Terbit online pada laman web jurnal: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

JURNAL DINDA

Kelompok Keahlian Rekayasa Data
Institut Teknologi Telkom Purwokerto

Vol.1 No.1 (2021) FEB 2021 ISSN Media Elektronik: 2809-8064

Analisis Sentimen Masyarakat Terhadap COVID-19 Pada Media Sosial Twitter

Ardianne Luthfika Fairuz[1], Rima Dias Ramadhani[2] , Nia Annisa Ferani Tanjung[3]
[1,2]Teknik Informatika, Fakultas Informatika, Institut Teknologi Telkom Purwokerto
[3]Rekayasa Perangkat Lunak, Fakultas Informatika, Institut Teknologi Telkom Purwokerto

[1]16102150@ittelkom-pwt.ac.id, [2]rima@ittelkom-pwt.ac.id, [3]nia@ittelkom-pwt.ac.id

Abstract

At the end of 2019, the world was shocked by the emergence of a disease caused by the SARS-CoV-2 virus which is the newest type of virus from the coronavirus. This disease is known as COVID-19. The spread of this disease is quite wide and fast. In a short time this disease began to spread to all corners of the world, including Indonesia. With such a high level of spread and no vaccine for COVID-19 has been found, it is causing chaos in the community. Not a few people are active in social media and write their opinions and thoughts on social media platforms such as Twitter. The occurrence of this pandemic has encouraged the public to write their opinions and thoughts on COVID-19 on Twitter. A sentiment analysis model is needed to classify public tweets into positive and negative. Sentiment analysis is part of Natural Language Processing which creates a system to recognize and extract opinions in text form. In this study, Naive Bayes and K-Nearest Neighbor algorithms were used to build a sentiment analysis model. The accuracy is 85% for Naïve Bayes and 82% for K-Nearest Neighbor at values of k = 6, 8, and 14.

Keywords: *COVID-19, Twitter, Sentiment Analysis, Naïve Bayes, K-Nearest Neighbor*

Abstrak

Akhir tahun 2019 lalu dunia digemparkan oleh munculnya suatu penyakit yang disebabkan oleh virus SARS-CoV-2 yang merupakan jenis virus terbaru dari coronavirus. Penyakit ini dikenal dengan nama COVID-19. Penyebaran penyakit ini terbilang cukup luas dan cepat. Dalam waktu singkat penyakit ini mulai menyebar ke segala penjuru dunia tak terkecuali Indonesia. Dengan tingkat penyebaran yang begitu tinggi dan belum ditemukannya vaksin untuk COVID-19, menyebabkan kekacauan di tengah masyarakat. Hal ini mempengaruhi banyak sektor kehidupan masyarakat. Tak sedikit masyarakat yang aktif bersosial media dan menuliskan pendapat, opini serta pemikirannya di platform media sosial seperti Twitter. Terjadinya pandemi ini mendorong masyarakat untuk menuliskan opini, pemikiran serta pendapatnya terhadap COVID-19 pada media sosial Twitter. Dibutuhkan suatu model sentiment analysis untuk mengklasifikasi tweet masyarakat di Twitter menjadi positif dan negatif. Sentiment analysis merupakan bagian dari Natural Language Processing yang membuat sebuah sistem guna mengenali serta mengekstraksi opini dalam bentuk teks. Pada penelitian ini digunakan algoritma Naive Bayes dan K-Nearest Neighbor untuk digunakan dalam membangun model sentiment analysis terhadap tweet pengguna Twitter terhadap COVID-19. Didapatkan akurasi sebesar 85% untuk algoritma Naïve Bayes dan 82% untuk algoritma K-Nearest Neighbor pada nilai k=6, 8, dan 14.

Kata kunci: *COVID-19, Twitter, Sentiment Analysis, Naïve Bayes, K-Nearest Neighbor*

Diterima Redaksi : 10-12-2020 | Selesai Revisi : 21-12-2020 | Diterbitkan Online : 05-02-2021

1. Pendahuluan

Sejak awal tahun 2020, ramai di masyarakat lingkup dunia tentang adanya sebuah virus mematikan yang dapat menyebar luas dalam waktu yang singkat bernama SARS-CoV-2. SARS-CoV-2 ini merupakan jenis virus terbaru dari coronavirus. Penyakit yang disebabkan oleh virus ini dinamakan Coronavirus disease 2019 (COVID-19). Berdasarkan data yang diambil dari World Health Organization (WHO), terdapat sebanyak 4,347,935 kasus yang terjadi di seluruh dunia, dengan 79,187 kasus baru per tanggal 15 Mei 2020 [1]. Dengan level percepatan penyebaran yang begitu tinggi dan belum ditemukannya vaksin untuk COVID-19, menyebabkan terjadinya kekacauan ditengah masyarakat. Hal ini mempengaruhi banyak sektor kehidupan, dari ekonomi, politik, industri [2], pendidikan, medis, dan juga kehidupan sosial.

