

TUGAS AKHIR - KA 184801

**PERBANDINGAN ANALISIS SENTIMEN MENGENAI BPJS
PADA MEDIA SOSIAL *TWITTER* MENGGUNAKAN *NAÏVE
BAYES CLASSIFIER* DAN *SUPPORT VECTOR MACHINE***

DIVA DURROTUN NADA

NRP 06311840000049

Dosen Pembimbing

Dr. Drs. Soehardjoepri, M.Si

NIP 19620504.198701.1.001

R. Mohamad Atok, S.Si, M.Si, Ph.D

NIP 19710915.199702.1.001

PROGRAM STUDI SARJANA SAINS AKTUARIA

DEPARTEMEN AKTUARIA

FAKULTAS SAINS DAN ANALITIKA DATA

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SURABAYA

2022



TUGAS AKHIR - KA 184801

**PERBANDINGAN ANALISIS SENTIMEN MENGENAI BPJS
PADA MEDIA SOSIAL *TWITTER* MENGGUNAKAN *NAÏVE
BAYES CLASSIFIER* DAN *SUPPORT VECTOR MACHINE***

DIVA DURROTUN NADA

NRP 06311840000049

Dosen Pembimbing

Dr. Drs. Soehardjoepri, M.Si

NIP 19620504.198701.1.001

R. Mohamad Atok, S.Si, M.Si, Ph.D

NIP 19710915.199702.1.001

PROGRAM STUDI SARJANA SAINS AKTUARIA

DEPARTEMEN AKTUARIA

FAKULTAS SAINS DAN ANALITIKA DATA

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SURABAYA

2022



FINAL PROJECT - KA 184801

**COMPARISON OF SENTIMENT ANALYSIS ABOUT BPJS
ON TWITTER SOCIAL MEDIA USING NAÏVE BAYES
CLASSIFIER AND SUPPORT VECTOR MACHINE**

DIVA DURROTUN NADA

NRP 06311840000049

Dosen Pembimbing

Dr. Drs. Soehardjoepri, M.Si

NIP 19620504.198701.1.001

R. Mohamad Atok, S.Si, M.Si, Ph.D

NIP 19710915.199702.1.001

UNDERGRADUATE STUDY PROGRAM ACTUARIAL SCIENCE

DEPARTMENT OF ACTUARIAL

FACULTY OF SCIENCE AND DATA ANALYTICS

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SURABAYA

2022

LEMBAR PENGESAHAN

PERBANDINGAN ANALISIS SENTIMEN MENGENAI BPJS PADA MEDIA SOSIAL *TWITTER* MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER* DAN *SUPPORT VECTOR MACHINE*

TUGAS AKHIR

Diajukan untuk Memenuhi Salah Satu Syarat

Memperoleh Gelar Sarjana Aktuaria pada

Program Studi Sarjana Sains Aktuaria

Departemen Aktuaria

Fakultas Sains dan Analitika Data





Institut Teknologi Sepuluh Nopember

Oleh: **DIVA DURROTUN NADA**

NRP. 063118 4000 0049

Disetujui oleh Tim Penguji Tugas Akhir:

- | | |
|--------------------------------------------|---------------|
| 1. Dr. Drs. Soehardjoepri, M.Si | Pembimbing |
| 2. R. Mohamad Atok, S.Si, M.Si, Ph.D | Ko-Pembimbing |
| 3. Pratnya Paramitha Oktaviana, S.Si, M.Si | Penguji |
| 4. Ulil Azmi, S.Si, M.Si | Penguji |

()
()
()
()

SURABAYA

Juli, 2022

APPROVAL SHEET

COMPARISON OF SENTIMENT ANALYSIS ABOUT BPJS ON TWITTER SOCIAL MEDIA USING NAÏVE BAYES CLASSIFIER AND SUPPORT VECTOR MACHINE





FINAL PROJECT

Submitted to fulfill one of the requirements
for obtaining a degree Bachelor of Actuarial Science at
Undergraduate Study Program of Actuarial Science
Department of Actuarial Science
Faculty of Science and Data Analytics
Institut Teknologi Sepuluh Nopember

By: **DIVA DURROTUN NADA**

NRP. 063118 4000 0049

Approved by Final Project Examiner Team:

- | | | |
|--------------------------------------------|------------|-------------------------------------------------------------------------------------------|
| 1. Dr. Drs. Soehardjoepri, M.Si | Advisor |  |
| 2. R. Mohamad Atok, S.Si, M.Si, Ph.D | Co-Advisor | () |
| 3. Pratnya Paramitha Oktaviana, S.Si, M.Si | Examiner | () |
| 4. Ulil Azmi, S.Si, M.Si | Examiner | () |

SURABAYA

July, 2022

PERNYATAAN ORISINALITAS

Yang bertanda tangan di bawah ini:

Nama mahasiswa / NRP : Diva Durrotun Nada / 0631184000049
Departemen : Aktuaria
Dosen Pembimbing / NIP : Dr. Drs. Soehardjoepri, M.Si /19620504.198701.1.001
R. Mohamad Atok, S.Si, M.Si, Ph.D
/19710915.199702.1.001

dengan ini menyatakan bahwa Tugas Akhir dengan judul “Perbandingan Analisis Sentimen Mengenai BPJS Pada Media Sosial *Twitter* Menggunakan Metode *Naïve Bayes Classifier* dan *Support Vector Machine*” adalah hasil karya sendiri, bersifat orisinal, dan ditulis dengan mengikuti kaidah penulisan ilmiah.

Bilamana di kemudian hari ditemukan ketidaksesuaian dengan pernyataan ini, maka saya bersedia menerima sanksi sesuai dengan ketentuan yang berlaku di institut Teknologi Sepuluh Nopember.

Surabaya, Juli 2022

Mengetahui

Dosen Pembimbing



(Dr. Drs. Soehardjoepri, M.Si)

NIP. 19620504.198701.1.001

Dosen Ko-Pembimbing



(R. Mohamad Atok, S.Si, M.Si, Ph.D)

NIP. 19710915.199702.1.001

Mahasiswa,



(Diva Durrotun Nada)

NRP. 0631184000049

STATEMENT OF ORIGINALITY

The undersigned below:

Name of student / NRP : Diva Durrotun Nada / 0631184000049
Department : Actuarial Science
Advisor / NIP : Dr. Drs. Soehardjoepri, M.Si /19620504.198701.1.001
R. Mohamad Atok, S.Si, M.Si, Ph.D
/19710915.199702.1.001

Hereby declare that the Final Project with the title of "Comparison of Sentiment Analysis About BPJS on Twitter Social Media Using Naïve Bayes Classifier and Support Vector Machine" is the result of my own work, is original, and is written by following the rules of scientific writing.

If in the future there is a discrepancy with statement then I am willing to accept sanctions in accordance with the provisions that apply at Institut Teknologi Sepuluh Nopember.

Surabaya, July 2022

Acknowledge

Advisor



(Dr. Drs. Soehardjoepri, M.Si)

NIP. 19620504.198701.1.001

Co-Advisor



(R. Mohamad Atok, S.Si, M.Si, Ph.D)

NIP. 19710915.199702.1.001

Student



(Diva Durrotun Nada)

NRP. 0631184000049

**PERBANDINGAN ANALISIS SENTIMEN MENGENAI BPJS PADA MEDIA
SOSIAL TWITTER MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER DAN
SUPPORT VECTOR MACHINE**

Nama Mahasiswa : Diva Durrotun Nada
NRP : 063118 4000 049
Departemen : Aktuaria
Dosen Pembimbing : Dr. Drs. Soehardjoepri, M.Si
R. Mohamad Atok, S.Si, M.Si, Ph.D

Abstrak

Seiring dengan perkembangan teknologi, masyarakat saat ini dapat mengungkapkan perasaan, pendapat, atau pandangannya kepada publik melalui jejaring sosial. Salah satu media sosial terpopuler saat ini adalah *Twitter* yang diluncurkan oleh Jack Dorsey pada tanggal 15 Juli 2006. Media sosial ini merupakan salah satu media sosial utama yang digunakan masyarakat Indonesia untuk memberikan opini kepada pengguna internet. Karena jumlah pengguna *Twitter* yang cukup besar, hal ini sering digunakan oleh pemerintah, pelaku bisnis, maupun masyarakat untuk melihat pendapat pengguna tentang suatu produk atau layanan. Karena sebagian besar masyarakat Indonesia menggunakan BPJS, maka hal ini menyebabkan banyak pengguna media sosial seperti *Twitter* mengunggah ulasan mereka terkait kinerja BPJS. Hal ini dikarenakan hasil penelitian diperoleh langsung dari opini publik atas apa yang mereka alami, maka hasil tersebut dapat digunakan sebagai pengoptimalisasian program kerja, dan peningkatan kualitas pelayanan bagi perusahaan tersebut. Penelitian ini menggunakan dua metode untuk membandingkan tingkat akurasi antara metode *Naïve Bayes Classifier* dan *Support Vector Machine* menggunakan data *Twitter* berupa *tweet* umum mengenai kinerja BPJS dengan kata kunci “BPJS”, “Badan Penyelenggara Jaminan Sosial”, “Klaim” sejak Januari 2019 sampai Desember 2021. Hasil penelitian menunjukkan bahwa metode *Support Vector Machine* Kernel RBF dengan parameter $C = 1000$ dan $\gamma = 100$ memiliki performa ketepatan klasifikasi yang paling baik dibanding *Naïve Bayes Classifier* dan *Support Vector Machine* Kernel Linear. Dengan hasil rata-rata ketepatan klasifikasi SVM Kernel RBF, SVM Kernel Linear, dan *Naïve Bayes Classifier* masing-masing sebesar 97,1%, 92,5%, dan 86,7%.

Kata kunci: Analisis Sentimen, BPJS, *Naïve Bayes Classifier*, *Support Vector Machine*, *Twitter*.

COMPARISON OF SENTIMENT ANALYSIS ABOUT BPJS ON TWITTER SOCIAL MEDIA USING NAÏVE BAYES CLASSIFIER AND SUPPORT VECTOR MACHINE

Student Name : Diva Durrotun Nada
NRP : 063118 4000 049
Departement : Aktuaria
Advisor : Dr. Drs. Soehardjoepri, M.Si
R. Mohamad Atok, S.Si, M.Si, Ph.D

Abstract

Along with the development of technology, today's society can express their feelings, opinions, or views to the public through social networks. One of the most popular social media today is Twitter, which was launched by Jack Dorsey on July 15, 2006. This social media is one of the main social media used by Indonesian people to give opinions to internet users. Because the number of Twitter users is quite large, it is often used by governments, businesses, and the public to see what users think about a product or service. Because most Indonesians use BPJS, this has caused many social media users such as Twitter to upload their reviews regarding the performance of BPJS. This is because the research results are obtained directly from public opinion on what they experience, so these results can be used as work optimization programs, and improving service quality for the company. This study uses two methods to compare the level of accuracy between the Naïve Bayes Classifier and Support Vector Machine methods using Twitter data in the form of general tweets about BPJS performance in submitting claims with the keyword "BPJS", "Badan Penyelenggara Jaminan Sosial, "Claim" from January 2019 to December 2021. The results show that the method The Support Vector Kernel RBF Machine with parameters $C = 1000$ and $\gamma = 100$ has the best classification accuracy performance compared to the Naïve Bayes Classifier and Support Vector Machine Kernel Linear. With the results of the average classification accuracy of SVM Kernel RBF, SVM Kernel Linear, and Naïve Bayes Classifier, respectively, which are 97.1%, 92.5%, and 86.7%.

Keywords: *Sentiment Analysis , BPJS, Naïve Bayes Classifier, Support Vector Machine, Twitter*

KATA PENGANTAR

Puji Syukur penulis panjatkan atas rahmat dan hidayah yang diberikan Allah SWT sehingga penulis dapat menyelesaikan laporan Tugas Akhir yang berjudul “Perbandingan Analisis Sentimen Mengenai BPJS Pada Sosial Media Twitter Menggunakan Metode *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM)”.

Tugas Akhir ini dapat terselesaikan tidak lepas dari bantuan dan dukungan dari berbagai pihak. Oleh karena itu, atas dukungan, saran, motivasi, semangat, serta bantuan yang telah diberikan kepada penulis, penulis menyampaikan terima kasih kepada.

1. Allah SWT yang telah memberikan petunjuk, kekuatan, kesabaran serta keteguhan hati kepada penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan sangat baik tanpa melalaikan perintah-Nya.
2. Bapak Dr. Drs. Soehardjoepri, M.Si selaku kepala Departemen Aktuaria ITS yang telah memberikan fasilitas untuk menunjang kelancaran penyelesaian Tugas Akhir
3. Ibu Pratnya Paramitha Oktaviana S.Si, M.Si, M.Sc selaku dosen wali yang telah memberikan pengarahan dan motivasi selama perkuliahan penulis.
4. Bapak Dr. Drs. Soehardjoepri, M.Si dan Bapak R. Mohamad Atok, S.Si, M.Si, Ph.D selaku dosen pembimbing Tugas Akhir, atas segala bimbingan, saran dan arahan yang diberikan sehingga Tugas Akhir ini dapat diselesaikan dengan baik
5. Ibu Pratnya Paramitha Oktaviana S.Si, M.Si, M.Sc dan Ibu Ulil Azmi S.Si, M.Si, M.Sc selaku dosen penguji yang senantiasa memberi arahan, bimbingan, dan waktu dalam menyelesaikan Tugas Akhir
6. Bapak dan Ibu Dosen Departemen Sains Aktuaria Fakultas Sains dan Analitika Data yang telah memberikan ilmu dan seluruh Tenaga Kependidikan Departemen Sains Aktuaria yang telah membantu administrasi
7. Orang tua dan keluarga penulis yang selalu memberikan doa, kasih sayang, dan segala dukungan baik moral maupun materi
8. Sahabat penulis, Audrey Fahdina, Dhea Kartika, Meidytania, dan Melinda Andriani yang selalu mendukung, memotivasi, dan memberikan saran
9. Teman-teman penulis yang selalu mengingatkan, menyemangati, dan menemani
10. *Last but not least, I wanna thank me, I wanna thank me for believing in me, I wanna thank me for doing all this hard work, I wanna thank me for having no days off, I wanna thank me for never quitting, I wanna thank me for always being a giver and tryna give more than I receive, I wanna thank me for tryna do more right than wrong, I wanna thank me for just being me at all times.*

Penulis menyadari berbagai keterbatasan yang dimiliki sehingga laporan Tugas Akhir ini masih jauh dari kata sempurna. Untuk itu, penulis terbuka menerima kritik dan masukan yang bersifat membangun untuk perbaikan penyusunan laporan Tugas Akhir agar menjadi lebih baik. Semoga dukungan, bimbingan, dan kebaikan yang telah diberikan kepada penulis mendapatkan rida Allah SWT.

Surabaya, Juli 2022
Hormat kami,

Penulis

DAFTAR ISI

LEMBAR PENGESAHAN	iii
APPROVAL SHEET	iv
PERNYATAAN ORISINALITAS	v
STATEMENT OF ORIGINALITY	vi
Abstrak	vii
Abstract	viii
KATA PENGANTAR	ix
DAFTAR ISI	xi
DAFTAR GAMBAR	xiii
DAFTAR TABEL	xv
DAFTAR LAMPIRAN	xvii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
BAB II TINJAUAN PUSTAKA	5
2.1 Penelitian Sebelumnya	5
2.2 Twitter	6
2.3 Badan Penyelenggara Jaminan Sosial	7
2.4 Analisis Sentimen	7
2.5 Data Mining	9
2.6 Text Mining	9
2.7 Text Preprocessing	10
2.7.1 Cleaning	10
2.7.2 Case Folding	10
2.7.3 <i>Stopword Removal</i>	11
2.7.4 <i>Stemming</i>	11
2.7.5 <i>Tokenization</i>	11
2.8 Term Weighting	11
2.9 Oversampling	13
2.10 K-fold Cross Validation	13
2.11 Teorema Bayes	13
2.12 Naïve Bayes Classifier	14
2.13 Support Vector Machine	16
2.14 Klasifikasi	18
2.15 Word Cloud	19
BAB III METODOLOGI PENELITIAN	21
3.1 Sumber Data	21
3.2 Variabel Penelitian	21
3.3 Metode Analisis	22

3.3.1	Studi Literatur	22
3.3.2	Pengumpulan Data	22
3.3.3	Pre-processing Data	22
3.3.4	Pelabelan Kelas Sentimen	22
3.3.5	<i>Oversampling</i> Data.....	22
3.3.6	<i>Term Frequency-Inverse Document Frequency</i>	22
3.3.7	Pembagian Data <i>Testing</i> dan <i>Training</i>	23
3.3.8	Klasifikasi <i>Naïve Bayes Classifier</i>	23
3.3.9	Klasifikasi <i>Support Vector Machine</i>	23
3.3.10	Perbandingan Klasifikasi <i>Naïve Bayes Classifier</i> dan <i>Support Vector Machine</i> ..	23
3.3.11	Kesimpulan dan Saran.....	23
BAB IV HASIL DAN PEMBAHASAN		27
4.1	Karakteristik Data	27
4.2	Preprocessing Text.....	28
4.2.1	Cleaning	28
4.2.2	Case Folding.....	29
4.2.3	Stopwords Removal	29
4.2.4	Stemming	30
4.2.5	Tokenization.....	30
4.3	Pelabelan.....	30
4.4	<i>Oversampling</i>	31
4.5	<i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	32
4.6	Analisis Klasifikasi Menggunakan <i>Naïve Bayes Classifier</i>	33
4.7	Analisis Klasifikasi Menggunakan <i>Support Vector Machine</i>	35
4.7.1	Analisis Klasifikasi Menggunakan <i>Support Vector Machine</i> Kernel Linear.....	35
4.7.2	Analisis Klasifikasi Menggunakan <i>Support Vector Machine</i> Kernel RBF.....	36
4.8	Perbandingan <i>Naïve Bayes Classifier</i> dan <i>Support Vector Machine</i>	37
4.9	Visualisasi Word Cloud.....	38
BAB V KESIMPULAN DAN SARAN		39
5.1	Kesimpulan	39
5.2	Saran	39
DAFTAR PUSTAKA		41
LAMPIRAN.....		43
BIODATA PENULIS.....		61

DAFTAR GAMBAR

Gambar 2.1 Logo Twitter.....	6
Gambar 2.2 Algoritma Cross Validation	13
Gambar 2.3 Ilustrasi linearly separable case.....	16
Gambar 2.4 Ilustrasi Non-Linear Case	18
Gambar 2.5 Visualisasi Word Cloud	20
Gambar 3.1 Diagram Alir Metode Analisis	24
Gambar 4.1 Grafik Ulasan BPJS Kesehatan dan BPJS Ketenagakerjaan.....	27
Gambar 4.2 Pie chart Sentimen Positif dan Negatif	31
Gambar 4.3 Pie chart Sentimen Positif dan Negatif Setelah Oversampling	32
Gambar 4.4 Visualisasi Word cloud	38

(Halaman ini sengaja dikosongkan)

