, vol. 2018-Janua, no. Iccmc, pp. 945–950. doi: 10.1109/ICCMC.2017.8282607.

[6] D. Vijayarani, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," *Int. J. Sci. Eng. Technol. Res.*, vol. 4, no. 4, pp. 816–820, 2015.

[7] I. Kuzborskij, F. M. Carlucci, and B. Caputo, "When Naïve Bayes Nearest Neighbors Meet Convolutional Neural Networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2100–2109, 2016, doi: 10.1109/CVPR.2016.231.

[8] A. Moreno-Ortiz, J. Fernández-Cruz, and C. Pérez-Hernández, "Design and evaluation of SentiEcon: A fine-grained economic/financial sentiment lexicon from a corpus of business news," *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, no. May, pp. 5065–5072, 2020.

[9] M. R, B. A, and S. K, "COVID-19 Outbreak: Tweet based Analysis and Visualization towards the Influence of Coronavirus in the World," *GEDRAG Organ. Rev.*, vol. 33, no. 02, pp. 534–549, 2020, doi: 10.37896/gor33.02/062.

[10] W. A. Social and Hootsuite, "Digital 2021 : INDONESIA," *Simon Kemp*, p. 103, 2021.

[11] I. P. Windasari, F. N. Uzzi, and K. I. Satoto, "Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek," *Proc. - 2017 4th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2017*, vol. 2018-Janua, pp. 266–269, 2017, doi: 10.1109/ICITACEE.2017.8257715.

[12] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *J. Big Data*, vol. 5, no. 1, pp. 1–10, 2018, doi: 10.1186/s40537-018-0164-1.

[13] A. M. Barik, R. Mahendra, and M. Adriani, "Normalization of Indonesian-English Code-Mixed Twitter Data," pp. 417–424, 2019, doi: 10.18653/v1/d19-5554.

[14] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 765–772, 2019, doi: 10.1016/j.procs.2019.11.181.

[15] Z. Al-Halah, A. Aitken, W. Shi, and J. Caballero, "Smile, be happy :) emoji embedding for visual sentiment analysis," *Proc. - 2019 Int. Conf. Comput. Vis. Work. ICCVW 2019*, pp. 4491–4500, 2019, doi: 10.1109/ICCVW.2019.00550.

[16] V. Cherian and M. S. Bindu, “Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique,” *Int. J. Comput. Sci. Trends Technol.*, vol. 5, no. 2, pp. 68–73, 2017.
[17] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, “Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 799–806, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14744.
[18] I. H. Sarker, M. A. Kabir, A. Colman, and J. Han, “An improved Naive Bayes classifier-based noise detection technique for classifying user phone call behavior,” *Commun. Comput. Inf. Sci.*, vol. 845, no. AusDM 2017, pp. 72–85, 2018, doi: 10.1007/978-981-13-0292-3_5.
[19] Y. Peng, Q. Chen, and Z. Lu, “An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining,” no. 1, pp. 205–214, 2020, doi: 10.18653/v1/2020.bionlp-1.22.
[20] E. I. W. M. X. L. Evmsyw, S. J. Sqtyxiv, R. Rxy, and I. H. Y. Wk, “IWFPE 2012 Poster Presentation,” p. 2012, 2012.