Dampak dari terjadinya pandemi ini membuat banyak kekacauan di berbagai belahan dunia, tidak terkecuali Indonesia. Awal bulan Maret 2020, Presiden Joko Widodo menyampaikan langsung temuan kasus COVID-19 pertama di Indonesia [3]. Total kasus terkonfirmasi pada bulan Maret yakni 1.528 kasus. Kemudian pada bulan selanjutnya jumlah kasus terkonfirnasi mengalami peningkatan sebanyak 8.590, yang berarti pada bulan April terdapat 10.118 kasus terkonfirmasi. Selanjutnya pada bulan Mei terdapat sebanyak 26.473 kasus terkonfirmasi. Pada bulan Juni terdapat sebanyak 56.385 kasus terkonfirmasi COVID-19, dan untuk data terbaru yakni bulan Juli jumlah kasus terkonfirmasi meningkat sangat tajam dari jumlah sebelumnya. Pada bulan Juli terjadi peningkatan sebanyak 51.991, total kasus terkonfirmasi yaitu 108.376 [4].

Banyak penutupan sekolah, fasilitas umum, pembatasan moda transportasi, pelayanan masyarakat, dan juga berbagai macam hal lainnya. Di Indonesia sendiri, pemerintah mengeluarkan kebijakan Pembatasan Sosial Berskala Besar (PSBB). PSBB meliputi pembatasan kegiatan pendidikan, pembatasan kegiatan kerja (dengan diberlakukannya Work From Home oleh berbagai instansi), pembatasan kegiatan beribadah, larangan masyarakat untuk berkumpul, pembatasan atau larangan bepergian keluar kota, dan lain sebagainya. Kebijakan PSBB ini diatur dalam Peraturan Pemerintah No. 21 Tahun 2020 tentang Pembatasan Sosial Berskala Besar (PP PSBB) dalam Rangka Percepatan Penanganan Corona Virus Disease (COVID-19) [5].

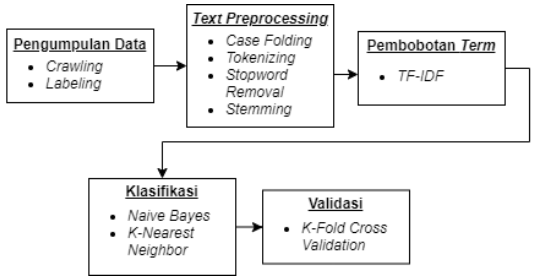
Indonesia merupakan salah satu negara dengan populasi pengguna internet terbesar di dunia. Berdasarkan sumber dari *We are Social* pada tahun 2020, Indonesia mencapai angka 175,4 juta orang pengakses internet. Hal ini menunjukan adanya pertumbuhan populasi pengakses internet sebesar 17 persen dalam satu tahun terakhir, yang berarti ada 25, 3 juta pengakses internet baru sejak tahun 2019 [6]. *Twitter* merupakan salah satu *platform* media sosial yang paling banyak digunakan oleh masyarakat Indonesia. Menurut sumber *We are Social* dan *Hootsuite* tahun 2020, Twitter menempati posisi kelima pada kategori media sosial yang sering digunakan dengan jumlah presentase pengguna sebesar 56% setelah *Youtube, Whatsapp, Facebook* dan juga *Instagram. Twitter* banyak digunakan oleh masyarakat Indonesia dikarenakan penggunaannya yang tergolong cukup mudah. Untuk melakukan pendaftaran, pengguna hanya membutuhkan *e-mail* dan juga nomor ponsel. Salah satu fitur dari *Twitter* yang paling sering digunakan yaitu fitur *tweet.* Dengan menggunakan fitur ini pengguna dapat menuliskan pemikiran, pendapat serta opininya. *Tweet-tweet* yang ditulis oleh para pengguna *platform* ini tentu dapat diolah menjadi satu informasi yang berguna.

Setiap individu memiliki pendapat dan opini yang berbeda-beda. Pendapat ini sangat penting dan juga merupakan salah satu hal yang mempengaruhi perilaku utama manusia. Sentimen analisis (*sentiment analysis*) atau yang juga sering disebut sebagai *opinion mining* merupakan bidang studi yang menganalisa opini masyarakat, sentimen, evaluasi, penilaian, sikap, dan emosi terhadap sebuah produk, pelayanan, organisasi dan perhimpunan, seorang tokoh, dan isu atau masalah serta peristiwa yang terjadi pada masyarakat itu sendiri [7]. Sentimen analisis banyak digunakan di berbagai bidang dalam kehidupan sehari-hari. Seperti misalnya dalam bidang bisnis, sentimen analisis digunakan untuk memprediksi harga saham, menganalisis kebutuhan pasar, hingga penyusunan strategi pemasaran berdasarkan sentimen masyarakat. Sentimen analisis juga digunakan untuk mengetahui opini publik terkait isu, kejadian, dan juga peristiwa yang terjadi di masyarakat.