DAFTAR TABEL

Tabel 2.1 Fitur Twitter dan Kegunaannya	6
Tabel 2.2 Proses Cleaning Pada Tweet.....	10
Tabel 2.3 Proses Case Folding Pada Tweet.....	10
Tabel 2.4 Proses Stopword Removal Pada Tweet	11
Tabel 2.5 Proses Stemming Pada Tweet.....	11
Tabel 2.6 Proses Tokenization Pada Tweet	11
Tabel 2.7 Ilustrasi tweet.....	12
Tabel 2.8 Ilustrasi TF-IDF	12
Tabel 2.9 Persamaan Fungsi Kernel	18
Tabel 2.10 Confusion Matrix.....	19
Tabel 3.1 Sumber Data	21
Tabel 3.2 Struktur Data Penelitian.....	21
Tabel 4.1 Data Tweet Sebelum Preprocessing Text	28
Tabel 4.2 Data Tweet Sebelum dan Sesudah Cleaning	28
Tabel 4.3 Data Tweet Sebelum dan Sesudah Case Folding.....	29
Tabel 4.4 Data Tweet Sebelum dan Sesudah Stopwords.....	29
Tabel 4.5 Data Tweet Sebelum dan Sesudah Stemming	30
Tabel 4.6 Data Tweet Sebelum dan Sesudah Tokenization.....	30
Tabel 4.7 Data tweet Setelah Pelabelan.....	31
Tabel 4.8 Tabel TF.....	32
Tabel 4.9 Tabel TF-IDF.....	33
Tabel 4.10 Hasil Probabilitas Naïve Bayes Classifier	33
Tabel 4.11 Pengukuran Ketepatan Klasifikasi Naïve Bayes Classifier	34
Tabel 4.12 Confusion Matrix Naïve Bayes Classifier	34
Tabel 4.13 Hasil Ketepatan Klasifikasi SVM Kernel Linear $C = 100$	35
Tabel 4.14 Confusion Matrix SVM Kernel Linear Parameter $C = 100$	36
Tabel 4.15 Hasil Perhitungan Ketepatan Klasifikasi SVM Kernel RBF $C = 1000, \gamma = 100$...	36
Tabel 4.16 Confusion Matrix SVM Kernel RBF $C = 1000 \gamma = 100$	37
Tabel 4.17 Perbandingan Ketepatan Klasifikasi	37

(Halaman ini sengaja dikosongkan)

DAFTAR LAMPIRAN

Lampiran 1. Data Tweet Setelah Pre-processing Text dan Pelabelan	43
Lampiran 2 Hasil Ketepatan Klasifikasi SVM Kernel Linear	44
Lampiran 3 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 0.001$	45
Lampiran 4 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 0.01$	46
Lampiran 5 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 0.1$	47
Lampiran 6 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 1$	48
Lampiran 7 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 10$	49
Lampiran 8 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 100$	50
Lampiran 9 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 1000$	51
Lampiran 10 Syntax scrapping data menggunakan Python	52
Lampiran 11 Syntax preprocessing data menggunakan Python	53
Lampiran 12 Syntax Pelabelan, <i>Oversampling</i> data dan TF-IDF Menggunakan Python	56
Lampiran 13 Naïve Bayes Classifier menggunakan Python	57
Lampiran 14 <i>Support Vector Machine</i> menggunakan <i>Python</i>	58
Lampiran 15 Syntax Word cloud	59

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam kehidupan sosial, mengungkapkan perasaan atau pendapat tentang sesuatu menjadi kegiatan rutin. Seiring dengan perkembangan teknologi, masyarakat saat ini dapat mengungkapkan perasaan, pendapat atau pandangannya kepada publik melalui jejaring sosial. Salah satu media sosial terpopuler saat ini ialah *Twitter* yang didirikan oleh Jack Dorsey dan resmi diluncurkan pada tanggal 15 Juli 2006, merupakan salah satu media sosial utama yang digunakan masyarakat Indonesia untuk memberikan opini kepada pengguna internet. Menurut data dari otoritas Informasi dan Komunikasi Publik Indonesia, Indonesia menduduki peringkat ke-5 dunia dalam hal pengguna *Twitter* setelah Amerika Serikat, Brasil, Jepang, dan Inggris (Kominfo, 2022).

Twitter adalah jejaring sosial online yang memungkinkan pengguna untuk mengirim pesan singkat 280 karakter yang disebut *tweets*. Menurut angka industri media sosial baru-baru ini, *Twitter* saat ini menempati peringkat sebagai salah satu sosial media terkemuka di seluruh dunia berdasarkan pengguna aktif. Pada kuartal keempat tahun 2020, *Twitter* memiliki 192 juta pengguna aktif harian dan pada kuartal kedua tahun 2021, *Twitter* memiliki 206 juta pengguna aktif harian (Statista Research Department, 2022). Mayoritas pengguna *Twitter* di Indonesia adalah orang-orang yang tidak memiliki blog atau tidak pernah meng-*upload* video ke *Youtube* tetapi sering meng-*update* statusnya di *Twitter*. *Twitter* sering digunakan untuk mengungkapkan perasaan tentang sesuatu, baik itu memuji atau mengkritik. Karena jumlah pengguna *Twitter* yang cukup besar, hal ini sering digunakan oleh pemerintah, pebisnis, maupun masyarakat untuk melihat pendapat pengguna tentang program pemerintah saat ini atau pendapat pengguna tentang suatu produk atau layanan. Informasi yang didapat di jejaring sosial bisa berupa opini positif maupun negatif. Proses untuk meninjau pendapat masyarakat ini sering disebut analisis sentimen (Nurulbaiti, dkk., 2018).

Di tahun 2020, seluruh dunia dilanda sebuah virus meresahkan yang disebut Covid-19 dan pada Maret 2020, pertama kali tercatat kasus positif Covid-19 di Indonesia. Saat itu, tercatat dua orang warga Depok positif terpapar virus tersebut. Dan kasus tersebut terus meningkat sejak diumumkannya dua kasus pertama. Indonesia sendiri sudah menghadapi gelombang pertama infeksi Covid-19 sejak Maret 2020, gelombang kedua Covid-19 terjadi usai adanya varian Delta Covid-19 pada pertengahan tahun 2021. Pada gelombang kedua ini, kasus yang dialami melonjak tinggi sehingga menyebabkan tingkat hunian dan persediaan oksigen meningkat tajam. Dampak yang diakibatkan dari virus ini adalah diberhentikannya kegiatan pemerintahan, ekonomi, maupun sekolah untuk sementara waktu sampai keadaan mulai kondusif dan menyebabkan banyak pekerja yang kehilangan pekerjaannya. Pemerintah dengan tegas memberi kepercayaan kepada BPJS untuk menjalankan tugas khusus dalam penanganan Covid-19 dengan melakukan proses verifikasi klaim kasus Covid-19 dan untuk pembiayaan dilakukan oleh Kementerian Kesehatan.

Karena hal ini, defisit yang telah dialami BPJS sejak 2017 semakin meningkat karena terjadinya lonjakan klaim. Sehingga untuk menekan defisit tersebut, BPJS membuat peraturan mengenai kenaikan iuran. Hal ini menyebabkan kemarahan dari para peserta karena dianggap BPJS tidak berperikemanusiaan. Dilihat dari kaca mata aktuarial, hal ini merupakan dampak dari perhitungan iuran yang terkesan tidak seimbang sehingga tidak dapat menutupi realisasi biaya kesehatan serta beban biaya dan menyebabkan kenaikan defisit yang besar dibanding tahun sebelumnya. Aktuarial merupakan bidang ilmu yang mempelajari segala perhitungan risiko,

maupun premi. Sehingga dalam hal ini BPJS membutuhkan aktuaris yang handal untuk mengatasi defisit tersebut.

BPJS atau Badan Penyelenggara Jaminan Sosial sendiri adalah salah satu dari lima rencana Sistem Jaminan Sosial Nasional (SJSN), yaitu jaminan Kesehatan, perlindungan kecelakaan kerja, perlindungan hari tua, perlindungan pensiun, dan perlindungan kematian. Dengan landasan hukum UU Nomor 24 Tahun 2011 tentang Badan Penyelenggara Jaminan Sosial, dan UU Nomor 40 Tahun 2004 tentang Sistem Jaminan Sosial. BPJS dibagi menjadi dua, yaitu BPJS Kesehatan dan BPJS Ketenagakerjaan (BPJS Kesehatan, 2021). Karena sebagian besar masyarakat Indonesia menggunakan BPJS, maka hal ini menyebabkan banyak pengguna media sosial seperti *Twitter* mengunggah ulasan mereka mengenai BPJS. Ulasan tersebut dapat berupa ulasan positif maupun negatif. Ulasan tersebut kemudian akan dianalisis menggunakan analisis sentimen untuk mengetahui kinerja pelayanan dari BPJS.

Analisis sentimen atau bisa juga disebut *opinion mining* adalah bidang ilmu komputer yang menganalisis pendapat, penilaian, sikap, emosi, sentimen, dan evaluasi terhadap suatu produk, layanan, organisasi, individu, tokoh publik, acara, maupun kegiatan tertentu (Liu, 2012). Manfaat analisis sentimen telah menyebabkan perkembangan pesat penelitian dan pengembangan aplikasi yang terkait dengan analisis sentimen. Hal ini dikarenakan hasil penelitian diperoleh langsung dari opini publik atas apa yang mereka alami dan hasil tersebut dapat digunakan sebagai optimalisasi produk, program kerja, atau isu yang beredar dalam rangka peningkatan kualitas pelayanan bagi perusahaan atau instansi itu sendiri.

Ada banyak metode yang dapat digunakan dalam analisis sentimen, antara lain *Naïve Bayes Classifier*, *Support Vector Machine*, *Artificial Neural Network*, *Decision Tree*, dan *K-Nearest Neighbor*. Pada penelitian ini peneliti ingin membandingkan dua metode yaitu *Naïve Bayes Classifier* dan *Support Vector Machine* untuk mengklasifikasikan ulasan mengenai BPJS kedalam kelas positif atau negatif. Hasil klasifikasi yang diperoleh menggunakan kedua metode tersebut, kemudian dievaluasi menggunakan ukuran ketepatan klasifikasi untuk melihat kinerja klasifikasi.

Adapun penelitian terdahulu yang berkaitan dengan analisis sentimen antara lain penelitian Gading Teguh Santoso (2021) berjudul “Analisis Sentimen Pada *Tweet* Dengan Tagar #BPJSRasaRentenir Menggunakan Metode *Support Vector Machine (SVM)*” dengan hasil akurasi sebesar 94%. Serta penelitian Hanafi Rahman (2021) berjudul “Klasifikasi Sentimen Masyarakat Terhadap Layanan Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan di *Twitter* Menggunakan Metode *K-Nearest Neighbor*” dengan hasil tingkat akurasi sebesar 87,33%. Dan penelitian Zia Ayu Nuansa Gumilang (2018) berjudul “Implementasi *Naïve Bayes Classifier* dan Asosiasi Untuk Analisis Sentimen Data Ulasan Aplikasi *E-Commerce Shopee* Pada Situs *Google Play*” diperoleh tingkat akurasi sebesar 97,4%.

Berdasarkan penjelasan latar belakang diatas, maka penulis melakukan penelitian analisis sentimen menggunakan data *twitter* mengenai BPJS, karena BPJS merupakan salah satu program jaminan sosial yang digunakan oleh sebagian masyarakat Indonesia dan dengan kondisi pandemi seperti sekarang ini penulis ingin mengetahui apakah kinerja BPJS menurun atau semakin baik dengan membandingkan dua metode yaitu *Naïve Bayes Classifier* dan *Support Vector Machine*. Karena pada penelitian sebelumnya kedua metode ini memiliki nilai akurasi yang lebih tinggi dibandingkan metode *K-Nearest Neighbor*.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang diatas, maka rumusan masalah pada penelitian ini adalah sebagai berikut.

1. Bagaimana hasil penerapan metode *Naïve Bayes Classifier* dan *Support Vector Machine* dalam mengklasifikasikan tanggapan atau ulasan pengguna *twitter* mengenai BPJS ?
2. Bagaimana hasil akurasi dari metode *Naïve Bayes Classifier* dan *Support Vector Machine* dalam mengklasifikasikan tanggapan atau ulasan pengguna *twitter* mengenai BPJS ?
3. Bagaimana perbandingan hasil ketepatan klasifikasi antara metode *Naïve Bayes Classifier* dan *Support Vector Machine* ?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut.

1. Sumber data yang digunakan berasal dari *Twitter* berupa data *Tweets* umum
2. Data *tweets* yang digunakan berbahasa Indonesia
3. *Tweets* yang digunakan dimulai dari tahun 2019-2021 dengan kata kunci “BPJS”, “Klaim BPJS”, dan “Badan Penyelenggara Jaminan Sosial”
4. Klasifikasi *tweet* akan digolongkan dalam dua kelompok yaitu positif dan negatif.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah pada penelitian ini maka tujuan dari penelitian ini adalah sebagai berikut.

1. Mendapatkan hasil penerapan metode *Naïve Bayes Classifier* dan *Support Vector Machine* dalam mengklasifikasikan tanggapan pengguna *Twitter* mengenai BPJS
2. Mendapatkan hasil akurasi dari metode *Naïve Bayes Classifier* dan *Support Vector Machine* dalam mengklasifikasikan tanggapan pengguna *twitter* mengenai BPJS
3. Mendapatkan perbandingan hasil ketepatan klasifikasi dari metode *Naïve Bayes Classifier* dan *Support Vector Machine*.

1.5 Manfaat Penelitian

Manfaat yang diharapkan pada penelitian ini adalah sebagai berikut.

1. Memberikan informasi mengenai persepsi pengguna *twitter* mengenai BPJS
2. Memberikan informasi mengenai penerapan metode *Naïve Bayes Classifier* dan *Support Vector Machine* pada analisis sentimen
3. Hasil dari penelitian ini dapat digunakan sebagai evaluasi dan acuan untuk mengoptimalkan layanan BPJS.

(Halaman ini sengaja dikosongkan)

BAB II TINJAUAN PUSTAKA

2.1 Penelitian Sebelumnya

Penelitian terdahulu sebagai bahan referensi penulis sesuai dengan topik tugas akhir adalah sebagai berikut.

1. Analisis Sentimen pada *Tweet* dengan Tagar #BPJSRasaRentenir Menggunakan Metode *Support Vector Machine* (SVM)

Jejaring sosial membantu pengguna internet dalam berkomunikasi. Hal ini dikarenakan pengguna jejaring sosial dapat menyampaikan pesan dengan memanfaatkan fasilitas yang disediakan oleh setiap media sosial. Pesan-pesan para pengguna media sosial dapat dimanfaatkan dalam berbagai hal, seperti *review* terhadap suatu produk atau *review* terhadap suatu masalah pada politik atau masalah sosial sekarang ini. Hal ini bisa dilakukan dengan menganalisis sentimen para pengguna sosial media. Penelitian tersebut menggunakan data *tweet* dengan tagar #BPJSRasaRentenir dan menggunakan metode *Support Vector Machine* yang dilakukan dengan mengklasifikasikan sentimen kedalam kelas positif dan negatif. Dengan menggunakan 200 data sampel diperoleh hasil akurasi sebesar 94% sehingga penelitian ini menghasilkan model yang dapat melakukan klasifikasi dan prediksi tentang tanggapan masyarakat terhadap BPJS. Dan pada penelitian ini analisis sentimen masyarakat lebih beranggapan negatif (Santoso, 2021)

2. Klasifikasi Sentimen Masyarakat Terhadap Layanan Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan di *Twitter* Menggunakan Metode *K-Nearest Neighbor*

BPJS adalah badan hukum yang dibentuk oleh Undang-Undang untuk menyelenggarakan program jaminan kesehatan. Melalui program ini pemerintah berupaya agar masyarakat dapat mendapatkan layanan kesehatan yang baik. Namun pada penerapannya masih menimbulkan sentimen positif dan negatif masyarakat tentang layanan BPJS Kesehatan. Penelitian tersebut dilakukan untuk menerapkan dan menguji akurasi dari metode *K-Nearest Neighbor* dalam mengklasifikasikan sentimen masyarakat terhadap layanan BPJS Kesehatan di *Twitter*. Dengan menggunakan pembagian dataset 90:10 dengan data latih 2700 dan data uji 300 dan *threshold* 18 memiliki akurasi tertinggi dengan *K* bernilai 9 diperoleh hasil akurasi sebesar 87,33% (Rahman, 2021)

3. Implementasi *Naïve Bayes Classifier* dan Asosiasi Untuk Analisis Sentimen Data Ulasan Aplikasi *E-Commerce Shopee* pada Situs *Google Play*

Salah satu pengaruh dari meningkatnya jumlah pengguna internet di Indonesia adalah semakin menjamurnya kegiatan berbelanja melalui media internet. Seiring perkembangannya, kini *e-commerce* dapat diakses dengan mudah melalui *mobile phone* dalam bentuk aplikasi yang dapat diunduh dengan mudah. Salah satu contoh aplikasi *e-commerce* yang memiliki reputasi baik di Indonesia ialah *Shopee*. Sehingga pada penelitian tersebut penulis tertarik untuk meneliti *Shopee* yang datanya didapatkan dari situs *google play* menggunakan metode *Naïve Bayes Classifier* untuk mengklasifikasikan ulasan berdasarkan kategori kelas sentimen positif dan negatif. Populasi yang digunakan pada penelitian tersebut adalah semua data ulasan *Shopee* pada bulan Januari 2018 hingga Maret 2018, dan untuk sampel yang digunakan adalah ulasan *Shopee* sejak pertengahan *upgrade* aplikasi *Shopee* yaitu tanggal 17 Januari 2018 hingga terakhir *upgrade* pada akhir Maret 2018. Hasil yang diperoleh dari penelitian tersebut dengan perbandingan data latih dan data uji sebesar 80 : 20 diperoleh hasil klasifikasi sentimen dengan tingkat akurasi sebesar 97,4% (Gumilang, 2018)

2.2 Twitter

Twitter adalah sebuah situs web yang dimiliki dan dioperasikan oleh Twitter Inc., yang menawarkan jaringan sosial berupa microblog sehingga memungkinkan penggunaanya untuk mengirim dan membaca pesan *Tweets*. Mikroblog adalah salah satu jenis alat komunikasi online dimana pengguna dapat memperbarui status tentang mereka yang sedang memikirkan dan melakukan sesuatu (Harijiatno, 2019). *Tweets* dapat berupa teks dan menautkan foto atau video. Melalui *tweeting*, pengguna *twitter* dapat berinteraksi dengan pengguna lain untuk membahas topik tertentu, seperti isu-isu sosial yang terjadi di sekitar mereka, mengenai opini tentang produk, produk yang sedang beredar, atau sekedar berbagi tentang kegiatan yang sedang berlangsung. Logo *twitter* ditampilkan pada Gambar 2.1



Gambar 2.1 Logo *Twitter*

(Sumber : <https://help.twitter.com/id/using-twitter/>)

Sejak peluncurannya, jumlah pengguna *Twitter* terus meningkat dari tahun ke tahun karena penggunaannya bisa dilakukan oleh siapapun, kapanpun, dan dimanapun. Selengkapnya fitur yang tersedia pada jejaring sosial *Twitter* beserta fungsinya dapat dilihat pada tabel 2.1

Tabel 2.1 Fitur *Twitter* dan Kegunaannya

Fitur	Keterangan
Laman Utama (<i>Home</i>)	Fitur <i>Home</i> berguna untuk melihat <i>tweets</i> kiriman dari pengguna lain yang diikuti oleh sebuah akun. Kumpulan <i>tweets</i> yang ditampilkan disebut linimasa.
Profil (<i>Profile</i>)	Halaman profil menampilkan data diri dari suatu akun. Selain itu juga menampilkan <i>tweets</i> yang pernah dibuat, media berupa foto dan video yang pernah diunggah, dan juga menampilkan kumpulan <i>tweets</i> yang ditandai sebagai favorit.
Pengikut (<i>Followers</i>)	Pengikut adalah pengguna lain yang menjadikan suatu akun sebagai teman. Bila pengguna lain menjadi pengikut akun seseorang, maka <i>tweets</i> seseorang yang diikuti tersebut akan masuk ke halaman utamanya.
<i>Following</i>	<i>Following</i> adalah akun seseorang yang mengikuti akun pengguna lain agar <i>tweets</i> yang dikirim oleh orang yang diikuti tersebut masuk ke dalam halaman utama.
<i>Mentions</i>	Fitur ini dilambangkan dengan '@' di depan nama pengguna (<i>username</i>). Berfungsi untuk menandai secara langsung pengguna lain yang akan diajak berinteraksi ataupun berbalas <i>tweet</i> dengan lebih dari satu pengguna.
<i>Favorite</i>	Fitur ini memfasilitasi untuk menyimpan <i>tweets</i> yang disukai dalam satu kelompok <i>tweets</i> pada dinding profil pengguna sehingga walau <i>tweets</i> tersebut sudah tertimbun <i>tweets</i> lain di linimasa, <i>tweets</i> masih bisa ditemukan di profil pengguna.