Dalam penelitian ini akan dibahas mengenai sentimen analisis masyarakat terhadap pandemi COVID-19 pada media sosial *Twitter.* Berdasarkan peristiwa yang saat ini sedang ramai di masyarakat, banyak pengguna media sosial yang memberikan opini, pendapat, serta pemikirannya terhadap COVID-19 pada *platform* media sosial *Twitter*. Hal ini menarik untuk diteliti guna mengetahui opini masyarakat tentang pandemi yang sedang terjadi sekarang ini. Untuk menunjang penelitian tentang hal tersebut, dibutuhkan algoritma untuk mengklasifikasikan komentar masyarakat di media sosial Twitter, baik itu komentar positif maupun komentar negatif. Ada beberapa algoritma yang dapat digunakan untuk

klasifikasi komentar pada media sosial Twitter, diantaranya adalah *Naïve Bayes* [8][9], *K-Nearest Neighbor* [10][11], *Decision Tree* [12][13], dan juga *Support Vectore Machine* [14][15]. Pada penelitian ini akan dipilih *Naïve Bayes* dan *K-Nearest Neighbor* sebagai algoritma yang akan digunakan untuk mengklasifikasi komentar pada media sosial *Twitter*.

2. Metode Penelitian



Gambar 1. Alur Penelitian

Dalam penelitian ini ada beberapa tahap yang harus dilalui terlebih dahulu sebelum akhirnya didapat nilai akurasi dari masing-masing algoritma yang digunakan. Tahap pertama yang dilakukan yaitu pengumpulan data. Pada tahap pengumpulan data, ada dua proses yang dilakukan yaitu proses *crawling* dan *labeling.* Proses *crawling* yaitu proses dimana kita mengambil data dari media sosial *Twitter* untuk nantinya digunakan dalam penelitian. Kemudian proses kedua yang dilakukan yaitu proses labeling. *Labeling* yakni proses melabeli data yang sudah diambil atau didapatkan dari proses pertama yang sudah dilakukan, *crawling*, dengan label positif dan negatif.

Pada tahap kedua penelitian ini, terdapat empat proses yang akan dilakukan. Dimana pada tahap ini akan memproses atau mengolah data yang telah diambil atau didapat dan diberi label. Proses pertama yang dilakukan yaitu *case folding*. Pada proses ini akan dilakukan pengubahan semua huruf yang terdapat dalam dokumen menjadi huruf kecil atau *lowercase* dan menghilangkan karakter selain huruf. Selanjutnya akan dilakukan proses *tokenizing*, dimana pada proses ini akan dilakukan pemotongan atau pemisahan setiap kata yang terdapat dalam dokumen yang kemudian disebut dengan token. Setelah dilakukan proses *tokenizing*, proses selanjutnya yang dilakukan yaitu *stopword removal*. Pada proses ini semua kosakata yang tidak memiliki makna seperti kata ‘dan’, ‘di’, ‘oleh’, akan dihilangkan dari dalam dokumen sehingga menyisakan kata yang bermakna saja di dalam dokumen. Selanjutnya, proses terakhir dalam tahapan *text preprocessing* yaitu proses *stemming*. Semua kata yang terdapat dalam dokumen akan diubah menjadi bentuk kata dasar dengan menghapus atau menghilangkan imbuhan yang terdepat pada kata tersebut.

Setelah melalui seluruh proses pada tahap kedua, akan didapatkan data atau dokumen yang sudah siap diolah serta diproses. Data yang sudah siap diproses itu kemudian dihitung seberapa banyak kemunculan atau frekuensi kemunculan setiap katanya di dalam dokumen. Tahap ini dinamakan tahap pembobotan *term*. Dalam prosesnya, akan digunakan metode pembobotan TF-IDF.

Tahap keempat yaitu tahap klasifikasi. Pada tahap ini data yang sudah melewati proses *labeling*, tahap *preprocessing* serta sudah dilakukan pembobotan dengan metode TF-IDF, akan diproses dengan algoritma yang telah dipilih yaitu algoritma *Naïve Bayes* dan *K-Nearest Neighbor*. Dalam tahapan ini mesin akan diajari untuk mengenal pola data atau dokumen yang ada untuk kemudian dapat mengklasifikasi sebuah data ke dalam dua kelas, yaitu kelas positif dan kelas negatif.

Selanjutnya tahap validasi. Pada tahap ini, proses validasi dilakukan dengan menggunakan *K-Fold Cross Validation*. Proses ini dilakukan untuk pengujian serta penilaian kinerja proses sebuah algoritma. Nantinya dari proses yang dilakukan pada tahapan ini akan didapatkan nilai akurasi dari masing-masing algoritma yang digunakan.

2.1. Landasan Teori

2.1.1 COVID-19

COVID-19 merupakan penyakit menular yang disebabkan oleh Coronavirus jenis terbaru. Coronavirus merupakan suatu kelompok virus yang dapat menyerang baik hewan maupun manusia. Biasanya Coronavirus menyebabkan terjadinya infeksi saluran pernafasan pada manusia, mulai batuk, pilek hingga Middle East Syndrome (MERS), serta Severe Acute Respiratory Syndrome (SARS). Penyakit COVID-19 pertama kali ditemukan pada akhir tahun 2019 tepatnya pada bulan Desember di Wuhan, Tiongkok. Penularan penyakit ini umumnya ditandai dengan demam, batuk kering, dan tubuh yang mudah terasa lelah. Gejala lain yang mungkin dialami oleh orang yang terjangkit atau tertular penyakit ini meliputi rasa nyeri dan sakit pada anggota tubuh tertentu, hidung tersumbat, sakit kepala, konjungtivis, sakit tenggorokan, diare, kehilangan indera penciuman atau perasa, ruam pada kulit.