Tabel 2.1 Fitur *Twitter* dan Kegunaannya (Lanjutan)

Fitur	Keterangan
Pesan Langsung (<i>Direct Message</i>)	Fitur yang familiar disebut dengan “Dm” ini digunakan untuk bertukar pesan secara pribadi antara dua pengguna. Fitur ini memfasilitasi pengiriman dalam bentuk teks dan gambar.
<i>Hashtag</i>	Fitur ini dilambangkan dengan “#” di depan topik tertentu. <i>Hashtag</i> digunakan untuk pencarian topik sejenis yang ditulis oleh orang lain.
<i>List</i>	Fitur yang memungkinkan untuk mengelompokkan akun-akun yang diikuti ke dalam satu grup sehingga memudahkan untuk dapat melihat secara keseluruhan nama para pengguna (<i>username</i>) yang diikuti.
Topik Terkini (<i>Trending Topic</i>)	Topik yang sedang banyak dibicarakan oleh banyak pengguna dalam suatu waktu yang bersamaan. <i>Trending topic</i> terbagi menjadi dua, yaitu <i>trending topic</i> dunia dan <i>trending topic</i> negara.

2.3 Badan Penyelenggara Jaminan Sosial

Berdasarkan UU Nomor 24 Tahun 2011 tentang Badan Penyelenggara Jaminan Sosial, BPJS adalah badan hukum yang dibentuk untuk menyelenggarakan program jaminan sosial. BPJS dibagi menjadi dua yaitu BPJS Kesehatan dan BPJS Ketenagakerjaan. BPJS Kesehatan berfungsi menyelenggarakan program jaminan kesehatan, dan BPJS Ketenagakerjaan berfungsi menyelenggarakan program jaminan kecelakaan kerja, jaminan hari tua, jaminan pensiun, dan jaminan kematian (BPJS Kesehatan, 2021)

BPJS bertujuan untuk mengelola dana publik, yaitu dana jaminan sosial untuk kepentingan peserta. Pada 1 Januari 2014, Pemerintah mengoperasikan BPJS Kesehatan atas perintah UU BPJS. Sejak BPJS Kesehatan beroperasi, penyelenggaraan program jaminan kesehatan sosial terhadap program-program kesehatan perorangan dialihkan kepada BPJS Kesehatan. Beberapa pengalihan program yang terjadi adalah sebagai berikut (Rahman, 2021)

1. Kementerian Kesehatan tidak lagi menyelenggarakan program Jaminan Kesehatan Masyarakat (Jamkesmas)
2. Kementerian Pertahanan, Tentara Nasional Indonesia, dan Kepolisian Republik Indonesia tidak lagi menyelenggarakan program pelayanan kesehatan bagi pesertanya, kecuali untuk pelayanan kesehatan tertentu berkaitan dengan kegiatan operasionalnya, yang ditetapkan dengan Peraturan Presiden
3. PT Jamsostek (Persero) tidak lagi menyelenggarakan program jaminan pemeliharaan kesehatan.

2.4 Analisis Sentimen

Sentimen dapat diartikan sebagai pendapat ataupun pandangan yang didasarkan pada perasaan berlebihan terhadap sesuatu. Sentimen biasanya terdapat dalam pernyataan serta kalimat yang memiliki pendapat. Sentimen juga berguna untuk mengetahui perasaan yang diberikan seseorang terhadap topik atau objek tertentu (Nugraha, 2020).

Analisis sentimen atau bisa juga disebut *opinion mining* adalah salah satu bidang ilmu komputer yang menganalisis pendapat, penilaian, sikap, emosi, sentimen, dan evaluasi terhadap suatu produk, layanan, organisasi, individu, tokoh publik, acara, maupun kegiatan tertentu (Liu, 2012). Menurut (Kristiyanti, 2015), analisis sentimen atau *opinion mining* adalah studi komputasi mengenai pendapat, perilaku, dan emosi seseorang terhadap entitas. Entitas tersebut dapat menggambarkan individu, kejadian atau topik.

Tugas yang dilakukan oleh analisis sentimen adalah mengelompokkan polaritas yang terdapat pada suatu teks, baik dalam dokumen maupun kalimat apakah pendapat tersebut bersifat positif, negatif, atau netral. Terdapat beberapa proses yang dapat dilakukan dalam analisis sentiment dan dibagi menjadi beberapa tingkatan, sebagai berikut (Nugraha, 2020).

1. *Document-level Sentiment Analysis*

Dalam kasus ini, sebuah asumsi implisit merupakan suatu dokumen yang menyatakan pendapat tentang target tertentu. Tujuan dari tingkatan ini adalah untuk mengetahui dokumen tertentu dapat menunjukkan suatu klasifikasi sentimen, baik itu positif maupun negatif,

2. *Sentence and Phrase-level Sentiment Analysis*

Seringkali *document-level sentiment analysis* dipecah menjadi unit terkecil sesuai dengan kata, frasa, ataupun kalimat individu. Dalam tingkatan ini setiap kalimat akan dianalisis satu persatu dan akan diklasifikasikan ke dalam kategori positif maupun negatif,

3. *Entity and Aspect-Level Opinions*

Pada tingkatan ini sentimen analisis diukur dengan melihat suatu pendapat atau opini terhadap karakteristik dari suatu entitas.

Ada beragam jenis analisis sentimen yang dapat digunakan untuk mengidentifikasi respon pengguna. Dari melihat polaritas pendapat hingga mengidentifikasi niat pengguna. Beberapa jenis analisis sentimen adalah sebagai berikut (Area, Universitas Medan, 2016)

1. *Fine-Grained Sentiment Analysis*

Merupakan salah satu jenis analisis sentiment yang paling umum. Fokusnya ada pada tingkat polaritas pendapat. Tipe analisis ini akan mengelompokkan respon atau pendapat ke dalam beberapa kategori seperti sangat positif, agak positif, netral, agak negatif, dan negatif

2. *Intent Sentiment Analysis*

Tipe ini bertujuan untuk mengidentifikasi dan menggali lebih dalam motivasi di balik pesan pengguna untuk melihat apakah itu termasuk keluhan, saran, pendapat, pertanyaan, atau justru penghargaan terhadap produk atau layanan

3. *Aspect-Based Sentiment Analysis*

Pada tipe ini lebih berfokus pada elemen-elemen yang lebih spesifik dari suatu produk atau layanan. Analisis sentimen berbasis aspek ini juga memungkinkan untuk menghubungkan sentimen spesifik dengan berbagai aspek produk atau layanan.

Cara kerja sentimen analisis dalam mengambil data dibagi menjadi tiga langkah, yaitu

1. Klasifikasi

Pertama perlu mengklasifikasikan data yang dinilai sebagai opini dari sebuah teks. Ada tiga klasifikasi dalam metode analisis sentimen yang dapat dilakukan, antara lain

- a. *Machine Learning*, fitur-fitur yang terdapat di dalamnya dapat mengenali sentimen dalam sebuah teks. Metode ini semakin populer saat ini karena dinilai representatif
- b. *Lexicon-Based*, menggunakan berbagai kata yang dinilai dengan skor polaritas untuk mengetahui tanggapan masyarakat atau konsumen mengenai suatu topik. Keunggulannya adalah tidak memerlukan data pelatihan, dan kelemahannya adalah masih banyak kata yang belum termuat dalam kamus ini
- c. Campuran, menggabungkan metode *machine learning* dan *lexicon* agar memberikan hasil yang menjanjikan. Tetapi metode ini jarang digunakan

2. Evaluasi

Setelah data terklasifikasi, metode selanjutnya adalah menggunakan matriks evaluasi seperti *Precision*, *Recall*, *F-Score*, dan *Accuracy*. Proses ini juga melibatkan pengukuran rata-rata untuk menangani data yang masuk ke dalam dua klasifikasi atau lebih

3. Visualisasi Data

Visualisasi data dilakukan dengan menggunakan bagan sesuai kebutuhan perusahaan biasanya menggunakan grafik, histogram, atau matriks. Namun, hasil akhir dari analisis sentimen bisa sangat bervariasi. Karena itulah teknik visualisasi data berupa *word cloud* cukup efektif untuk menampilkan hasil analisis.

2.5 Data Mining

Data mining adalah proses menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Data mining merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Hal-hal penting yang terkait dengan data mining adalah sebagai berikut (Bramer, 2007)

1. Merupakan suatu proses otomatis terhadap data yang sudah ada
2. Data yang diproses merupakan data yang sangat besar
3. Tujuan dari data mining adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Data mining mempunyai tujuan sebagai *Explanatory* yaitu untuk menjelaskan beberapa kondisi terkait dengan suatu penelitian, *Confirmatory* digunakan untuk mempertegas hipotesis, dan *Exploratory* yang berguna dalam menganalisis data untuk hubungan baru yang tidak diharapkan (Mustika & Ardilla, 2021). Data mining dibagi menjadi beberapa kelompok berdasarkan tugas dapat dilakukan, yaitu

1. *Estimation*, untuk menerka sebuah nilai yang belum diketahui, seperti menerka penghasilan seseorang ketika informasi mengenai orang tersebut diketahui
2. *Prediction*, untuk memperkirakan nilai masa mendatang, seperti memprediksi stik barang satu tahun ke depan
3. *Classification*, proses penemuan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui
4. *Clustering*, pengelompokan mengidentifikasi data yang memiliki karakteristik tertentu
5. *Association*, dinamakan juga analisis keranjang pasar dimana fungsi ini mengidentifikasi item-item produk yang kemungkinan dibeli konsumen bersamaan dengan produk lain.

2.6 Text Mining

Text mining dapat didefinisikan secara luas sebagai proses intensif pengetahuan dimana pengguna berinteraksi dengan kumpulan dokumen dari waktu ke waktu dengan menggunakan seperangkat alat analisis. *Text mining* berusaha untuk mengekstrak informasi penting melalui identifikasi dan eksplorasi pola yang menarik (Findawati & Rosidi, 2020)

Menurut Syakuro (2017) *text mining* adalah proses menambang data berupa teks dimana sumber data biasanya didapat dari dokumen dan bertujuan untuk mencari kata-kata yang dapat mewakili isi dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen tersebut. *Text mining* merupakan penerapan konsep dan teknik data mining untuk mencari pola dalam teks untuk mencari informasi yang bermanfaat untuk tujuan tertentu. Perbedaan *text mining* dan *data mining* yaitu pada penggunaan sumber data. Pada *data mining*, sumber data yang digunakan berasal dari pola-pola tertentu dan terstruktur. Sedangkan pada *text mining*,

sumber data berasal dari teks yang relatif tidak terstruktur karena menggunakan tata bahasa manusia atau biasa disebut *natural language* (Subagyo, 2021).

Text mining dapat menganalisa dokumen, mengelompokkan dokumen berdasarkan kata-kata yang terkandung di dalamnya, serta menentukan kesamaan di antara dokumen untuk mengetahui hubungan variabel satu dengan variabel lainnya. Penerapan *text mining* yang paling umum dilakukan saat ini adalah penyaringan spam, analisis sentimen, mengukur preferensi pelanggan, meringkas dokumen, pengelompokan topik penelitian, dan lainnya (Findawati & Rosidi, 2020). Ada beberapa manfaat dari penggunaan *text mining* seperti mengetahui kepuasan pelanggan sehingga dapat membantu perusahaan untuk mendeteksi masalah produk dan bisnis sebelum masalah menjadi besar dan mempengaruhi penjualan, mendeteksi kecurangan, penipuan, dan manajemen risiko (Stedman, Copyright 2010 - 2022).

2.7 Text Preprocessing

Text preprocessing adalah langkah awal dari *text mining* yang bertujuan untuk mempersiapkan dokumen menjadi data yang akan diolah pada tahap selanjutnya. Karena *text mining* memiliki struktur yang tidak baku, maka diperlukan proses perubahan menjadi data terstruktur sesuai kebutuhan, yang biasanya akan menjadi nilai-nilai numerik (Triawati, 2009). Penggunaan *text preprocessing* yang tepat dapat meningkatkan akurasi pada kasus klasifikasi.

2.7.1 Cleaning

Merupakan proses pembersihan tweet dari kata yang tidak diperlukan seperti HTML, *emoticons*, *hashtag*, *username*, simbol, angka, tanda baca, *duplicate*, dan url. Untuk *emoticons*, diubah terlebih dahulu kedalam format html (Permatasari, 2021). Berikut merupakan ilustrasi dari *cleaning*.

Tabel 2.2 Proses *Cleaning* Pada *Tweet*

Sebelum	Sesudah
Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan. #disnakerlampung #KPU #KPUMelayani #bpjskesehatan #lampostco https://t.co/2nUaexGu5i	Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan

Setelah data *tweet* melewati proses *cleaning* dapat dilihat bahwa terdapat beberapa komponen yang hilang seperti *hashtag*, *url*, tanda baca, dan angka.

2.7.2 Case Folding

Case folding merupakan proses mengubah kata ke dalam format yang sama, dalam hal ini menjadi format *lowercase* atau *uppercase* (Hidayatullah, 2016). Pada penelitian ini menggunakan format *lowercase*. Berikut merupakan ilustrasi dari *case folding*.

Tabel 2.3 Proses *Case Folding* Pada *Tweet*

Sebelum	Sesudah
Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan	disnaker lampung mengusulkan petugas komisi pemilihan tingkat desa provinsi dilindungi bpjs ketenagakerjaan

2.7.3 Stopword Removal

Merupakan proses menghapus kata-kata umum dan sering muncul tetapi tidak memiliki pengaruh yang signifikan terhadap makna dari sebuah kalimat. Pada penelitian ini kamus yang digunakan merupakan kamus bawaan dari *Python* yaitu Sastrawi. Contoh *stopword* dalam Bahasa Indonesia adalah penggunaan kata “dan”, “atau”, “yang”, “itu”, dan lainnya. Selain itu pada proses ini juga terjadi penghapusan kata kunci seperti “bpjs”, “klaim”, “badan”, “penyelenggara”, “jaminan”, “sosial”, “kesehatan”, dan “ketenagakerjaan”. Berikut merupakan ilustrasi dari *stopword removal*.

Tabel 2.4 Proses *Stopword Removal* Pada *Tweet*

Sebelum	Sesudah
disnaker lampung mengusulkan petugas komisi pemilihan tingkat desa provinsi dilindungi bpjs ketenagakerjaan	disnaker lampung mengusulkan petugas komisi pemilihan tingkat desa provinsi dilindungi

2.7.4 Stemming

Merupakan proses mengubah kata menjadi kata dasar dengan menghilangkan awalan, akhiran, sisipan, dan *confixes* (kombinasi awalan dan akhiran) (Ariadi & Fithriasri, 2015). Tujuan dari tahapan ini adalah agar suatu kata sesuai dengan kaidah Bahasa Indonesia yang benar. Dalam penelitian ini, kamus yang digunakan adalah kamus bawaan dari bahasa pemrograman *Python* yaitu Sastrawi. Berikut merupakan ilustrasi dari proses *stemming*.

Tabel 2.5 Proses *Stemming* Pada *Tweet*

Sebelum	Sesudah
disnaker lampung mengusulkan petugas komisi pemilihan tingkat desa provinsi dilindungi	disnaker lampung usul tugas komisi pilih tingkat desa provinsi lindung

2.7.5 Tokenization

Merupakan proses memecah kalimat menjadi kata-kata yang lebih berarti dan bermakna. Hasil dari proses ini hanyalah berisi kumpulan kata saja (Putri, 2016). Berikut merupakan ilustrasi dari proses *tokenization*.

Tabel 2.6 Proses *Tokenization* Pada *Tweet*

Sebelum	Sesudah
disnaker lampung usul tugas komisi pilih tingkat desa provinsi lindung	['disnaker', 'lampung', 'usul', 'tugas', 'komisi', 'pilih', 'tingkat', 'desa', 'provinsi', 'lindung']

2.8 Term Weighting

Term weighting atau proses pembobotan pada tiap kata adalah metode yang digunakan untuk mendapatkan nilai frekuensi dari masing-masing kata dalam suatu dokumen sehingga diperoleh perbandingan antar kata satu dengan lainnya karena setiap kata memiliki tingkat kepentingan yang berbeda (Abualigah, dkk., 2016). Salah satu metode yang sering digunakan untuk menghitung bobot dari suatu kata adalah *Term Frequency-Inverse Document Frequency* (TF-IDF).

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan pengukuran yang digunakan untuk mengukur seberapa penting suatu kata terhadap kumpulan dokumen. *Term Frequency* (TF) adalah frekuensi kemunculan sebuah kata dalam satu dokumen, sedangkan

Inverse Document Frequency (IDF) adalah frekuensi kemunculan sebuah kata pada seluruh kumpulan dokumen (Jannah, dkk, 2018). Rumus untuk menentukan pembobotan dengan TF-IDF ditunjukkan pada persamaan berikut (Permatasari, 2021),

$$TF_{ij} = \frac{f_{ij}}{\sum_{i' \in j} f_{ij}} \quad (2.1)$$

$$IDF = \log\left(\frac{N}{DF_{ij}}\right) \quad (2.2)$$

$$W_{ij} = TF_{ij} \times IDF \quad (2.3)$$

dimana :

f_{ij} : jumlah kata i untuk setiap *tweet*,

$\sum_{i' \in j} f_{ij}$: jumlah *tweet* yang mengandung kata i ,

W_{ij} : bobot TF-IDF pada kata kunci ke- i dan *tweet* ke- j ,

TF_{ij} : jumlah kemunculan kata i pada *tweet* ke- j ,

DF_{ij} : jumlah *tweet* ke- j yang mengandung kata i ,

N : jumlah seluruh *tweet*.

Berikut merupakan ilustrasi pembobotan TF-IDF.

Tabel 2.7 Ilustrasi *tweet*

<i>Tweet</i> ke-	<i>Tweet</i>
T1	perokok bayar premi mahal
T2	iur premi naik defisit kurang

Dari Tabel 2.2 diatas, langkah selanjutnya adalah menghitung bobot per kata pada *tweet* yang ada di Tabel 2.2 dengan menggunakan Persamaan 2.1, 2.2, dan 2.3. Berikut merupakan tabel ilustrasi perhitungan TF-IDF

Tabel 2.8 Ilustrasi TF-IDF

Kata	<i>Count</i>		DF_{ij}	TF_{ij}		IDF $\log\left(\frac{N}{DF_{ij}}\right)$	$TF-IDF$	
				$\frac{f_{ij}}{\sum_{i' \in j} f_{ij}}$			$TF_{ij} \times IDF$	
	T1	T2		T1	T2		T1	T2
bayar	1	0	1	0,25	0,00	0,90	0,23	0,00
defisit	0	1	1	0,00	0,20	0,90	0,00	0,18
iur	0	1	1	0,00	0,20	0,90	0,00	0,18
kurang	0	1	1	0,00	0,20	0,90	0,00	0,18
mahal	1	0	1	0,25	0,00	0,90	0,23	0,00
naik	0	1	1	0,00	0,20	0,90	0,00	0,18
perokok	1	0	1	0,25	0,00	0,90	0,23	0,00
premi	1	1	2	0,25	0,20	0,60	0,15	0,12
$\sum_{i' \in j} f_{ij}$	4	5						

2.9 Oversampling

Metode *oversampling* adalah salah satu metode yang bertujuan untuk menambahkan jumlah data pada kelas minoritas dengan memanfaatkan teknik *sampling* pada data training kelas minoritas sehingga diharapkan rasio antar kelas minoritas dan mayoritas dapat lebih seimbang (Mahmood, 2015). Konsep penambahan data pada *oversampling* dibagi menjadi dua, yaitu *oversampling* menggunakan data asli seperti metode *Random Oversampling* dan penambahan menggunakan data sintetik seperti *Synthetic Minority Over-sampling Technique* (SMOTE) (Chawla, 2002). Pada penelitian ini metode yang akan digunakan dalam *oversampling* adalah *Random Oversampling*. Pada algoritma ROS (*Random Oversampling*), data kelas minoritas dipilih secara acak kemudian ditambahkan ke dalam data *training*. Proses pemilihan dan penambahan ini diulang-ulang sampai jumlah data kelas minoritas sama dengan jumlah kelas mayoritas.

2.10 K-fold Cross Validation

K-fold cross validation adalah teknik memvalidasi keakuratan sebuah model yang dibentuk berdasarkan data set tertentu. Data yang digunakan dalam proses pembentukan model disebut sebagai data latih atau *training* dan data yang digunakan untuk memvalidasi disebut sebagai data uji atau *testing* (Davidson & Hinkley, 1997). Metode ini banyak digunakan peneliti karena dapat mengurangi bias yang terjadi dalam pengambilan sampel. *K-fold cross validation* secara berulang membagi data menjadi data latih dan data uji sehingga tiap data mendapat kesempatan yang sama untuk menjadi data uji (Gokgoz & Subasi, 2015). Masing-masing *fold* memiliki jumlah data dengan ukuran yang sama atau mendekati sama. Selama *k* subset akan dipilih satu *fold* sebagai data uji dan sisanya menjadi data latih. Menurut Kohavi (1995), penggunaan jumlah *fold* terbaik untuk uji validitas dianjurkan menggunakan *10-fold cross validation*.



Gambar 2.2 Algoritma *Cross Validation*

(Sumber : <https://drzinph.com/nested-cross-validation-cross-validation-series-part-1/>)

2.11 Teorema Bayes

Teorema *bayes* merupakan teorema yang mengacu pada konsep probabilitas bersyarat (Tan, dkk, 2006). Metode ini merupakan pendekatan statistic untuk melakukan inferensi induksi pada persoalan klasifikasi. Larose (2006) menyatakan probabilitas bersyarat dalam Persamaan 2.3

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.3)$$

dimana :

A : kelas ulasan (positif atau negatif),

B : kata pada setiap ulasan,

$P(A \cap B)$: probabilitas interaksi kelas ulasan dan kata pada setiap ulasan,

$P(B)$: probabilitas kata pada setiap ulasan.

Demikian pula $P(B|A) = \frac{P(A \cap B)}{P(A)}$, sehingga

$$P(A \cap B) = P(B|A) \times P(A) \quad (2.4)$$

Kemudian substitusi Persamaan 2.3 ke Persamaan 2.4, maka diperoleh

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2.5)$$

dengan :

$P(A|B)$: probabilitas kelas ulasan berdasarkan kondisi kata pada setiap ulasan,

$P(B|A)$: probabilitas kata pada setiap ulasan berdasarkan kondisi pada kelas ulasan,

$P(A)$: probabilitas kelas ulasan,

$P(B)$: probabilitas kata pada setiap ulasan.

2.12 Naïve Bayes Classifier

Naïve Bayes Classifier merupakan sebuah algoritma pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari data yang diberikan (Subagyo, 2021). Menurut Han, Kamber, & Pei (2012) *Naïve Bayes Classifier* adalah salah satu teknik *data mining* yang sering digunakan untuk mengklasifikasikan data dalam jumlah yang besar dan dapat untuk memprediksi probabilitas keanggotaan suatu *class*. Kelebihan *Naïve Bayes Classifier* adalah algoritma sederhana tetapi memiliki akurasi yang tinggi. Secara umum model probabilitas untuk sebuah klasifikasi adalah probabilitas bersyarat berdasarkan Persamaan 2.5 yang dapat disesuaikan seperti Persamaan 2.6.

$$P(C_j|X_1, X_2, \dots, X_n) \quad (2.6)$$

dengan :

C_j : kelas ulasan, dimana

J_1 : sentimen positif

J_2 : sentimen negatif

X_1, X_2, \dots, X_n : kata pada setiap ulasan, dimana

X_1 : kata pertama

X_2 : kata kedua

X_n : kata ke- n

$P(C_j|X_1, X_2, \dots, X_n)$: probabilitas X_1, X_2, \dots, X_n pada kelas C_j . Sehingga, berdasarkan Persamaan 2.6, maka

$$P(C_j|X_1, X_2, \dots, X_n) = \frac{P(X_1, X_2, \dots, X_n|C_j)P(C_j)}{P(X_1, X_2, \dots, X_n)} \quad (2.7)$$

dimana,

$$P(X_1, X_2, \dots, X_n|C_j) = P(X_1|C_j) \times P(X_2|C_j) \times \dots \times P(X_n|C_j) \quad (2.8)$$

dengan :

$P(C_j|X_1, X_2, \dots, X_n)$: probabilitas kelas C_j pada X_1, X_2, \dots, X_n (posterior)

$P(X_1, X_2, \dots, X_n|C_j)$: probabilitas X_1, X_2, \dots, X_n pada kelas C_j (likelihood)

$P(C_j)$: probabilitas dari kelas C_j (prior)

$P(X_1, X_2, \dots, X_n)$: probabilitas dari X_1, X_2, \dots, X_n (evidence)

Nilai dari $P(X_1, X_2, \dots, X_n)$ selalu tetap untuk setiap kelas pada satu ulasan. Sehingga Persamaan 2.7 dapat ditulis sebagai berikut.

$$P(C_j|X_1, X_2, \dots, X_n) = P(X_1, X_2, \dots, X_n|C_j) \times P(C_j) \quad (2.9)$$

Nilai $P(C_j|X_1, X_2, \dots, X_n)$ nantinya akan dibandingkan dengan nilai $P(C_j|X_1, X_2, \dots, X_n)$ lainnya untuk menentukan klasifikasi kelas ulasan tersebut. Klasifikasi dokumen bertujuan untuk menentukan kelas terbaik untuk suatu dokumen. Kelas terbaik ditentukan dengan mencari *maximum a posterior* (MAP) melalui Persamaan 2.10

$$C_{MAP} = \operatorname{argmax}_{C_j=C} P(C_j|X_1, X_2, \dots, X_n) \quad (2.10)$$

Menurut Persamaan 2.9 maka Persamaan 2.10 dapat disesuaikan menjadi Persamaan 2.11

$$C_{MAP} = \operatorname{argmax}_{C_j=C} P(X_1, X_2, \dots, X_n|C_j) \times P(C_j) \quad (2.11)$$

Persamaan 2.11 dapat disederhanakan menjadi Persamaan 2.12

$$C_{MAP} = \operatorname{argmax}_{C_j=C} P(C_j) \times \prod P(X_n|C_j) \quad (2.12)$$

Nilai $P(C_j)$ dapat dihitung dengan Persamaan 2.13

$$P(C_j) = \frac{N_c}{N} \quad (2.13)$$

dimana,

N_c : jumlah dokumen dengan kelas c

N : jumlah seluruh dokumen

Untuk setiap probabilitas X_n untuk tiap kelas C_j dihitung dengan Persamaan 2.14

$$P(X_n|C_j) = \frac{n_i}{n} \quad (2.14)$$

dimana,

- n_i : jumlah kemunculan term X_n dalam kelas C_j
- n : jumlah frekuensi seluruh term dalam kelas C_j

Karena suatu kata dalam kelas akan memiliki nilai 0 hal ini dapat menyebabkan perhitungan $P(X_n|C_j)$ bernilai 0, maka untuk mengatasi masalah tersebut, diterapkan teknik *add-one* atau *Laplace smoothing*, sehingga Persamaan 2.14 berubah menjadi Persamaan 2.15

$$P(X_n|C_j) = \frac{|n_i + 1|}{|n + \text{kosakata}|} \tag{2.15}$$

dimana,

kosakata : jumlah seluruh term pada vocabulary

Dan untuk perhitungan menggunakan pembobotan TF-IDF dapat dilihat pada Persamaan 2.16

$$P(X_n|C_j) = \frac{|W_{ij} + 1|}{|W + \text{kosakata}|} \tag{2.16}$$

dimana,

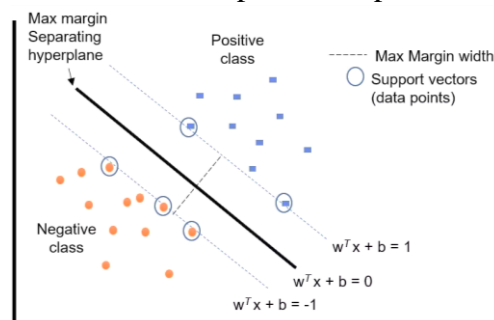
W_{ij} : bobot TF-IDF term X_n dalam kelas C_j

W : jumlah bobot TF-IDF seluruh term pada kelas C_j

kosakata : Jumlah IDF seluruh *term* pada *vocabulary*.

2.13 Support Vector Machine

Menurut Santos (2015) *Support Vector Machine* adalah teknik prediksi yang digunakan untuk klasifikasi dan regresi. Menurut Williams (2011), *Support Vector Machine* adalah salah satu metode statistika yang mempelajari area yang memisahkan antar kategori dalam sebuah observasi. SVM adalah salah satu dari metode yang dikembangkan untuk mengatasi permasalahan yang tidak bisa diselesaikan dengan metode statistika klasik, terutama pada kasus klasifikasi dan prediksi. Dalam SVM, jarak atau *margin* antar kategori dimana nilai dan targetnya sama. Tujuan dari metode ini adalah membangun pemisah optimum yang disebut *Optimal Separating Hyperplane* (OSH) sehingga dapat digunakan untuk klasifikasi. Metode SVM mampu menyelesaikan permasalahan *linear* maupun *non linear*. Pada kasus data yang dapat dipisahkan secara linier, konsep dari SVM adalah dengan menemukan *hyperplane* yang optimum pada *input space*. Fungsi dari *hyperplane* tersebut adalah sebagai pemisah dua buah kelas pada *input space* yang sering disimbolkan dengan -1 dan +1 (Permatasari, 2021). Ilustrasi SVM pada data yang terpisah secara linear dapat dilihat pada Gambar 2.3



Gambar 2.3 Ilustrasi *linearly separable case*

(Sumber : https://www.reneshbedre.com/assets/posts/svm/svm_linear.webp)

Pada Gambar 2.3, kedua kelas data dipisahkan oleh sepasang bidang pembatas yang sejajar atau linear. Data yang berada pada bidang pembatas disebut *support vector*. Persamaan *hyperplane* dapat ditulis sebagai berikut.

$$w^T x + b = 0 \quad (2.17)$$

Dimana w adalah vektor bobot (*wight vector*), x adalah data *training* dan b adalah bias atau konstanta. Dengan adanya kelas positif (+1) dan negative (-1), maka diperoleh Persamaan 2.18 dan 2.19

$$w \cdot x_1 + b \geq +1 \quad (2.18)$$

$$w \cdot x_1 + b \leq -1 \quad (2.19)$$

dengan nilai margin antara bidang pembatas adalah $\frac{2}{\|w\|}$

Nilai margin ini dimaksimalkan dengan tetap memenuhi persamaan 2.18 dan 2.19. Karena memaksimalkan nilai margin sama dengan meminimumkan $\|w\|^2$. Jika kedua bidang pembatas direpresentasikan dalam pertidaksamaan, maka akan menjadi sebagai berikut.

$$y_i(w \cdot x_i + b) - 1 \geq 0 \quad (2.20)$$

Maka pencarian bidang pemisah terbaik dengan nilai margin terbesar dapat dirumuskan sebagai berikut.

$$\min \frac{1}{2} \|w\|^2 \quad (2.21)$$

Dengan $y_i(w \cdot x_i + b) - 1 \geq 0$ dan fungsi batasan sebagai berikut.

$$\sum_{m=1}^M \alpha_m [y_m (W \cdot X_m + b) - 1] \quad (2.22)$$

Kemudian menggunakan *Lagrange Multiplier* didapatkan persamaan berikut.

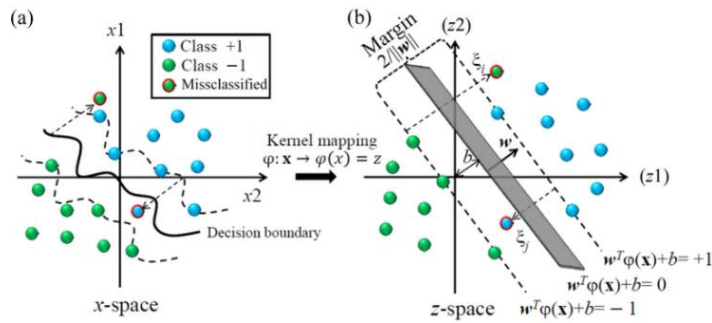
$$L_d = \sum_{m=1}^M \alpha_m - \frac{1}{2} \sum_{m_1=1}^M \sum_{m_2=1}^M \alpha_{m_1} \alpha_{m_2} y_{m_1} y_{m_2} (X_{m_1}^T X_{m_2}) \quad (2.23)$$

Persamaan L_d didapat dengan mensubstitusikan nilai $y_{m_1}, y_{m_2}, X_{m_1}$ dan X_{m_2} ke Persamaan 2.23. Persamaan 2.23 digunakan untuk mencari nilai α_m dengan mengoptimumkan Persamaan 2.23 dengan mencari turunan parsial L_d terhadap α . Setelah mendapat nilai α , selanjutnya adalah mencari nilai w dan b dengan Persamaan 2.24 dan 2.25.

$$w = \sum_{i=1} \alpha_i y_i x_i \quad (2.24)$$

$$b = 1 - w^T x \quad (2.25)$$

Pada kasus riil, sangat jarang ditemukan masalah yang bersifat *linear separable*. Oleh karena itu, SVM membutuhkan fungsi yang mampu membuat pemisah yang tidak linier. Fungsi yang sering digunakan untuk mengatasi hal tersebut adalah dikenal dengan *Kernel Trick*.



Gambar 2.4 Ilustrasi *Non-Linear Case*

(Sumber : www.researchgate.net/figure/Graphical-presentation-of-the-support-vector-machine-classifier-with-a-non-linear-kernel_fig1_299529384)

Dimana kegunaan *kernel trick* adalah untuk menghitung *scalar product* yang ditunjukkan oleh fungsi kernel sebagai berikut.

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \quad (2.26)$$

Dengan persamaan *Lagrange Multiplier* adalah sebagai berikut.

$$L_d = \sum_{m=1}^M \alpha_m - \frac{1}{2} \sum_{m_1=1}^M \sum_{m_2=1}^M \alpha_{m_1} \alpha_{m_2} y_{m_1} y_{m_2} K(X_{m_1} X_{m_2}) \quad (2.27)$$

Ada beberapa fungsi kernel yang sering digunakan dalam literatur SVM antara lain sebagai berikut (Karatzoglou, Smola, Hornik, & Zeileis, 2004)

1. Kernel Linear, merupakan kernel yang paling sederhana dari semua fungsi kernel. Biasa digunakan dalam kasus klasifikasi teks,
2. Kernel *Gaussian* RBF, merupakan kernel yang umum digunakan untuk data yang sudah *valid* dan merupakan default dalam SVM,
3. Kernel Polynomial, merupakan kernel yang sering digunakan untuk klasifikasi gambar,
4. Kernel Sigmoid (*Tangent Hyperbolic*), merupakan kernel yang sering digunakan untuk *neural network*.

Fungsi kernel yang umum digunakan adalah sebagai berikut (Permatasari, 2021)

Tabel 2.9 Persamaan Fungsi Kernel

Fungsi Kernel	Rumus K (x_{m1}, x_{m2})	Parameter
Linier	$x_i^T x_j$	C
RBF	$\exp(-\gamma \ x_i - x_j\ ^2), \gamma > 0$	γ dan C

Fungsi kernel *Gaussian* RBF memiliki kelebihan yaitu secara otomatis menentukan nilai, lokasi dari *center* serta nilai pembobot dan dapat mencakup nilai rentang tak terhingga. *Gaussian* RBF juga efektif dalam menghindari *overfitting* dengan memilih nilai yang tepat untuk parameter C dan γ . Fungsi kernel yang direkomendasikan adalah fungsi kernel RBF karena dapat memetakan hubungan tidak linear dan RBF lebih baik terhadap *outlier* karena kernel RBF berada diantara selang $(-\infty, +\infty)$ sedangkan fungsi kernel yang lain memiliki rentang -1 sampai $+1$ (Scholkopf & Smola, 2002).

2.14 Klasifikasi

Klasifikasi adalah melakukan penilaian terhadap suatu objek data untuk masuk ke dalam suatu kelas tertentu dari sejumlah kelas yang tersedia (Eko Prasetyo, 2013). Hasil klasifikasi

dan prediksi dapat dievaluasi menggunakan pengukuran ketepatan klasifikasi. Semakin tinggi akurasi klasifikasi maka performansi teknik klasifikasi juga semakin baik. Dalam mengukur ketepatan klasifikasi, diperlukan suatu *tools* yang disebut *confusion matrix*. *Confusion matrix* merupakan salah satu *tools* penting dalam metode visualisasi yang digunakan pada mesin pembelajaran yang biasanya memuat dua kategori atau lebih (Han, Kamber, & Pei, 2012). Contoh hasil *confusion matrix* dapat dilihat pada Tabel 2.10

Tabel 2.10 *Confusion Matrix*

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Keterangan Tabel 2.10 :

TP : *True Positive* (jumlah prediksi benar pada kelas positif)

FP : *False Positive* (jumlah prediksi salah pada kelas positif)

FN : *False Negative* (jumlah prediksi salah pada kelas negatif)

TN : *True Negative* (jumlah prediksi benar pada kelas negatif)

Pengukuran yang sering digunakan untuk menghitung ketepatan klasifikasi adalah *accuracy*, *sensitivity*, dan *specivicity* (Hotho dkk, 2005).