Penyebaran COVID-19 terbilang cepat. Penyakit ini dapat menyebar dan juga menular lewat cairan dari hidung ataupun mulut yang keluar saat seseorang yang terinfeksi berbicara, batuk atau bersin. Cairan ini juga dapat menempel pada benda dan permukaan lainnya di sekitar kita seperti meja, gagang

pintu, pegangan tangan, bahkan uang. Penularan dapat terjadi apabila seseorang yang baru saja menyentuh benda-benda yang terkena cairan dari orang yang sudah terinfeksi COVID-19 kemudian menyentuh mata, hidung, atau mulut. Hal terburuk yang dapat disebabkan oleh penyakit COVID-19 yaitu kematian. Hingga sekarang belum ditemukan vaksin untuk penyakit ini [16].

2.1.2 *Twitter*

Twitter merupakan sebuah situs media sosial yang mulai dikembangkan pada tahun 2006. Situs ini pertama kali ditemukan oleh Jack Dorsey dan Evan Williams. Twitter merupakan social networking dimana memungkinkan penggunanya dapat saling berkomunikasi satu sama lain melalui fitur yang bernama tweet. Dengan fitur tweet pengguna dapat membuat tulisan atau teks sebanyak 280 karakter [17]. Tidak hanya tweet, saat ini Twitter memiliki banyak fitur lainnya seperti direct message yang memungkinkan pengguna berkomunikasi satu sama lain dengan lebih privat, story yang memungkinkan pengguna dapat merekam momen baik itu foto atau video secara langsung atau realtime, live yang memungkinkan pengguna melakukan siaran langsung (melalui aplikasi pihak ketiga Periscope), voice note memungkinkan pengguna untuk merekam suara, dan masih banyak fitur lainnya.

Twitter menyediakan API (Application Programming Interface). Twitter API diperuntukkan bagi pengembang. Dengan Twitter API memungkinkan pengguna dapat membaca, menulis dan mengambil data dari Twitter. Penggunaan Twitter API ini juga memungkinkan pengembang untuk mengambil informasi atau data pengguna di Twitter atau suatu subjek di lokasi tertentu [18].

2.1.3 *Sentiment Analysis*

Sentiment analysis merupakan merupakan kajian tentang cara menyelesaikan dan memecahkan masalah dari berdasarkan opini masyarakat, sikap serta emosi suatu entitas, dimana entitas tersebut dapat mewakili individu [19]. Sentiment analysis atau yang juga disebut *opinion mining* merupakan proses memahami, mengekstrak serta mengolah data tekstual secara otomatis guna mendapatkan informasi yang terkandung dalam suatu kalimat opini. Dilakukannya analisis sentimen ini bertujuan untuk melihat pendapat atau kecenderungan opini terhadap suatu masalah ataupun objek oleh seseorang, apa memiliki kecenderungan positif, negatif, atau netral [20].

2.1.4 *Naïve Bayes*

Naïve Bayes merupakan sebuah metode klasifikasi yang berdasar pada teorema *Bayes*. Metode ini memprediksi data di masa yang akan mendatang berdasarkan data sebelumnya atau data yang sudah ada. Ciri utama dari metode *Naïve Bayes* ini adalah asumsi yang kuat akan independensi dari masing-masing kondisi. Berikut formula/rumus yang digunakan untuk perhitungan *Naïve Bayes*:

$$P(C|X) = \frac{P(x|c)P(c)}{P(x)} \qquad (1)$$

Keterangan:

- x : Data dengan class yang belum diketahui.
- c : Hipotesis data yang merupakan suatu *specific class.*
- P(C|X) : *Posterior probability*.
- P(c) : *Prior probability*.
- P(x|c) : Probabilitas berdasarkan kondisi hipotesis.

Nilai evidence selalu sama untuk tiap kelas pada suatu sampel. Nilai posterior nantinya akan dibandingkan dengan nilai posterior kelas lain untuk menentukan ke kelas mana suatu sampel akan diklasifikasikan [21].

2.1.5 *K-Nearest Neighbor*

K-nearest Neighbor (KNN) merupakan sebuah algoritma yang digunakan untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang memiliki jarak terdekat dengan objek tersebut. Nilai K terbaik tergantung pada data, secara umum nilai K yang tinggi akan mengurangi noise pada klasifikasi akan tetapi hal ini membuat batasan antar setiap klasifikasi menjadi kabur [22]. Untuk menghitung tingkat kemiripan tetangga antar 2 objek dapat menggunakan persamaan:

$$Sim(d_i.q_i) = \frac{\Sigma_{j=1}^{t}(q_{ij}.d_{ij})}{\sqrt{\Sigma_{j=1}^{t}(q_{ij})^2.\Sigma_{j=1}^{t}(d_{ij})^2}} \qquad (2)$$

Keterangan :

- q_{ij} = Bobot istilah *j* pada dokumen *i* = tf_{ij} . idf_j
- d_{ij} = Bobot istilah *j* pada dokumen *i* = tf_{ij} . idf_j

3. Hasil dan Pembahasan

Pada penelitian ini terdapat lima tahapan yang pada masing-masing tahapan terdapat beberapa proses di dalamnya, diantaranya tahap pengumpulan data, tahap *text preprocessing*, tahap pembobotan *term*, tahap klasifikasi dan tahap validasi. Berikut hasil dari masing-masing tahapan yang sudah dilakukan:

3.1. Pengumpulan Data

Pada tahap pengumpulan data terdapat dua proses yang dilakukan yaitu *crawling data* dan *labeling.* Proses *crawling data* dilakukan menggunakan bahasa pemrograman *Python* dengan *library Tweepy*. Proses ini dilakukan untuk mengambil data *tweet* pengguna media sosial *Twitter* yang

memiliki topik bahasan COVID-19. Pengambilan data dilakukan dengan *keyword* 'COVID' dan *hastag* '#COVID19'. Data *tweet* diambil dalam jangka waktu 4 September 2020 hingga 12 November 2020. Selanjutnya dipilih sebanyak 1000 data dari total semua data *tweet* yang didapatkan untuk kemudian dilakukan pelabelan data positif serta negatif.

Setelah didapatkan *tweet* dari media sosial *Twitter* dengan topik bahasan COVID-19, selanjutnya dilakukan pelabelan. Pelabelan dilakukan dengan pengisian kuisioner oleh pengguna aktif media sosial *Twitter* dengan kriteria usia responden tidak kurang dari 17 tahun, pengguna aktif media sosial *Twitter* setidaknya selama dua tahun, menggunakan *Twitter* paling singkat/sebentar 2 jam/hari, dan mengetahui isu yang sedang ramai diperbincangkan pada media sosial *Twitter*. Berikut contoh hasil data yang sudah dilabeli:

Tabel 1. Data *Labeling*

id	*Tweet*	label
1	b'Semoga Vaksin Covid19 yg efektif dan aman segera ditemukan https://t.co/GO9lRUbBjZ'	1
2	b'Pencegahan penularan virus corona, agar mematuhi protokol kesehatan Covid19 yaitu selalu memakai masker, menjaga jarak dan mencuci tangan dengan sabun.\n\n#BersamaLawanCovid19\n#KampungTangguhSemeru\n#boba\n#bojonegorobahagia https://t.co/hMZHLidBVD'	1
3	"b'Everybody is a covid Ranger, Ayo bersama-sama berjuang melawan COVID-19. \n #pertaminaemployeejurnalism #covidranger #bumnuntuk75tahun #COVID19 #berjuangmelawancovid19 https://t.co/vjmtJwhTcm'"	1
…	…	…
1000	b'Heleh persetan dengan WHO lah ya, bacot aja terus bikin blunder. kami tidak percaya dengan kalian World Hoax Organization !! \xf0\x9f\x91\x8e\nkalau kalian bilang "mungkin takkan pernah ada" artinya covid19 juga sebenarnya tidak pernah ada , semua akal-akalan kalian beuh https://t.co/fKAAuiXKAH'	0

3.2. *Text Preprocessing*

Setelah didapatkan data yang telah memiliki label positif dan negatif, tahap selanjutnya yang dilakukan adalah *text processing.* Pada tahapan ini terdapat beberapa proses yang akan dilakukan terhadap data yang ada, yaitu proses *case folding, tokenizing, stopword removal* dan *stemming.* Proses pertama yang dilakukan pada tahapan *text preprocessing* adalah *case folding*. Pada proses *case folding*, semua huruf kapital (*uppercase*) pada dokumen akan diubah menjadi huruf kecil (*lowercase*). Karakter lain selain huruf juga akan dihilangkan.

Tabel 2. Proses *Case Folding*

Sebelum *Case Folding*	Sesudah *Case Folding*
Semoga Vaksin Covid19 yg efektif dan aman segera ditemukan	semoga vaksin covid yang efektif dan aman segera ditemukan
@Hafidz_AR1924 Covid buatan.....zionis.....TUNGGU AZAB SANG PENCIPTA.....	hafidzar covid buatan zionis tunggu azab sang pencipta
…	…
SEMOGA ALLAH MEMUDAHKAN ORANG ORANG YANG SEDANG BERJUANG MENANGANI COVID	semoga allah memudahkan orang orang yang sedang berjuang menangani covid

Setelah dilakukan proses *case folding*, proses selanjutnya yaitu *tokenizing*. Pada proses ini dilakukan pemotongan/pemisahan setiap kata dalam teks yang disebut sebagai token.

Tabel 3. Proses *Tokenizing*

Sebelum *Tokenizing*	Sesudah *Tokenizing*
semoga vaksin covid yang efektif dan aman segera ditemukan	'semoga', 'vaksin', 'covid', 'yang', 'efektif', 'dan', 'aman', 'segera', 'ditemukan'
hafidzar covid buatan zionis tunggu azab sang pencipta	'hafidzar', 'covid', 'buatan', 'zionis', 'tunggu', 'azab', 'sang', 'pencipta'
…	…
semoga allah memudahkan orang orang yang sedang berjuang menangani covid	'semoga', 'allah', 'memudahkan', 'orang', 'orang', 'yang', 'berjuang', 'menangani', 'covid'

Proses selanjutnya yang dilakukan setelah *tokenizing* yakni *stopword removal*. Pada proses *stopword removal* dilakukan penghapusan kosakata tak bermakna yang terdapat dalam dokumen. Kosakata tak bermakna meliputi "dan", "di", "oleh", "karena", "yang", "ini", dan lain sebagainya.