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (2.21)$$

Sensitivity adalah tingkat positif benar atau ukuran performansi untuk mengukur kelas positif.

$$Sensitivity = \frac{TP}{TP + FN} \quad (2.22)$$

Specivicity adalah tingkat negatif benar atau ukuran performansi untuk mengukur kelas negatif.

$$Specivicity = \frac{TN}{TN + FP} \quad (2.23)$$

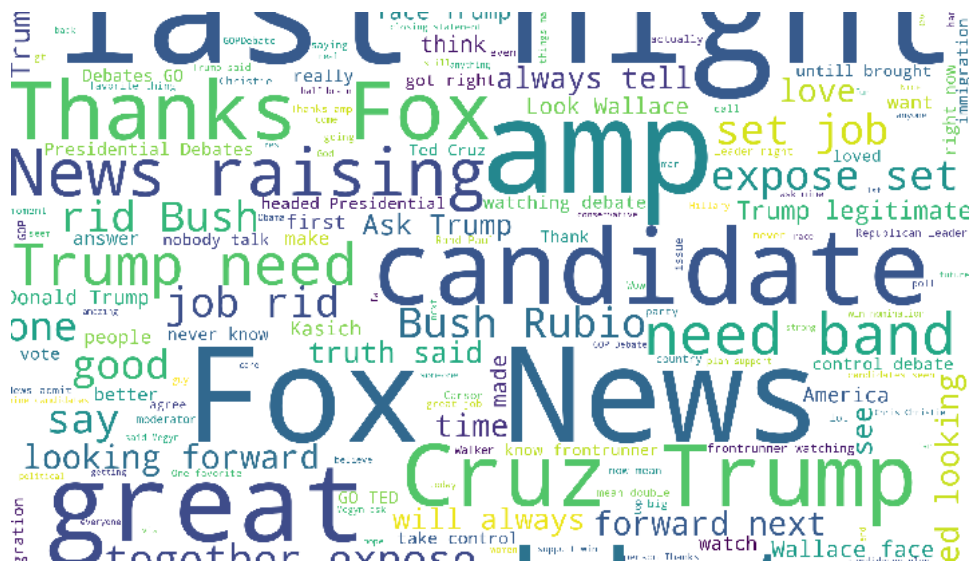
Tingkat akurasi dari suatu model klasifikasi dapat ditentukan berdasarkan kriteria sebagai berikut (Gorunescu, 2011)

1. 90% - 100% : sempurna (*excellent classification*),
2. 80% - 90% : baik (*good classification*),
3. 70% - 80% : adil (*fair classification*),
4. 60% - 70% : buruk (*poor classification*),
5. 50% - 60% : gagal (*failure*)

2.15 Word Cloud

Word cloud merupakan sebuah sistem yang memunculkan susunan kata sebagai citra visual terkait frekuensi kemunculan kata dalam suatu teks verbal. Menurut McNaught dan Lam (2010) visualisasi *word cloud* dari teks akan memudahkan pengamat dalam melihat gagasan sehingga dapat menjadi alat bantu dalam melakukan analisis terhadap sebuah wacana tertulis. Menurut Castella & Sutton (2014) *word cloud* bertujuan untuk menampilkan grafis dari sebuah

dokumen yang dilakukan dengan *plotting* kata-kata yang sering muncul pada ruang dua dimensi.



Gambar 2.5 Visualisasi Word Cloud

(Sumber : <https://www.kaggle.com/code/ngyptr/python-nltk-sentiment-analysis>)

Word cloud menampilkan kata-kata populer atau sering muncul dari sebuah dokumen. Frekuensi dari kata yang sering muncul ditunjukkan melalui ukuran huruf kata tersebut. Semakin besar ukuran kata, maka semakin besar pula frekuensi kata tersebut muncul dalam dokumen (Castella & Sutton, 2014).

BAB III METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder berupa kumpulan *tweet* mengenai ulasan atau tanggapan pengguna *Twitter* mengenai kinerja BPJS. Data tersebut berupa *tweets* berbahasa Indonesia dengan kata kunci “BPJS”, “Badan Penyelenggara Jaminan Sosial”, “Klaim BPJS” periode Januari 2019 sampai Desember 2021 yang diambil menggunakan teknik *scrapping* dengan bahasa pemrograman *Python 2.7.6*. Berikut merupakan data *tweet* yang diperoleh dari proses *scrapping*.

Tabel 3.1 Sumber Data

Tanggal	Text
30-12-2021	Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan.#disnakerlampung #KPU #KPUMelayani #bpjskesehatan #lampostco https://t.co/2nUaexGu5i
29-12-2021	@urhjsm kita adalah anggota paguyuban BPJS
23-12-2021	Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan mendapatkan 3 penghargaan dalam Penghargaan Top Digital Awards 2021 yang diselenggarakan oleh Majalah ItWorks.
⋮	⋮
03-01-2019	@Erlina84829866 Salam Sehat Ibu Erlina. Benar, sehubungan dengan diterbitkan Peraturan Badan Penyelenggara Jaminan Sosial Kesehatan Nomor 6 tahun 2018 tentang Administrasi Kepesertaan dalam Program Jaminan Kesehatan, yang mulai berlaku ditanggal 18 Desember 2018
02-01-2019	Masyarakat yang tidak memiliki kartu anggota Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan, terpaksa mulai Januari tahun 2019, tidak akan mendapat layanan subsidi https://t.co/j19epZhq0Z
02-01-2019	Kalangan serikat pekerja meminta Presiden Joko Widodo segera bertindak mengganti anggota Dewan Pengawas Badan Penyelenggara Jaminan Sosial Ketenagakerjaan yang diduga melakukan pelecehan seksual terhadap stafnya, RA. #Humaniora #AdadiKompas https://t.co/pgpkItmlIj

3.2 Variabel Penelitian

Setelah dilakukan *preprocessing* pada data *tweet* terdiri dari dua variabel yaitu variabel predictor (X) dan variabel respon (Y). Dimana X adalah frekuensi kata dasar setiap *tweet* dan Y adalah klasifikasi sentimen *tweet* positif dan negatif. Berikut merupakan struktur data penelitian untuk analisis sentimen pada penelitian ini.

Tabel 3.2 Struktur Data Penelitian

<i>Tweet</i>	Klasifikasi Sentimen (Y)	Kata Dasar (X_1)	Kata Dasar (X_2)	⋯	Kata Dasar (X_n)
1	Negatif	$X_{1,1}$	$X_{2,1}$		$X_{n,1}$
2	Negatif	$X_{1,2}$	$X_{2,2}$		$X_{n,2}$
⋮	⋮	⋮	⋮	⋮	⋮
2662	Positif	$X_{1,2662}$	$X_{2,2662}$		$X_{n,2662}$

3.3 Metode Analisis

Langkah analisis yang digunakan dalam penelitian ini dengan metode *Naïve Bayes Classifier* dan *Support Vector Machine* adalah sebagai berikut

3.3.1 Studi Literatur

Pada tahap ini dilakukan studi literatur dengan pencarian Pustaka dan referensi yang bersumber dari artikel, jurnal, maupun *e-book* terkait Analisis Sentimen menggunakan *Naïve Bayes Classifier* dan *Support Vector Machine*.

3.3.2 Pengumpulan Data

Tahapan selanjutnya adalah pengumpulan data *tweet* yang mengandung beberapa kata kunci, yaitu ‘bpjs’, ‘badan’, ‘penyelenggara’, ‘jaminan’, ‘sosial’, ‘klaim’, ‘kesehatan’, dan ‘ketenagakerjaan’ dengan periode Januari 2019 sampai Desember 2020.

3.3.3 Pre-processing Data

Tahap selanjutnya adalah melakukan *pre-processing* data menggunakan *Python 2.7.6*. Ada beberapa tahapan yang harus dilakukan pada proses ini seperti.

1. *Cleaning*, proses ini adalah proses membersihkan *tweet* dari *emoticons*, *hashtag*, *duplicate*, *username*, simbol, angka, tanda baca, dan *url*.
2. *Case folding*, proses merubah format *tweet* menjadi seragam. Dalam hal ini menjadi format *lowercase*,
3. *Stopwords removal*, proses menghapus kata-kata yang sering muncul tetapi tidak mempengaruhi makna. Kamus yang digunakan adalah kamus bawaan *Python* yaitu Sastrawi,
4. *Stemming*, adalah proses mengubah kata menjadi kata dasar dengan menghilangkan awalan, akhiran, dan sisipan. Kamus yang digunakan adalah kamus bawaan *Python* yaitu Sastrawi,
5. *Tokenization*, proses memecah kalimat menjadi per kata.

3.3.4 Pelabelan Kelas Sentimen

Setelah data *tweet* menjadi per kata, tahap selanjutnya adalah memberi label pada *tweet* secara manual dengan melihat banyaknya kata positif dan negatif yang terdapat dalam suatu *tweet*. Menurut KBBI yang dimaksud dengan kata positif adalah kata yang bersifat nyata dan membangun. Dan kata negatif adalah kata yang bersifat buruk atau tidak diinginkan. Kata negatif biasanya mengandung kata ‘tidak’, ‘bukan’, ‘tapi’. Dalam penelitian ini sentimen akan dibagi menjadi dua kategori yaitu kategori sentimen positif dan negatif. Suatu *tweet* dikatakan bersentimen positif apabila jumlah kata positif lebih banyak dibanding jumlah kata negatif. Untuk kategori sentimen positif nantinya akan diberi label +1 dan untuk kategori kelas negatif akan diberi label -1.

3.3.5 Oversampling Data

Setelah melabelkan *tweet* menjadi kelas positif dan negatif akan diketahui proporsi dari masing-masing kelas sentimen. Apabila salah satu kelas sentimen memiliki proporsi yang kurang dari 35% maka perlu dilakukan *oversampling* data karena data tersebut cenderung *imbalance* atau tidak seimbang. Hal tersebut perlu dilakukan untuk menghindari bias pada hasil ketepatan klasifikasi.

3.3.6 Term Frequency-Inverse Document Frequency

Tahap selanjutnya adalah membobot setiap kata pada *tweet* dengan menghitung frekuensi kemunculan kata di tiap *tweet* dengan menggunakan Persamaan 2.1. Lalu menghitung jumlah frekuensi kemunculan kata yang sama pada tiap *tweet* dengan Persamaan 2.2. Setelah

kedua tahap tersebut, barulah menghitung bobot untuk tiap kata dalam *tweet* dengan Persamaan 2.3.

3.3.7 Pembagian Data *Testing* dan *Training*

Setelah melakukan pembobotan kata tahap selanjutnya adalah membagi data *testing* dan *training* menggunakan teknik *stratified 10-fold cross validation*. Dengan teknik ini, data akan dibagi menjadi 10% data *testing* dan 90% data *training* kedalam 10 *fold* yang berbeda beda di tiap *fold* nya dengan mengambil data secara acak pada 2662 data.

3.3.8 Klasifikasi *Naïve Bayes Classifier*

Pada tahap ini dilakukan analisis klasifikasi *Naïve Bayes Classifier* menggunakan data *testing* dan *training* dengan langkah-langkah sebagai berikut.

1. Menghitung probabilitas C_j atau prior menggunakan Persamaan 2.13,
2. Menghitung probabilitas X_n pada kelas C_j menggunakan Persamaan 2.16 dan 2.8,
3. Menghitung probabilitas posterior menggunakan Persamaan 2.9,
4. Menentukan kelas terbaik dengan mencari nilai *maximum a posterior* (MAP) menggunakan Persamaan 2.12,
5. Menghitung ketepatan klasifikasi dari model dengan Persamaan 2.21.

3.3.9 Klasifikasi *Support Vector Machine*

Setelah melakukan klasifikasi dengan *Naïve Bayes Classifier*, tahap selanjutnya adalah melakukan klasifikasi *Support Vector Machine* menggunakan data *training* dan data *testing* yang sama dengan langkah-langkah sebagai berikut.

1. Melakukan analisis klasifikasi menggunakan *Support Vector Machine* kernel Linear, dan *Radial Basis Function* (RBF),
2. Menentukan nilai parameter C dan γ untuk masing-masing kernel dengan rentang nilai parameter $C = (0,001; 0,01; 0,1; 1; 10; 100; 1000)$ dan nilai $\gamma = (0,001; 0,01; 0,1; 1; 10; 100; 1000)$. Nilai tersebut dipilih berdasarkan pada percobaan *trial and error*,
3. Melakukan perbandingan dan pemilihan yang terbaik dari hasil klasifikasi masing-masing kernel dan parameter,
4. Melakukan analisis hasil model klasifikasi pada metode *Support Vector Machine*,
5. Melakukan perhitungan ketepatan klasifikasi untuk masing-masing kernel.

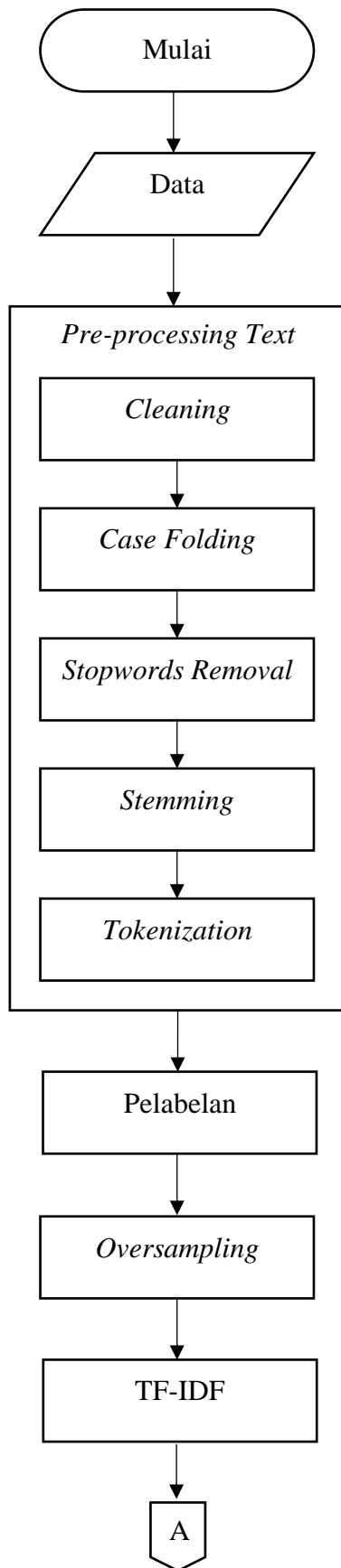
3.3.10 Perbandingan Klasifikasi *Naïve Bayes Classifier* dan *Support Vector Machine*

Setelah dilakukan analisis klasifikasi dengan dua metode diatas, tahap selanjutnya adalah membandingkan hasil dari kedua metode tadi. Dalam penelitian ini aspek yang akan digunakan untuk perbandingan adalah nilai rata-rata ketepatan klasifikasi, karena nilai tersebut merupakan nilai tengah dari 10 subset dan nilai tersebut cenderung mewakili nilai dari kesepuluh subset.

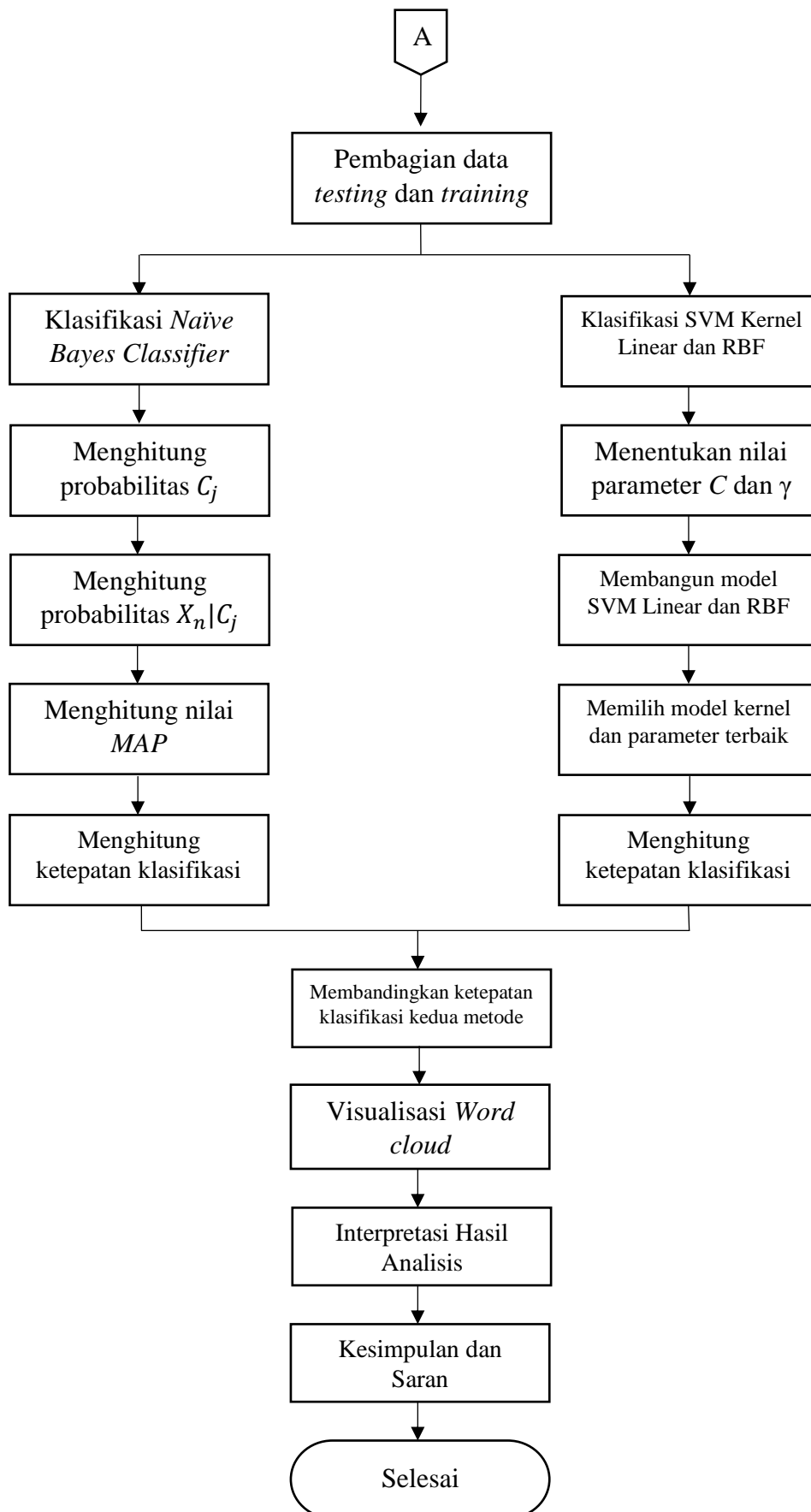
3.3.11 Kesimpulan dan Saran

Setelah diketahui perbandingan diantara *Naïve Bayes Classifier* dan *Support Vector Machine* maka dari hasil tersebut dapat ditarik kesimpulan sebanyak rumusan masalah dan beberapa saran untuk mengembangkan penelitian lebih lanjut.

Berdasarkan tahapan analisis diatas, dapat digambarkan dengan diagram alir sebagai berikut.