Tabel 4. Proses *Stopword Removal*

Sebelum *Stopword Removal*	Sesudah *Stopword Removal*
semoga vaksin covid yang efektif dan aman segera ditemukan	semoga vaksin covid efektif aman segera ditemukan
hafidzar covid buatan zionis tunggu azab sang pencipta	hafidzar covid buatan zionis tunggu azab pencipta
…	…
semoga allah memudahkan	semoga allah memudahkan

orang orang yang sedang berjuang menangani covid	orang orang sedang berjuang menangani covid

Proses selanjutnya yang dilakukan dalam tahap *text preprocessing* adalah *stemming.* Pada proses *stemming* dilakukan penguraian atau pemetaan suatu kata menjadi bentuk dasar. Dalam proses ini akan dilakukan penghapusan berbagai macam imbuhan baik berupa sufiks, prefiks, serta konfiks yang terdapat pada kata di dalam dokumen. Seperti misalnya penghapusan imbuhan pada kata "menjalankan" yang kemudian menjadi "jalan", "meraih" menjadi "raih", "menghimbau" menjadi "himbau", dan lain sebagainya.

Tabel 5. Proses *Stemming*

Sebelum *Stemming*	Sesudah *Stemming*
semoga vaksin covid yang efektif dan aman segera ditemukan	semoga vaksin covid efektif aman segera temu
hafidzar covid buatan zionis tunggu azab sang pencipta	hafidzar covid buat zionis tunggu azab cipta
…	…
semoga allah memudahkan orang orang yang sedang berjuang menangani covid	semoga allah mudah orang orang sedang juang tangan covid

3.3. Pembobotan TF-IDF

Setelah melewati semua proses pada tahap *text preprocessing*, didapatkan dokumen yang sudah siap untuk diproses pada tahap selanjutnya, yaitu pembobotan. Pada tahap ini akan dilakukan perhitungan untuk menghitung banyak kemunculan atau frekuensi kemunculan kata yang terdapat dalam dokumen. Dari total 1000 data/dokumen yang ada terdapat 4027 *term* unik.

Berikut merupakan contoh perhitungan TF-IDF secara manual. Perhitungan dilakukan terhadap empat dokumen yang sudah melalui tahap *text preprocessing* dan siap untuk diproses lebih lanjut.

Tabel 6. Dokumen yang Akan Digunakan Dalam Perhitungan

Dokumen 1	semoga pandemi covid cepet hilang aamiin
Dokumen 2	insyaallah semoga kita semua jaga jauh covid aamiin
Dokumen 3	vaksin harap setiap manusia semoga pandemi segera akhir
Dokumen 4	bodoh manusia lebih tular virus covid

Tabel 7. Perhitungan Manual TF-IDF

Token	tf				df	D/df	IDF	W			
	D1	D2	D3	D4				D1	D2	D3	D4
semoga	1	1	1	0	3	1,3	0,1	0,1	0,1	0,1	0
pandemi	1	0	1	0	2	2	0,3	0,3	0	0,3	0
covid	1	1	0	1	3	1,3	0,1	0,1	0,1	0	0,1
cepet	1	0	0	0	1	4	0,6	0,6	0	0	0
hilang	1	0	0	0	1	4	0,6	0,6	0	0	0
aamiin	1	1	0	0	2	2	0,3	0,3	0,3	0	0
insyaallah	0	1	0	0	1	4	0,6	0	0,6	0	0
kita	0	1	0	0	1	4	0,6	0	0,6	0	0
semua	0	1	0	0	1	4	0,6	0	0,6	0	0
jaga	0	1	0	0	1	4	0,6	0	0,6	0	0
jauh	0	1	0	0	1	4	0,6	0	0,6	0	0
vaksin	0	0	1	0	1	4	0,6	0	0	0,6	0
harap	0	0	1	0	1	4	0,6	0	0	0,6	0
manusia	0	0	1	1	2	2	0,3	0	0	0,3	0
segera	0	0	1	0	1	4	0,6	0	0	0,6	0
akhir	0	0	1	0	1	4	0,6	0	0	0,6	0
bodoh	0	0	0	1	1	4	0,6	0	0	0	0,6
tular	0	0	0	1	1	4	0,6	0	0	0	0,6
virus	0	0	0	1	1	4	0,6	0	0	0	0,6

3.4. *Naïve Bayes*

Pada tahapan ini akan dilakukan perhitungan manual algoritma klasifikasi *Naïve Bayes*. Perhitungan akan dilakukan terhadap lima dokumen, tiga dokumen diantaranya merupakan dokumen berlabel positif satu dokumen berlabel negatif dan sisa satunya lagi merupakan dokumen yang belum dilabeli. Dokumen yang digunakan dalam perhitungan ini diambil dari dataset yang digunakan. Perhitungan ini dilakukan dengan tujuan untuk menentukan dokumen yang belum dilabeli tersebut apakah masuk dalam klasifikasi dokumen positif atau negatif. Berikut kelima dokumen yang akan diproses dengan perhitungan manual *Naïve Bayes*:

Tabel 8. Dokumen yang Digunakan

Dokumen ke-	Isi Dokumen	Label
1	semoga pandemi covid cepet hilang aamiin	1
2	insyaallah semoga kita jaga jauh covid aamiin	1
3	vaksin harap manusia semoga pandemi segera akhir	1

4	bodoh manusia tular virus covid	0
5	banyak bodoh rakyat indonesia sejak covid	?

Prior probability untuk *class* positif :

$$P(1) = \frac{3}{4} = 0{,}75 \qquad (3)$$

Keterangan :

P : Probabilitas kemunculan.

3 : Jumlah *class* positif dalam dokumen.

4 : Jumlah seluruh dokumen yang berlabel.

Prior probability untuk *class* negatif :

$$P(0) = \frac{1}{4} = 0{,}25 \qquad (4)$$

Keterangan :

P : Probabilitas kemunculan.

3 : Jumlah *class* negatif dalam dokumen.

4 : Jumlah seluruh dokumen yang berlabel.

Setelah didapatkan angka *prior*, langkah selanjutnya yaitu melakukan perhitungan untuk mencari probabilitas kemunculan *term* dalam dokumen *class*. Perhitungan dilakukan terhadap setiap kata yang terdapat dalam dokumen yang akan diujikan.

$$P(banyak|1) = \frac{0+1}{15+18} = \frac{1}{33} = 0{,}03 \qquad (5)$$

$$P(bodoh|1) = \frac{0+1}{15+18} = \frac{1}{33} = 0{,}03 \qquad (6)$$

$$P(rakyat|1) = \frac{0+1}{15+18} = \frac{1}{33} = 0{,}03 \qquad (7)$$

$$P(indonesia|1) = \frac{0+1}{15+18} = \frac{1}{33} = 0{,}03 \qquad (8)$$

$$P(sejak|1) = \frac{0+1}{15+18} = \frac{1}{33} = 0{,}03 \qquad (9)$$

$$P(covid|1) = \frac{2+1}{15+18} = \frac{3}{33} = 0{,}09 \qquad (10)$$

$$P(banyak|0) = \frac{0+1}{3+18} = \frac{1}{21} = 0{,}04 \qquad (11)$$

$$P(bodoh|0) = \frac{1+1}{3+18} = \frac{2}{21} = 0{,}09 \qquad (12)$$

$$P(rakyat|0) = \frac{0+1}{3+18} = \frac{1}{21} = 0{,}04 \qquad (13)$$

$$P(indonesia|0) = \frac{0+1}{3+18} = \frac{1}{21} = 0{,}04 \qquad (14)$$

$$P(sejak|0) = \frac{0+1}{3+18} = \frac{1}{21} = 0{,}04 \qquad (15)$$

$$P(covid|0) = \frac{1+1}{3+18} = \frac{2}{21} = 0{,}09 \qquad (16)$$

$$P(1|d5) = 0{,}75 \times 0{,}03 \times 0{,}03 \times 0{,}03 \times 0{,}03 \times 0{,}03 \times 0{,}09 = 0{,}000000001640 \qquad (17)$$

$$P(0|d5) = 0{,}25 \times 0{,}04 \times 0{,}09 \times 0{,}04 \times 0{,}04 \times 0{,}04 \times 0{,}09 = 0{,}000000005184 \qquad (18)$$

Dari hasil perhitungan yang telah dilakukan di atas, dapat disimpulkan bahwa dokumen yang diujikan masuk ke dalam kategori dokumen berlabel negatif karena hasil perhitungan *class* negatif menunjukkan nilai yang lebih besar daripada hasil perhitungan *class* positif.

3.5. *K-Nearest Neighbor*

Langkah pertama yang dilakukan adalah menghitung bobot *term* dari setiap dokumen yang terlibat. Perhitungan ini menggunakan metode TF-IDF seperti yang sudah dilakukan sebelumnya, namun pada perhitungan kali ini pembobotan juga dilakukan terhadap data yang akan diklasifikasi.

Tabel 9. Pembobotan Dokumen TF-IDF

Token	tf					df	D/df	IDF	W				
	D1	D2	D3	D4	D5				D1	D2	D3	D4	D5
semoga	1	1	1	0	0	3	1,6	0,2	0,2	0,2	0,2	0	0
pandemi	1	0	1	0	0	2	2,5	0,3	0,3	0	0,3	0	0
covid	1	1	0	1	1	4	1,2	0,09	0,09	0,09	0	0,09	0,09
cepet	1	0	0	0	0	1	5	0,6	0,6	0	0	0	0
hilang	1	0	0	0	0	1	5	0,6	0,6	0	0	0	0
aamiin	1	1	0	0	0	2	2,5	0,3	0,3	0,3	0	0	0
insyaallah	0	1	0	0	0	1	5	0,6	0	0,6	0	0	0
kita	0	1	0	0	0	1	5	0,6	0	0,6	0	0	0
semua	0	1	0	0	0	1	5	0,6	0	0,6	0	0	0
jaga	0	1	0	0	0	1	5	0,6	0	0,6	0	0	0
jauh	0	1	0	0	0	1	5	0,6	0	0,6	0	0	0
vaksin	0	0	1	0	0	1	5	0,6	0	0	0,6	0	0
harap	0	0	1	0	0	1	5	0,6	0	0	0,6	0	0
manusia	0	0	1	1	0	2	2,5	0,3	0	0	0,3	0,3	0
segera	0	0	1	0	0	1	5	0,6	0	0	0,6	0	0
akhir	0	0	1	0	0	1	5	0,6	0	0	0,6	0	0
bodo	0	0	0	1	1	2	2,	0,	0	0	0	0,	0,