Gambar 3.1 Diagram Alir Metode Analisis



Gambar 3.1 Diagram Alir Metode Analisis (Lanjutan)

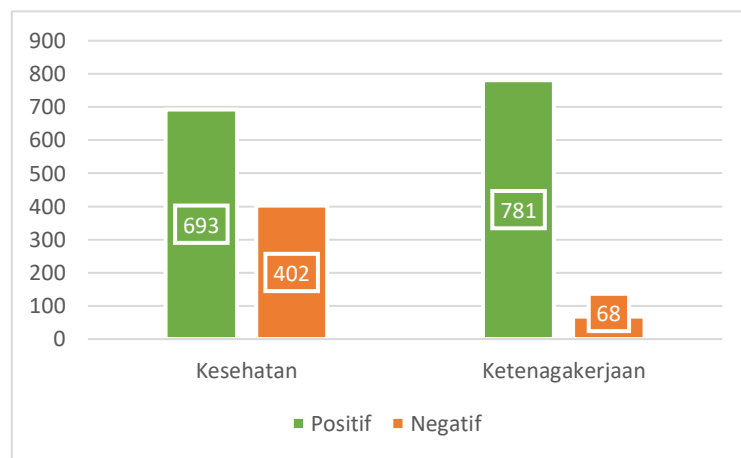
(Halaman ini sengaja dikosongkan)

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini akan dibahas hasil analisis berdasarkan pengolahan data yang telah dilakukan menggunakan metode analisis klasifikasi *Naïve Bayes Classifier* dan *Support Vector Machine* menggunakan bahasa pemrograman Python.

4.1 Karakteristik Data

Dari data *tweet* dengan kata kunci ‘bpjs’, ‘badan’, ‘klaim’, ‘penyelenggara’, ‘jaminan’, ‘sosial’ periode Januari 2019 sampai Desember 2021 didapatkan bahwa ulasan BPJS Kesehatan lebih banyak daripada BPJS Ketenagakerjaan. Hal ini dapat dilihat dari gambar 4.1 yang menunjukkan frekuensi ulasan tiap kelas sentimen pada BPJS Kesehatan maupun BPJS Ketenagakerjaan.



Gambar 4.1 Grafik Ulasan BPJS Kesehatan dan BPJS Ketenagakerjaan

Dari Gambar 4.1 dapat dilihat bahwa jumlah seluruh ulasan BPJS Kesehatan lebih banyak dibanding BPJS Ketenagakerjaan, hal ini dikarenakan jumlah peserta BPJS Kesehatan pada November 2021 tercatat sebanyak 229,51 juta orang. Jumlah tersebut mengalami kenaikan sebesar 3,16% dibanding tahun sebelumnya yang mencapai 222,46 juta orang. Jika dibandingkan dengan populasi Indonesia, maka 83,89% penduduk telah menjadi peserta BPJS. Sedangkan untuk BPJS Ketenagakerjaan tercatat memiliki 30,66 juta peserta aktif hingga kuartal IV tahun 2021. Jumlah tersebut mengalami kenaikan sebesar 2,27% dibandingkan dengan periode yang sama tahun lalu. Jumlah ulasan yang diperoleh BPJS Kesehatan adalah sebanyak 1095 ulasan dari total 2662 ulasan. Dengan ulasan positif sebanyak 693 ulasan dari total 2662 ulasan atau sebanyak 36% dari seluruh jumlah ulasan, dan ulasan negatif sebanyak 402 ulasan dari total 2662 ulasan atau sebanyak 21% dari seluruh jumlah ulasan. Jumlah ulasan yang diperoleh BPJS Ketenagakerjaan adalah sebanyak 849 ulasan dari total 2662 ulasan. Dengan ulasan positif sebanyak 781 ulasan dari total 2662 ulasan atau sebanyak 40% dari seluruh jumlah ulasan, dan untuk ulasan negatif sebanyak 68 ulasan dari total 2662 ulasan atau sebanyak 3% dari seluruh jumlah ulasan. Untuk ulasan positif, baik BPJS Kesehatan maupun BPJS Ketenagakerjaan memiliki jumlah ulasan positif yang lebih banyak dibandingkan ulasan negatif. Hal ini berarti kinerja dari kedua BPJS ini sama-sama baik dimata pengguna *twitter*. Tetapi ulasan positif BPJS Ketenagakerjaan lebih tinggi dibanding BPJS Kesehatan, hal ini dapat terjadi akibat adanya pandemi covid-19 yang melanda Indonesia dan ditunjukknya BPJS Kesehatan menjadi badan yang bertugas memverifikasi klaim covid-19.

4.2 Preprocessing Text

Data *tweet* mengenai BPJS pada media sosial *twitter* yang telah dikumpulkan kemudian dilakukan *preprocessing text* yang meliputi *cleaning*, *case folding*, *labeling*, *stemming*, *stopword removal*, dan *tokenizing*. Berikut beberapa *tweet* sebelum dilakukan *preprocessing text*.

Tabel 4.1 Data Tweet Sebelum *Preprocessing Text*

<i>Tweet</i>
Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan.#disnakerlampung #KPU #KPUMelayani #bpjskesehatan #lampostco https://t.co/2nUaexGu5i Pemerintah menetapkan akan memberikan suntikan dosis ketiga atau vaksin booster gratis bagi lansia dan peserta BPJS (Badan Penyelenggara Jaminan Sosial) Kesehatan mulai Januari 2022. https://t.co/h8zefEiYd2 https://t.co/dvNVZnouq9 Premi BPJS Kesehatan Bakal Ditentukan sesuai Jumlah Harta Kekayaan https://t.co/toea1sk714 Iuran Naik, Ribuan Peserta BPJS Turun Kelas https://t.co/7xELoRQLFw

Dapat dilihat dari Tabel 4.1 masih terdapat url, angka, tanda baca, simbol, serta format yang masih belum seragam pada *tweet*. Hal ini harus dihilangkan untuk dapat dilanjutkan ke tahap yang lebih jauh. Tahap pertama yang harus dilakukan adalah menghilangkan semua tanda baca, angka, simbol dan url. Berikut merupakan langkah-langkah yang harus dilakukan di *preprocessing text*.

4.2.1 Cleaning

Tahap *preprocessing text* yang pertama dilakukan adalah *cleaning* data. *Cleaning* data merupakan proses menghilangkan *HTML*, *emoticons*, *hashtag*, *username*, *duplicate*, angka, simbol, tanda baca, dan *url*. Berikut merupakan hasil *tweet* sebelum dan setelah dilakukannya *cleaning* data.

Tabel 4.2 Data *Tweet* Sebelum dan Sesudah *Cleaning*

Sebelum	Sesudah
Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan.#disnakerlampung #KPU #KPUMelayani #bpjskesehatan #lampostco https://t.co/2nUaexGu5i Pemerintah menetapkan akan memberikan suntikan dosis ketiga gratis bagi lansia dan peserta BPJS Kesehatan mulai Januari 2022. https://t.co/h8zefEiYd2 https://t.co/dvNVZnouq9	Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan Pemerintah menetapkan akan memberikan suntikan dosis ketiga gratis bagi lansia dan peserta BPJS Kesehatan mulai Januari
Premi BPJS Kesehatan Bakal Ditentukan sesuai Jumlah Harta Kekayaan https://t.co/toea1sk714	Premi BPJS Kesehatan Bakal Ditentukan sesuai Jumlah Harta Kekayaan
Iuran Naik, Ribuan Peserta BPJS Turun Kelas https://t.co/7xELoRQLFw	Iuran Naik Ribuan Peserta BPJS Turun Kelas

Dapat dilihat pada Tabel 4.2 sebelah kiri *tweet* sebelum dilakukan *cleaning* masih terdapat url, angka, tanda baca, dan simbol. Dan pada tabel sebelah kanan, setelah dilakukan *cleaning*, url, angka, tanda baca, dan simbol sudah tidak ada lagi.

4.2.2 Case Folding

Setelah *cleaning* data, tahap selanjutnya adalah *case folding* yang bertujuan untuk mengubah keseluruhan teks kedalam format yang sama yaitu *lowercase* karena dalam suatu *tweet* atau dokumen tidak selalu terstruktur dan konsisten dalam penggunaan huruf kapital. Berikut merupakan hasil *tweet* sebelum dan sesudah dilakukannya *case folding*.

Tabel 4.3 Data *Tweet* Sebelum dan Sesudah *Case Folding*

Sebelum	Sesudah
Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan	disnaker lampung mengusulkan petugas komisi pemilihan tingkat desa provinsi dilindungi bpjs ketenagakerjaan
Pemerintah menetapkan akan memberikan suntikan dosis ketiga gratis bagi lansia dan peserta BPJS Kesehatan mulai Januari Premi BPJS Kesehatan Bakal Ditentukan sesuai Jumlah Harta Kekayaan	pemerintah menetapkan akan memberikan suntikan dosis ketiga gratis bagi lansia dan peserta bpjs kesehatan mulai januari premi bpjs kesehatan bakal ditentukan sesuai jumlah harta kekayaan
Iuran Naik Ribuan Peserta BPJS Turun Kelas	iuran naik ribuan peserta bpjs turun kelas

4.2.3 Stopwords Removal

Setelah dilakukan *case folding* proses selanjutnya adalah *stopword removal* dengan menggunakan kamus Sastrawi yang tersedia di *Python* dengan menambahkan beberapa kata seperti 'bpjs', 'badan', 'penyelenggara', 'jaminan', 'sosial', dan 'klaim'. *Stopwords removal* adalah proses menghapus kata-kata umum dan sering muncul tetapi tidak memiliki pengaruh yang signifikan terhadap makna dari sebuah kalimat seperti kata 'dan', 'itu', 'atau', 'yang'. Berikut merupakan *tweet* sebelum dan sesudah dilakukannya *stopwords removal*.

Tabel 4.4 Data *Tweet* Sebelum dan Sesudah *Stopwords*

Sebelum	Sesudah
disnaker lampung mengusulkan petugas komisi pemilihan tingkat desa provinsi dilindungi bpjs ketenagakerjaan	disnaker lampung mengusulkan petugas komisi pemilihan tingkat desa provinsi dilindungi
pemerintah menetapkan akan memberikan suntikan dosis ketiga gratis bagi lansia dan peserta bpjs kesehatan mulai januari premi bpjs kesehatan bakal ditentukan sesuai jumlah harta kekayaan	pemerintah menetapkan suntikan dosis ketiga gratis lansia peserta januari premi ditentukan sesuai harta kekayaan
iuran naik ribuan peserta bpjs turun kelas	iuran naik ribuan peserta turun kelas

Dapat dilihat dari Tabel 4.4 sebelah kanan, kata-kata seperti bpjs, kesehatan, ketenagakerjaan, akan, mulai, bagi, bakal, dan jumlah telah dihapus dalam proses ini. Sehingga hasil dari *stopwords removal* adalah *tweet* yang berisi kata-kata yang lebih bermakna. Dan hal tersebut dapat berpengaruh pada pembobotan kata TF-IDF.

4.2.4 Stemming

Setelah melewati proses *stopwords removal*, proses selanjutnya adalah *stemming*. Dimana *tweet* diubah menjadi kata dasar dengan menghilangkan awalan, akhiran, sisipan, dan *confixes* (awalan dan akhiran) dengan menggunakan kamus bawaan *Python* yaitu Sastrawi. Berikut merupakan hasil *tweet* sebelum dan setelah dilakukannya *stemming*.

Tabel 4.5 Data *Tweet* Sebelum dan Sesudah *Stemming*

Sebelum	Sesudah
disnaker lampung mengusulkan petugas komisi pemilihan tingkat desa provinsi dilindungi pemerintah menetapkan suntikan dosis ketiga gratis lansia peserta januari	disnaker lampung usul tugas komisi pilih tingkat desa provinsi lindung pemerintah tetap suntik dosis tiga gratis lansia peserta januari
premi ditentukan sesuai harta kekayaan	premi tentu sesuai harta kaya
iuran naik ribuan peserta turun kelas	iur naik ribu peserta turun kelas

Setelah dilakukan proses *stemming* dapat dilihat bahwa kata-kata yang menggunakan awalan dan akhiran seperti ‘mengusulkan’, ‘menetapkan’, ‘ditentukan’, ‘kekayaan’, ‘pemilihan’ berubah menjadi kata dasar dengan menghilangkan awalan dan akhiran.

4.2.5 Tokenization

Tahap terakhir dalam *prerprocessing text* adalah *tokenization*. Dimana pada tahap ini kalimat akan dipecah menjadi kata per kata untuk memudahkan proses analisis. Berikut merupakan hasil *tweet* sebelum dan setelah dilakukannya proses *tokenization*.

Tabel 4.6 Data *Tweet* Sebelum dan Sesudah *Tokenization*

Sebelum	Sesudah
disnaker lampung usul tugas komisi pilih tingkat desa provinsi lindung bpjs ketenagakerjaan pemerintah tetap suntik dosis tiga gratis lansia peserta januari	['disnaker', 'lampung', 'usul', 'tugas', 'komisi', 'pilih', 'tingkat', 'desa', 'provinsi', 'lindung']
premi tentu sesuai harta kaya	['premi', 'tentu', 'sesuai', 'harta', 'kaya']
iur naik ribu peserta turun kelas	['iur', 'naik', 'ribu', 'peserta', 'turun', 'kelas']

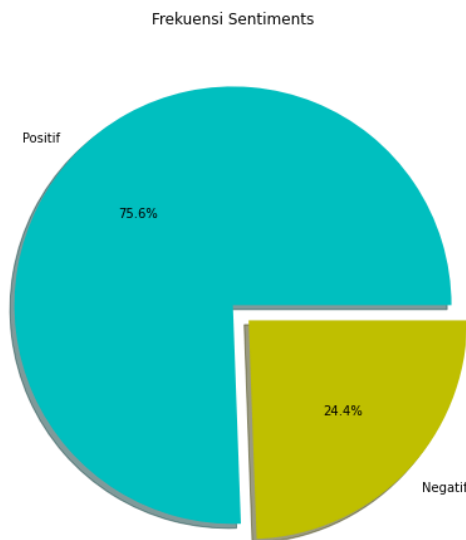
4.3 Pelabelan

Setelah melakukan *preprocessing text* tahap selanjutnya adalah melabel data tweet menjadi kelas positif dan negatif. Pelabelan ini dilakukan secara manual dengan melihat banyak kata negatif atau positif dalam satu *tweet*. Menurut KBBI yang dimaksud dengan kata positif adalah kata yang bersifat nyata dan membangun. Dan kata negatif adalah kata yang bersifat buruk atau tidak diinginkan. Kata negatif biasanya mengandung kata ‘tidak’, ‘bukan’, ‘tapi. Suatu *tweet* dikatakan bersentimen positif apabila jumlah kata positif lebih banyak dibanding jumlah kata negatif. Begitu juga sebaliknya, apabila dalam satu *tweet* memiliki jumlah kata negatif lebih banyak maka *tweet* tersebut dikatakan bersentimen negatif. Nantinya *tweet* yang masuk kedalam kelas positif akan diberi nilai 1 dan kelas negatif akan diberi nilai -1. Berikut merupakan hasil *tweet* setelah diberi label.

Tabel 4.7 Data *tweet* Setelah Pelabelan

<i>Tweet</i>	Sentimen	Label
'disnaker', 'lampung', 'usul', 'tugas', 'komisi', 'pilih', 'tingkat', 'desa', 'provinsi', 'lindung'	Positif	1
'pemerintah', 'tetap', 'suntik', 'dosis', 'tiga', 'gratis', 'lansia', 'peserta', 'januari'	Positif	1
'premi', 'tentu', 'sesuai', 'harta', 'kaya'	Positif	1
'iur', 'naik', 'ribu', 'peserta', 'turun', 'kelas'	Negatif	-1

Pada contoh pelabelan yang terakhir banyak kata positif dan negatif sama besar, tetapi apabila diteliti *tweet* tersebut bermakna negatif dimana akibat dari iuran yang naik banyak peserta memilih untuk turun kelas, itulah mengapa *tweet* tersebut diberi label negatif atau -1. Dari 2662 data *twitter* mengenai BPJS, diperoleh data dengan kategori sentimen negatif sebanyak 650 dan data dengan kategori sentimen positif sebanyak 2012 data. Data kategori sentimen tersebut akan disajikan dalam bentuk diagram atau grafik untuk mengetahui perbandingan jumlah antar kategori sentimen.

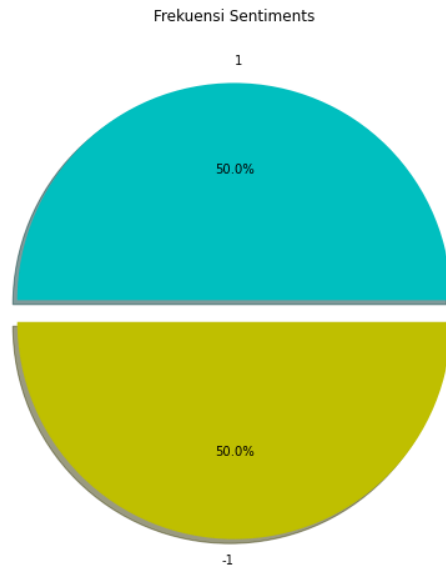


Gambar 4.2 Pie chart Sentimen Positif dan Negatif

Dapat dilihat dari Gambar 4.2 sebanyak 75,6% merupakan kategori sentimen positif dan sebanyak 24,4% merupakan sentimen negatif. Karena proporsi kategori kelas negatif kurang dari 35%, maka hal ini menunjukkan bahwa data cenderung *imbalance*. Keadaan ini akan berpengaruh pada perhitungan ketepatan klasifikasi dan menghasilkan hasil yang bias. Maka untuk melakukan pemodelan klasifikasi diperlukan adanya metode untuk mengatasi keadaan *imbalance* tersebut.

4.4 Oversampling

Karena setelah dilakukannya pelabelan diketahui bahwa kelas kategori negatif kurang dari 35%, sehingga cenderung *imbalance* hal tersebut dapat berakibat pada hasil ketepatan klasifikasi yang bias, maka untuk mengatasi hal tersebut diperlukan metode *oversampling*, dengan menambahkan jumlah data pada kelas minoritas dimana dalam penelitian ini adalah kelas kategori negatif sehingga diharapkan rasio antar kelas dapat lebih seimbang. Berikut merupakan hasil setelah dilakukannya *oversampling*.



Gambar 4.3 Pie chart Sentimen Positif dan Negatif Setelah *Oversampling*

Dapat dilihat dari Gambar 4.3 bahwa proporsi kelas positif dan negatif setelah dilakukan *oversampling* menjadi sama besar. Hal ini sejalan dengan konsep algoritma dari *oversampling* dimana data kelas minoritas dipilih secara acak kemudian ditambahkan ke dalam data *training*. Proses pemilihan dan penambahan ini diulang-ulang sampai jumlah data kelas minoritas sama dengan jumlah kelas mayoritas. Dan karena hal tersebut terjadi penambahan data yang semula hanya 2662 data, setelah dilakukannya *oversampling* data menjadi 4024 data.

4.5 Term Frequency-Inverse Document Frequency (TF-IDF)

Setelah dilakukan *preprocessing text* langkah selanjutnya adalah melakukan pembobotan kata (*term*). Dalam proses ini, setiap kata yang telah di *preprocessing* akan dipecah terlebih dahulu dan disimpan dalam *database* kemudian dihitung jumlah kemunculan setiap katanya. Pada penelitian ini, pembobotan kata menggunakan metode *term frequency inverse document* (TF-IDF), dimana perhitungan dilakukan dengan menghitung jumlah *term frequency* (TF) dokumen terlebih dahulu dengan Persamaan 2.1 kemudian menghitung nilai jumlah dokumen yang memiliki *term* (DF), lalu menghitung nilai IDF dengan Persamaan 2.2. Setelah nilai TF dan IDF didapatkan maka langkah selanjutnya adalah menentukan bobot kata dengan mengalikan TF dan IDF seperti di Persamaan 2.3. Berikut merupakan contoh kata-kata yang telah dihitung frekuensinya.

Tabel 4.8 Tabel TF

<i>tweet</i>	desa	disnaker	dosis	gratis	harta		tugas	turun	usul
1	1	1	0	0	0		1	0	1
2	0	0	1	1	0	...	0	0	0
3	0	0	0	0	1		0	0	0
4	0	0	0	0	0		0	1	0

Dari Tabel 4.8 diatas, didapatkan frekuensi kemunculan kata dari masing-masing *tweet*. Pada *tweet* pertama kata ‘desa’ muncul sebanyak satu kali. Di *tweet* kedua, ketiga, dan keempat tidak terdapat kemunculan kata desa. Pada *tweet* pertama muncul kata ‘disnaker’ sebanyak satu

kali dan pada *tweet* kedua, ketiga, dan keempat tidak terdapat kata ‘disnaker’. Pada *tweet* kedua muncul kata ‘dosis’ dan ‘gratis’ sebanyak satu kali. Perhitungan frekuensi ini dilakukan sampai *tweet* terakhir yang kemudian dari hasil frekuensi inilah kemudian dapat diketahui besar bobot dari masing-masing kata tersebut. Dengan menggunakan Persamaan 2.3 Berikut merupakan contoh dari kata-kata yang telah dibobot menggunakan TF-IDF.

Tabel 4.9 Tabel TF-IDF

<i>tweet</i>	desa	disnaker	dosis	gratis	harta	tugas	turun	usul
1	0,32	0,44	0,00	0,00	0,00	0,32	0,00	0,30
2	0,00	0,00	0,41	0,33	0,00	...	0,00	0,00
3	0,00	0,00	0,00	0,00	0,57	0,00	0,00	0,00
4	0,00	0,00	0,00	0,00	0,00	0,00	0,55	0,00

Setelah dilakukan perhitungan bobot tiap kata dengan bantuan *Python* didapatkan hasil bobot untuk tiap kata dapat dilihat pada Tabel 4.9. Diketahui bobot untuk kata ‘desa’ pada *tweet* pertama sebesar 0,32 untuk kata ‘disnaker’ sebesar 0,44. Untuk kata ‘dosis’ dan ‘gratis’ pada *tweet* kedua masing-masing sebesar 0,41 dan 0,33. Pembobotan ini dilakukan untuk semua *tweet* dimana nantinya hasil dari pembobotan tersebut akan digunakan untuk analisis klasifikasi.

4.6 Analisis Klasifikasi Menggunakan Naïve Bayes Classifier

Pada analisis klasifikasi penelitian ini, metode yang pertama digunakan adalah *Naïve Bayes Classifier*. Setelah membagi data menjadi data *training* dan *testing* menggunakan *stratified 10-fold cross validation*. Dimana terdapat 10 *subset* yang tiap *subset* nya terdiri dari 10% data *testing* dan 90% data *training* yang berbeda. Pada pengklasifikasi dengan *Naïve Bayes Classifier* akan menghasilkan probabilitas yang dapat digunakan untuk menentukan apakah *tweet* tersebut masuk ke dalam kategori sentimen positif atau negatif. Berikut merupakan beberapa nilai probabilitas yang dihasilkan dari model *Naïve Bayes Classifier*.

Tabel 4.10 Hasil Probabilitas *Naïve Bayes Classifier*

Negatif	Positif	Keputusan
0,605983	0,394017	Negatif
0,513188	0,486812	Negatif
0,196549	0,803451	Positif
	⋮	
0,882198	0,117802	Negatif
0,868786	0,131214	Negatif
0,602315	0,397685	Negatif
0,647539	0,352461	Negatif

Nilai probabilitas pada Tabel 4.10 menunjukkan bahwa *tweet* memiliki peluang untuk masuk ke dalam kategori sentimen sebesar nilai yang terdapat pada kedua kolom. Suatu *tweet* akan masuk ke dalam salah satu kategori dengan nilai probabilitas terbesar. Sehingga apabila suatu *tweet* memiliki nilai probabilitas dengan kategori positif lebih besar dibanding kategori

negatif maka *tweet* tersebut masuk ke dalam kategori positif. Begitu juga sebaliknya, apabila suatu *tweet* memiliki nilai probabilitas kategori negatif lebih besar maka *tweet* tersebut masuk ke dalam kategori negatif. Setelah mendapatkan nilai probabilitas dari setiap *tweet* yang akan digunakan untuk mengklasifikasikan *tweet* tersebut, langkah selanjutnya adalah menghitung ketepatan klasifikasi dari model. Berikut merupakan tabel pengukuran ketepatan klasifikasi dari *Naïve Bayes Classifier*.

Tabel 4.11 Pengukuran Ketepatan Klasifikasi *Naïve Bayes Classifier*

Subset	Accuracy (%)	Sensitivity (%)
1	89,6	89,2
2	87,3	91,7
3	84,6	88,0
4	86,1	94,0
5	87,1	94,6
6	86,6	93,5
7	86,6	92,0
8	88,1	93,7
9	86,8	93,0
10	84,6	90,2
Rata-Rata	86,7	92,0

Dari subset sebanyak 10 kali dimana setiap subsetnya terdiri dari 10% data *testing* dan 90% data *training* yang berbeda-beda di tiap subsetnya didapatkan rata-rata pengukuran ketepatan klasifikasi *accuracy* sebesar 86,7% nilai *accuracy* ini menunjukkan rasio prediksi benar untuk kelas positif dan negatif dengan keseluruhan data. Dan nilai rata-rata *sensitivity* dari model ini sebesar 90% dimana nilai *sensitivity* ini merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Nilai rata-rata ini dinilai dapat mewakili seluruh nilai dari 10 subset tersebut. sMaka dari hasil pengukuran ketepatan klasifikasi tersebut dapat dikatakan bahwa metode *Naïve Bayes Classifier* baik digunakan dalam mengklasifikasikan sentimen data *twitter* mengenai BPJS. Berikut merupakan *confusion matrix* untuk rata-rata ketepatan klasifikasi.

Tabel 4.12 *Confusion Matrix Naïve Bayes Classifier*

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	1869	143
Negatif	391	1621

Dapat dilihat pada Tabel 4.12 bahwa dari seluruh jumlah sentimen terdapat 1869 sentimen positif yang terklasifikasi dengan benar dan terdapat kesalahan prediksi sebanyak 143 sentimen yang masuk kedalam sentimen negatif. Dan sebanyak 1621 sentimen negatif telah terklasifikasi dengan benar dan terdapat kesalahan prediksi sebanyak 391 sentimen yang masuk kedalam sentimen positif. Maka berdasarkan *confusion matrix*, diatas didapatkan nilai *accuracy* sebesar 86,7%, yang artinya dari seluruh jumlah data sentimen terdapat sebanyak 3490 sentimen yang benar diklasifikasikan oleh model *Naïve Bayes Classifier*.

4.7 Analisis Klasifikasi Menggunakan Support Vector Machine

Setelah melakukan analisis klasifikasi menggunakan metode *Naïve Bayes Classifier*, metode kedua yang digunakan adalah *Support Vector Machine*. Pada penelitian ini terdapat dua jenis kernel yang digunakan, pertama *Support Vector Machine* dengan kernel linear dan yang kedua dengan kernel *Radial Basis Function* (RBF). Setelah membagi data *training* dan *testing* langkah selanjutnya adalah menentukan parameter C dan γ . Parameter C atau *Cost* adalah parameter yang bekerja sebagai pengoptimalan SVM untuk menghindari misklasifikasi di setiap sampel. Semakin tinggi nilai C maka kemungkinan terjadinya kesalahan dalam penentuan solusi akan semakin kecil, begitu juga sebaliknya semakin kecil nilai C maka semakin tinggi proporsi kesalahan yang terjadi pada solusi. Parameter γ adalah parameter yang menentukan seberapa jauh pengaruh dari satu sampel. Semakin besar nilai γ maka semakin tinggi model tersebut akan menghasilkan hasil yang bias. Untuk SVM dengan kernel linear hanya mempertimbangkan satu parameter, yaitu C dan untuk SVM kernel RBF mempertimbangkan kedua parameter C dan γ . Nilai untuk kedua parameter ini tidak terdapat patokan yang pasti sehingga harus melalui proses *trial and error*. Pada penelitian ini parameter C dan γ yang digunakan mulai dari rentang 10^{-3} sampai 10^3 .

4.7.1 Analisis Klasifikasi Menggunakan Support Vector Machine Kernel Linear

Pada penelitian ini, parameter C yang akan dicoba adalah mulai dari 10^{-3} sampai 10^3 . Selanjutnya data yang telah dibagi menjadi data *training* dan *testing* menggunakan *stratified cross validation* digunakan untuk klasifikasi pada masing-masing parameter C . Hasil perhitungan ketepatan klasifikasi pada tiap parameter C dapat dilihat pada Lampiran 2. Setelah dilakukan perhitungan, diperoleh hasil rata-rata ketepatan klasifikasi terbaik pada data *twitter* mengenai BPJS adalah model klasifikasi SVM dengan parameter $C = 100$ karena menghasilkan nilai rata-rata ketepatan klasifikasi paling besar di antara lainnya. Berikut merupakan hasil pengukuran ketepatan klasifikasi SVM Kernel Linear.

Tabel 4.13 Hasil Ketepatan Klasifikasi SVM Kernel Linear $C = 100$

Subset	Accuracy (%)	Sensitivity (%)
1	91,3	90,5
2	91,8	90,0
3	90,8	85,6
4	91,8	85,6
5	93,0	86,1
6	93,8	88,1
7	95,0	90,5
8	93,5	88,1
9	92,5	85,6
10	91,0	82,6
Rata-Rata	92,5	87,3

Dari subset sebanyak 10 kali dimana setiap subsetnya terdiri dari 10% data testing dan 90% data training yang berbeda-beda di tiap subsetnya didapatkan rata-rata pengukuran ketepatan klasifikasi accuracy sebesar 92,5% nilai accuracy ini menunjukkan rasio prediksi benar untuk kelas positif dan negatif dengan keseluruhan data. Dan nilai rata-rata sensitivity dari model ini sebesar 87,3% dimana nilai sensitivity ini merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Nilai rata-rata tersebut dinilai dapat

mewakili keseluruhan nilai dari 10 subset tersebut. Maka dari hasil pengukuran ketepatan klasifikasi tersebut dapat dikatakan bahwa metode Support Vector Machine Kernel Linear dengan parameter $C = 100$ baik digunakan dalam mengklasifikasikan sentimen data twitter mengenai BPJS. Berikut merupakan *confusion matrix* untuk rata-rata ketepatan klasifikasi.

Tabel 4.14 *Confusion Matrix* SVM Kernel Linear Parameter $C = 100$

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	1965	47
Negatif	256	1756

Dapat dilihat pada Tabel 4.14 bahwa dari seluruh jumlah sentimen terdapat 1965 sentimen positif yang terklasifikasi dengan benar dan terdapat kesalahan prediksi sebanyak 47 sentimen yang masuk kedalam sentimen negatif. Dan sebanyak 1756 sentimen negatif telah terklasifikasi dengan benar dan terdapat kesalahan prediksi sebanyak 256 sentimen yang masuk kedalam sentimen positif. Maka berdasarkan *confusion matrix*, diatas didapatkan nilai *accuracy* sebesar 92,5% yang artinya dari seluruh jumlah data sentimen terdapat sebanyak 3721 sentimen yang benar diklasifikasikan oleh model *Support Vector Machine* Kernel Linear dengan parameter $C = 100$.

4.7.2 Analisis Klasifikasi Menggunakan *Support Vector Machine* Kernel RBF

Tidak jauh berbeda dengan SVM Kernel Linear, pada klasifikasi SVM Kernel RBF setelah melakukan *oversampling* dan pembobotan tiap kata, langkah selanjutnya adalah membagi data menjadi data *training* dan *testing* dengan metode *stratified 10 fold cross validation*. Untuk metode SVM Kernel RBF, ada dua parameter yang digunakan, yaitu C dan γ . Setelah dilakukan perhitungan, diperoleh hasil bahwa nilai rata-rata ketepatan klasifikasi pada parameter $C = 1, 10, 100, 1000$ dan $\gamma = 100, \text{ dan } 1000$ sama besar. Karena semakin tinggi nilai C maka kemungkinan terjadinya kesalahan dalam penentuan solusi akan semakin kecil dan semakin besar nilai γ maka semakin tinggi model tersebut akan menghasilkan hasil yang bias. Oleh karena itu model SVM Kernel RBF yang dipilih adalah model dengan menggunakan parameter $C = 1000$ dan $\gamma = 100$. Untuk hasil ketepatan klasifikasi parameter yang lain akan ditampilkan pada Lampiran 3 hingga Lampiran 9. Berikut merupakan hasil pengukuran SVM Kernel RBF $C = 1000$ dan $\gamma = 100$.

Tabel 4.15 Hasil Perhitungan Ketepatan Klasifikasi SVM Kernel RBF $C = 1000, \gamma = 100$

Subset	Accuracy (%)	Sensitivity (%)
1	92,8	99,0
2	94,0	100
3	93,5	99,0
4	97,8	99,0
5	99,0	98,0
6	99,5	99,5
7	98,8	98,5
8	98,3	97,5
9	99,0	98,5
10	98,5	98,0
Rata-rata	97,1	98,7

Dari subset sebanyak 10 kali dimana setiap subsetnya terdiri dari 10% data testing dan 90% data training yang berbeda-beda di tiap subsetnya didapatkan rata-rata pengukuran ketepatan klasifikasi *accuracy* sebesar 97,1% nilai *accuracy* ini menunjukkan rasio prediksi benar untuk kelas positif dan negatif dengan keseluruhan data. Dan nilai rata-rata *sensitivity* dari model ini sebesar 98,7% dimana nilai *sensitivity* ini merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Nilai rata-rata tersebut dinilai dapat mewakili keseluruhan nilai 10 subset tersebut. Maka dari hasil pengukuran ketepatan klasifikasi tersebut dapat dikatakan bahwa metode Support Vector Machine Kernel Linear dengan parameter $C = 1000$ dan $\gamma = 100$ baik digunakan dalam mengklasifikasikan sentimen data twitter mengenai BPJS. Berikut merupakan confusion matrix untuk rata-rata ketepatan klasifikasi. Berikut merupakan *confusion matrix* untuk rata-rata ketepatan klasifikasi.

Tabel 4.16 *Confusion Matrix* SVM Kernel RBF $C = 1000$ $\gamma = 100$

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	1922	90
Negatif	26	1986

Dapat dilihat pada Tabel 4.16 bahwa dari seluruh jumlah sentimen terdapat 1992 sentimen positif yang terklasifikasi dengan benar dan terdapat kesalahan prediksi sebanyak 90 sentimen yang masuk kedalam sentimen negatif. Dan sebanyak 1986 sentimen negatif telah terklasifikasi dengan benar dan terdapat kesalahan prediksi sebanyak 26 sentimen yang masuk kedalam sentimen positif. Maka berdasarkan *confusion matrix*, diatas didapatkan nilai *accuracy* sebesar 97,1% yang artinya dari seluruh jumlah data sentimen terdapat sebanyak 3908 sentimen yang benar diklasifikasikan oleh model *Support Vector Machine* Kernel RBF dengan parameter $C = 1000$ $\gamma = 100$.

4.8 Perbandingan Naïve Bayes Classifier dan Support Vector Machine

Setelah memperoleh hasil ketepatan klasifikasi ketiga metode, maka langkah selanjutnya adalah membandingkan hasil ketepatan klasifikasi dari kedua metode tersebut. Perbandingan metode *Naïve Bayes Classifier* dan *Support Vector Machine* pada penelitian ini adalah membandingkan hasil terbaik dari rata-rata pengukuran ketepatan klasifikasi dari tiap metode karena nilai rata-rata tersebut dinilai dapat mewakili keseluruhan nilai dari 10 subset.

Tabel 4.17 Perbandingan Ketepatan Klasifikasi

Model	<i>Accuracy</i> (%)	<i>Sensitivity</i> (%)
<i>Naïve Bayes Classifier</i>	86,7	92,0
SVM Kernel Linear	92,5	87,3
SVM Kernel RBF	97,1	98,7

Perbandingan ketepatan klasifikasi dapat dilihat pada Tabel 4.17 menunjukkan bahwa secara keseluruhan hasil ketepatan klasifikasi dengan menggunakan metode SVM Kernel Linear maupun Kernel RBF lebih baik dibandingkan metode *Naïve Bayes Classifier*. Hal ini sejalan dengan pernyataan Scolkopf dan Smola (2002) bahwa fungsi kernel RBF dapat memetakan hubungan tidak linear dan RBF lebih baik terhadap *outlier* karena kernel RBF berada diantara selang $(-\infty, +\infty)$. Hal tersebut dilihat dari nilai rata-rata *accuracy* dan *sensitivity*

dari masing-masing metode. Jika dibandingkan dengan SVM Kernel Linear, hasil ketepatan klasifikasi SVM Kernel RBF yang paling baik diantara ketiga metode tersebut, karena nilai rata-rata *accuracy* dan *sensitivity* untuk SVM Kernel RBF lebih besar dibanding yang lain yaitu masing-masing sebesar 97,1% dan 98,7%. Sedangkan untuk SVM Kernel Linear nilai rata-rata *accuracy* dan *sensitivity* masing-masing sebesar 92,5% dan 87,3%. Dan untuk *Naïve Bayes Classifier* nilai rata-rata *accuracy* dan *sensitivity* masing-masing sebesar 86,7% dan 92,0%. Maka berdasarkan hasil rata-rata tersebut dapat dikatakan bahwa metode *Support Vector Machine* Kernel RBF dengan parameter $C = 1000$ $\gamma = 100$ merupakan metode yang paling baik digunakan pada penelitian ini.

4.9 Visualisasi Word Cloud

Visualisasi data teks menggunakan *word cloud* digunakan untuk menemukan kata yang paling sering muncul dalam data teks. Dalam penelitian ini, *word cloud* digunakan untuk memvisualisasikan *tweet* berdasarkan klasifikasi sentimen untuk melihat kata mana yang sering muncul dalam data *tweet*. Ukuran *font* pada *word cloud* menunjukkan frekuensi kemunculan kata. Semakin besar ukuran *font* maka semakin besar frekuensi kemunculan kata tersebut. Begitupun sebaliknya, semakin kecil ukuran *font* maka semakin kecil pula frekuensi kemunculan kata tersebut. Berikut merupakan visualisasi *word cloud* pada penelitian ini.