h							5	3				3	3
tular	0	0	0	1	0	1	5	0,6	0	0	0	0,6	0
virus	0	0	0	1	0	1	5	0,6	0	0	0	0,6	0
banyak	0	0	0	0	1	1	5	0,6	0	0	0	0	0,6
rakyat	0	0	0	0	1	1	5	0,6	0	0	0	0	0,6
indonesia	0	0	0	0	1	1	5	0,6	0	0	0	0	0,6
semenjak	0	0	0	0	1	1	5	0,6	0	0	0	0	0,6

Setelah melakukan pembobotan *term* yang terdapat dalam setiap dokumen, langkah selanjutnya yaitu menghitung kemiripan dokumen yang akan diklasifikasi terhadap dokumen lainnya. Perhitungan ini terlebih dahulu akan dilakukan dengan menggunakan rumus *cosine similiarity* untuk mendapatkan nilai panjang vektor.

Tabel 10. Perhitungan *Cosine Similiarity*

Term	WD5*Wdi					Panjang Vektor			
	D1	D2	D3	D4	D5	D1	D2	D3	D4
semoga	0	0	0	0	0	0,048	0,048	0,048	0
pandemi	0	0	0	0	0	0,158	0	0,158	0
covid	0,009	0,009	0	0,009	0,009	0,009	0,009	0	0,009
cepet	0	0	0	0	0	0,489	0	0	0
hilang	0	0	0	0	0	0,489	0	0	0
aamiin	0	0	0	0	0	0,158	0,158	0	0
insyaallah	0	0	0	0	0	0	0,489	0	0
kita	0	0	0	0	0	0	0,489	0	0
semua	0	0	0	0	0	0	0,489	0	0
dijaga	0	0	0	0	0	0	0,489	0	0
dijauhkan	0	0	0	0	0	0	0,489	0	0
vaksin	0	0	0	0	0	0	0	0,489	0
harapan	0	0	0	0	0	0	0	0,489	0
manusia	0	0	0	0	0	0	0	0,158	0,158
segera	0	0	0	0	0	0	0	0,489	0
berakhir	0	0	0	0	0	0	0	0,489	0
kebodohan	0	0	0	0,158	0,158	0	0	0	0,158
menular	0	0	0	0	0	0	0	0	0,489
virus	0	0	0	0	0	0	0	0	0,489
banyak	0	0	0	0	0,489	0	0	0	0
rakyat	0	0	0	0	0,489	0	0	0	0
indonesia	0	0	0	0	0,489	0	0	0	0
semenjak	0	0	0	0	0,489	0	0	0	0
JUMLAH	0,009	0,009	0	0,167	2,121	1,351	2,658	2,319	1,303
						1,45	1,16	1,63	1,52

Setelah didapatkan nilai panjang vektor, selanjutnya dilakukan kemiripan dari D5 terhadap setiap dokumen yang ada. Berikut merupakan perhitungannya:

$$Cos(D5, D1) = \frac{0,009}{(1,45 \times 1,16)} = 0,99998441 \quad (19)$$

$$Cos(D5, D2) = \frac{0,009}{(1,45 \times 1,63)} = 0,999992105 \quad (20)$$

$$Cos(D5, D3) = \frac{0}{(1,45 \times 1,52)} = 0 \quad (21)$$

$$Cos(D5, D4) = \frac{0,1677}{(1,45 \times 1,14)} = 0,994855227 \quad (22)$$

Dapat disimpulkan,data D5 masuk ke kategori atau kelas positif. Hasil ini berbanding terbalik dengan hasil perhitungan yang didapatkan dari perhitungan metode *Naïve Bayes*. Hal ini terjadi dikarenakan metode *K-Nearest Neighbor* sangat bergantung dengan data atau dokumen tetangganya, yang mana pada kasus ini jumlah data atau dokumen dengan kelas positif lebih banyak daripada data dengan kelas negatif. Ini membuat metode *K-Nearest Neigbor* mengklasifikasikan data D5 sebagai kelas positif, karena banyaknya jumlah tetangga dengan kelas positif.

3.6. Validasi *K-Fold Cross Validation*

Tahap validasi ini dilakukan untuk mengukur dan atau mengevaluasi kinerja sebuah algoritma. Cara kerja validasi ini yaitu dengan memisahkan data menjadi dua subset yakni data *training* dan data *testing*. Pada tahap ini, data yang ada dibagi menjadi 10 *fold* dan akan dilakukan randomisasi atau pengacakan data agar tidak terjadi pengelompokan data.

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Gambar 2. Cara Kerja *K-Fold Cross Validation*

Pada tabel di atas, diperlihatkan bahwa area yang diarsir merupakan data *testing*, sementara area yang tidak diarsir merupakan data *training*. Setiap dilakukan validasi, data akan digilir seperti pada gambar tabel di atas.