Gambar 4. 4 Visualisasi Word cloud

Pada Gambar 4.4 dapat dilihat bahwa kata yang memiliki frekuensi yang besar adalah kata ‘peserta’, ‘program’, dan ‘iuran’. Kata ‘iuran’ memiliki frekuensi kemunculan yang sering dikarenakan banyak *tweet* mengungkapkan ketidaksetujuan peserta akan kenaikan iuran BPJS di masa pandemi. Sehingga kata ‘peserta’ juga memiliki frekuensi kemunculan yang sering. Dan untuk kata ‘program’, BPJS sedang mensosialisasikan program-program yang ada di BPJS Kesehatan maupun Ketenagakerjaan kepada masyarakat umum maupun mahasiswa. Dan BPJS berencana untuk mengadakan program kerjasama di beberapa kampus. Pada Gambar 4.4 dapat dilihat terdapat kata ‘defisit’ yang memiliki frekuensi kemunculan lumayan sering. Hal ini karena terdapat kabar bahwa BPJS Kesehatan mengalami defisit yang mengakibatkan terjadinya kebijakan untuk menaikkan iuran oleh pemerintah untuk mengatasi defisit tersebut.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil analisis dan pembahasan, didapatkan kesimpulan sebagai berikut.

1. Penerapan Metode *Naïve Bayes Classifier* memiliki performa yang baik dalam pengklasifikasian dan *Support Vector Machine* Kernel Linear maupun RBF memiliki performa yang sempurna dalam pengklasifikasian data sentimen mengenai BPJS pada media sosial *Twitter*. Hal ini berdasarkan pada kriteria nilai akurasi pada Sub bab 2.14.
2. Hasil rata-rata ketepatan klasifikasi pada metode *Naïve Bayes Classifier*, *Support Vector Machine* Kernel Linear, dan *Support Vector Machine* Kernel RBF masing-masing sebesar 86,7%; 92,5%; dan 97,1%
3. Perbandingan ketepatan klasifikasi dari *Naïve Bayes Classifier* dan *Support Vector Machine* menunjukkan bahwa *Support Vector Machine* Kernel Linear dan RBF memiliki ketepatan klasifikasi yang sama-sama lebih baik dalam mengklasifikasikan data sentimen pada media sosial *twitter* mengenai BPJS.

5.2 Saran

Saran yang dapat diberikan dari hasil penelitian ini adalah sebagai berikut.

1. Penelitian ini dapat dijadikan referensi dengan menggunakan data *facebook*. Karena *facebook* memiliki perbedaan rentang usia dengan pengguna *twitter*. Pengguna *facebook* berkisar di usia 18-34 tahun untuk perempuan dan 25-34 tahun untuk laki-laki. Sedangkan pengguna *twitter* berkisar di usia 16-24 tahun dan mayoritas merupakan pengguna laki-laki.

(Halaman ini sengaja dikosongkan)

DAFTAR PUSTAKA

- Area, Universitas Medan. (2016). *LP2M UMA: ANALISIS SENTIMEN (SENTIMENT ANALYSIS) : DEFINISI, TIPE DAN CARA KERJANYA*. Retrieved from LP2M UMA: <https://lp2m.uma.ac.id/2022/02/21/analisis-sentimen-sentiment-analysis-definisi-tipe-dan-cara-kerjanya/>
- BPJS Kesehatan. (2021). *BPJS Kesehatan*. Retrieved February 8, 2022, from <https://www.bpjs-kesehatan.go.id/bpjs/home>
- Bramer, M. (2007). *Principles of Data Mining*.
- Chawla, N. V. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*.
- Davidson, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press. doi:<https://doi.org/10.1017/CBO9780511802843>
- Findawati, Y., & Rosidi, A. (2020). *Buku Ajar Text Mining*. Sidoarjo: UMSIDA Press.
- Gokgoz, E., & Subasi, A. (2015). Comparison of Decision Tree Algorithms for EMG Signal Classification Using DWT. *Biomedical Signal Processing and Control*. doi:10.1016/j.bspc.2014.12.005
- Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques* (Vol. 12). Berlin: Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-19721-5
- Gumilang, Z. A. (2018). Implementasi Naive Bayes Classifier dan Asosiasi Untuk Analisis Sentimen Data Ulasan Aplikasi E-Commerce Shopee pada Situs Google Play.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Technique*. Morgan Kaufman.
- Harijianto, S. D. (2019). Analisis Sentimen Pada Twitter Menggunakan Multinomial Naive Bayes. Retrieved from https://repository.usd.ac.id/35993/2/145314060_full.pdf
- Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab-An S4 Package for Kernel Methods in R. *Journal of Statistical Software*.
- Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*.
- Kominfo. (2022). *Indonesia Peringkat Lima Pengguna Twitter*. Retrieved from https://kominfo.go.id/content/detail/2366/%20indonesia-peringkat-lima-penggunatwitter/0/sorotan_media
- Kristiyanti, D. A. (2015). Analisis Sentimen Review Produk Kosmetik Menggunakan Algoritma Support Vector Machine dan Particle Swarm Optimization Sebagai Metode Seleksi Fitur. *Seminar Nasional Inovasi dan Tren (SNIT)*.
- Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*. doi:10.2200/S00416ED1V01Y201204HLT016
- Mahmood, A. M. (2015). Class Imbalance Learning in Data Mining - A Survey. *International Journal of Communication Technology for Social Networking Services*. doi:10.21742/ijctsns.2015.3.2.2
- McNaught, C., & Lam, P. (2010). Using Wordle as a Supplementary Research. *The Qualitative Report*.

- Mustika, & Ardilla, Y. (2021). *Data Mining dan Aplikasinya*. Bandung: Widina Bhakti Persada Bandung.
- Nugraha, F. A. (2020). *Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning*. Bandung: Kreatif Industri Nusantara.
- Nurulbaiti, Subekti, and, F., & Retno. (2018). Analisis Sentimen Terhadap Data Tweet untuk Badan Penyelenggara Jaminan Sosial (BPJS) Menggunakan Program R. *Jurnal Pendidikan Matematika dan Sains UNY*. Retrieved from <http://eprints.uny.ac.id/id/eprint/56367>
- Permatasari, R. W. (2021). Analisis Sentimen Masyarakat Indonesia Mengenai Vaksin COVID-19 Pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier dan Support Vector Machine.
- Rahman, H. (2021). Klasifikasi Sentimen Masyarakat Terhadap Layanan Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan di Twitter Menggunakan Metode K-Nearest Neighbor. Retrieved from <http://repository.uin-suska.ac.id/54988/2>
- RUU BPJS. (2011). *Undang Undang Republik Indonesia Nomor 24 Tahun 2011 Tentang Badan Penyelenggara Jaminan Sosial*. Jakarta. Retrieved from <https://bpjs-kesehatan.go.id/bpjs/dmdocuments/20e67493084e6d2e600888b1dd9f94f4.pdf>
- Santoso, G. T. (2021). Analisis Sentimen Pada Tweet Dengan Tagar #BPJSRasaRentenir Menggunakan Metode Support Vector Machine (SVM).
- Scholkopf, B., & Smola, A. J. (2002). *Learning with Kernel: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press.
- Statista Research Department. (2022). *Topic : Twitter*. Retrieved from https://www.statista.com/topics/737/twitter/#topicHeader__wrapper
- Stedman, C. (Copyright 2010 - 2022). *Text Mining (Text Analytics)*. Retrieved from SearchBusinessAnalytics: <https://www.techtarget.com/searchbusinessanalytics/definition/text-mining#:~:text=Text%20mining%20is%20the%20process,other%20attributes%20in%20the%20data.>
- Subagyo, A. R. (2021). Analisis Sentimen Untuk Menentukan Popularitas Marketplace Menggunakan Metode Naive Bayes Classifier.
- Syakuro, A. (2017). Analisis Sentimen Masyarakat Terhadap E-Commerce Pada Media Sosial Menggunakan Metode Naive Bayes Classifier (NBC) Dengan Seleksi Fitur Information Gain (IG). Retrieved from <http://etheses.uin-malang.ac.id/id/eprint/11706>
- Triawati, C. (2009). Metode Pembobotan Statistical Concept Based untuk Klustering dan Kategorisasi Dokumen Berbahasa Indonesia.
- Xhemali, D., J. Hinde, C., & G. Stone, R. (2009). Naive Bayes vs Decision Tree vs Neural Networks in the Classification of Training Web Pages. *IJCSI International Journal of Computer Science Issues*, 4(1).

LAMPIRAN

Lampiran 1. Data *Tweet* Setelah *Pre-processing Text* dan Pelabelan

<i>Tweet</i>	Sentimen
kebocoran data bpjs kesehatan pada mei data sejumlah peserta bpjs dijual di raid forums	Negatif
seorang pegawai dan lima tenaga alih daya bpjs kesehatan bondowoso terkonfirmasi positif covid	Negatif
disnaker lampung mengusulkan petugas komisi pemilihan umum tingkat desa hingga provinsi dilindungi bpjs ketenagakerjaan	Positif
pekerja terdampak covid di kota surabaya mendapat bantuan beras dari badan penyelenggara bp jaminan sosial tenaga kerja jamsostek	Positif
kepala bpjs kesehatan cabang jember mengimbau badan usaha memanfaatkan secara maksimal elektronik data badan usaha e dabu mobile yang sudah diluncurkan beberapa waktu lalu	Positif
bpjs ketenagakerjaan manokwari menggelar platinum gathering badan usaha selasa kemarin	Positif
bpjs kesehatan mendapatkan penghargaan top digital awards yang diselenggarakan oleh majalah itworks	Positif
pt timah tbk memberikan jaminan perlindungan sosial bagi nelayan dan kelompok rentan di wilayah lingkaran tambang perusahaan melalui bpjs ketenagakerjaan	Positif
tingkatkan kualitas pelayanan bagi para peserta bpjs ketenagakerjaan sulawesi tenggara melaksanakan kegiatan monitoring evaluasi pusat layanan kecelakaan kerja jumat	Positif
kepala staf kepresidenan menyebutkan selama ini para perokok merupakan penyumbang beban terbesar di bpjs	Negatif
⋮	⋮
klaim peserta non formal rugikan bpjs kesehatan	Negatif
presiden joko widodo mengakui belum semua rumah sakit swasta mau bekerja sama dalam program badan penyelenggara jaminan sosial kesehatan	Negatif
ketiadaan solusi dari kementerian kesehatan dan badan penyelenggara jaminan sosial kesehatan terkait dengan kenaikan iuran bagi semua kelas membuat komisi dpr sebagai mitra merasa geram	Negatif
bpjs kesehatan tidak mau gabung dengan bpjs syariah	Negatif
badan penyelenggara jaminan sosial kesehatan akan mengalami defisit likuiditas	Negatif
rumah sakit banyak menolak kartu bpjs kesehatan	Negatif
bpjs kesehatan rawan penyelewengan besaran pembayaran	Negatif
badan penyelenggara jaminan sosial bpjs kesehatan keluarga saya ada yang mengalami kecelakaan lalu lintas	Negatif
saat ini data peserta badan penyelenggara jaminan sosial bpjs kesehatan yang bermasalah telah diserahkan	Negatif
pasien bpjs masih dianaktirikan bandar lampung	Negatif
salah satu kenaikan yang harus dihadapi masyarakat pada mendatang adalah iuran badan penyelenggara jaminan sosial bpjs kesehatan	Negatif

Lampiran 2 Hasil Ketepatan Klasifikasi SVM Kernel Linear

<i>C</i>	<i>Accuracy (%)</i>	<i>C</i>	<i>Accuracy (%)</i>
0.001	49,9	10	91,3
	49,9		90,8
	49,9		90,1
	49,9		91,3
	59,7		92,0
	57,0		92,0
	59,0		94,5
	59,5		94,3
	58,5		92,0
	55,0		89,8
Rata-rata	54,8	Rata-rata	91,8
0.01	50,9	100	91,3
	51,9		91,8
	74,7		90,8
	74,7		91,8
	59,7		93,0
	57,0		93,8
	59,0		95,0
	59,5		93,5
	58,5		92,5
	55		91,0
Rata-rata	60,1	Rata-rata	92,5
0.1	83,4	1000	90,1
	83,4		88,1
	80,1		87,3
	81,6		90,1
	81,8		91,5
	82,3		90,8
	81,3		92,3
	83,8		91,8
	81,6		91,5
	79,6		88,6
Rata-rata	81,9	Rata-rata	90,2
1	90,6		
	89,6		
	87,3		
	88,1		
	90,3		
	91,5		
	93,0		
	93,0		
	90,3		
	88,6		
Rata-rata	90,2		

Lampiran 3 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 0.001$

<i>C</i>	<i>Accuracy (%)</i>	<i>C</i>	<i>Accuracy (%)</i>
0,001	49,9	10	61,0
	49,9		63,5
	49,9		66,7
	49,9		66,0
	59,7		65,2
	57,0		63,7
	59,0		67,2
	59,5		67,2
	58,5		65,7
	55,0		62,4
Rata-rata	54,8	Rata-rata	64,9
0,01	49,9	100	86,1
	49,9		85,4
	49,9		82,4
	49,9		84,4
	59,7		85,6
	57,0		86,3
	59,0		86,6
	59,5		87,1
	58,5		85,8
	55,0		83,1
Rata-rata	54,8	Rata-rata	85,3
0,1	49,9	1000	91,6
	49,9		90,1
	49,9		89,6
	49,9		88,1
	59,7		91,5
	57,0		91,0
	59,0		94,8
	59,5		91,8
	58,5		91,0
	55,0		89,1
Rata-rata	54,8	Rata-rata	90,9
1	49,9		
	49,9		
	49,9		
	49,9		
	59,7		
	57,0		
	59,0		
	59,5		
	58,5		
	55,0		
Rata-rata	54,8		

Lampiran 4 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 0.01$

<i>C</i>	<i>Accuracy (%)</i>	<i>C</i>	<i>Accuracy (%)</i>
0,001	49,9	10	86,1
	49,9		85,4
	49,9		82,4
	49,9		84,4
	59,7		85,6
	57,0		86,3
	59,0		86,6
	59,5		87,1
	58,5		86,1
	55,0		83,1
Rata-rata	54,8	Rata-rata	85,3
0,01	49,9	100	91,6
	49,9		90,1
	49,9		89,6
	49,9		88,1
	59,7		91,5
	57,0		91,5
	59,0		95,0
	59,5		91,8
	58,5		91,5
	55,0		88,8
Rata-rata	54,8	Rata-rata	91,0
0,1	49,9	1000	91,6
	49,9		89,6
	49,9		90,1
	49,9		92,8
	59,7		92,0
	57,0		93,5
	59,0		94,8
	59,5		93,8
	58,5		92,3
	55,0		90,0
Rata-rata	54,8	Rata-rata	92,0
1	60,5		
	63,3		
	66,5		
	66,0		
	65,9		
	63,7		
	66,9		
	67,4		
	65,7		
	62,4		
Rata-rata	64,8		

Lampiran 5 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 0.1$

<i>C</i>	<i>Accuracy (%)</i>	<i>C</i>	<i>Accuracy (%)</i>
0,001	49,9	10	91,6
	49,9		90,8
	49,9		89,6
	49,9		89,3
	59,7		92,5
	57,0		93,0
	59,0		95,5
	59,5		93,0
	58,5		91,5
	55,0		89,6
Rata-rata	54,8	Rata-rata	91,7
0,01	49,9	100	93,1
	49,9		92,3
	49,9		90,8
	49,9		92,3
	59,7		93,8
	57,0		94,3
	59,0		97,0
	59,5		95,0
	58,5		93,3
	55,0		92,3
Rata-rata	54,8	Rata-rata	93,4
0,1	55,3	1000	92,6
	58,6		92,6
	61,3		90,6
	60,3		92,1
	61,7		93,5
	59,5		94,5
	60,2		97,3
	59,7		94,3
	59,5		92,3
	59,0		91,8
Rata-rata	59,5	Rata-rata	93,1
1	85,9		
	85,4		
	82,1		
	84,1		
	84,8		
	85,8		
	86,6		
	87,3		
	85,1		
	83,6		
Rata-rata	85,1		

Lampiran 6 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 1$

<i>C</i>	<i>Accuracy (%)</i>	<i>C</i>	<i>Accuracy (%)</i>
0,001	49,9	10	93,3
	49,9		95,0
	49,9		93,3
	49,9		93,5
	59,7		96,0
	57,0		97,0
	59,0		97,5
	59,5		97,0
	58,5		95,3
	55,0		94,8
Rata-rata	54,8	Rata-rata	95,3
0,01	49,9	100	93,3
	49,9		95,0
	49,9		93,3
	49,9		93,5
	59,7		96,0
	57,0		97,0
	59,0		97,5
	59,5		97,0
	58,5		95,3
	55,0		94,8
Rata-rata	54,8	Rata-rata	95,3
0,1	74,9	1000	93,3
	80,9		95,0
	79,2		93,3
	80,6		93,5
	81,1		96,0
	80,6		97,0
	80,6		97,5
	82,8		97,0
	79,4		95,3
	79,1		94,8
Rata-rata	79,9	Rata-rata	95,3
1	93,8		
	95,8		
	91,3		
	93,1		
	96,0		
	96,0		
	96,3		
	96,0		
	95,3		
	94,3		
Rata-rata	94,8		

Lampiran 7 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 10$

<i>C</i>	<i>Accuracy (%)</i>	<i>C</i>	<i>Accuracy (%)</i>
0,001	49,9	10	92,8
	49,9		94,0
	49,9		93,5
	49,9		97,8
	53,2		98,8
	52,5		99,5
	55,0		98,8
	52,5		98,3
	51,7		99,0
	54,0		98,5
Rata-rata	51,8	Rata-rata	97,1
0,01	49,9	100	92,8
	49,9		94,0
	49,9		93,5
	49,9		97,8
	53,2		98,8
	52,5		99,5
	55,0		98,8
	52,5		98,3
	51,7		99,0
	54,0		98,5
Rata-rata	51,8	Rata-rata	97,1
0,1	50,6	1000	92,8
	53,6		94,0
	55,6		93,5
	54,8		97,8
	53,2		98,8
	53,5		99,5
	55,7		98,8
	53,0		98,3
	52,2		99,0
	54,0		98,5
Rata-rata	53,6	Rata-rata	97,1
1	92,8		
	94,0		
	93,5		
	97,8		
	98,8		
	99,5		
	98,8		
	98,3		
	99,0		
98,5			
Rata-rata	97,1		

Lampiran 8 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 100$

<i>C</i>	<i>Accuracy (%)</i>	<i>C</i>	<i>Accuracy (%)</i>
0,001	49,9	10	92,8
	49,9		94,0
	49,9		93,5
	49,9		97,8
	52,7		99,0
	52,5		99,5
	55,0		98,8
	52,5		98,3
	51,7		99,0
	53,7		98,5
Rata-rata	51,8	Rata-rata	97,1
0,01	49,9	100	92,8
	49,9		94,0
	49,9		93,5
	49,9		97,8
	52,7		99,0
	52,5		99,5
	55,0		98,8
	52,5		98,3
	51,7		99,0
	53,7		98,5
Rata-rata	51,8	Rata-rata	97,1
0,1	50,6	1000	92,8
	53,6		94,0
	55,6		93,5
	54,8		97,8
	53,5		99,0
	53,5		99,5
	55,0		98,8
	53,0		98,3
	51,7		99,0
	53,7		98,5
Rata-rata	53,5	Rata-rata	97,1
1	92,8		
	94,0		
	93,5		
	97,8		
	99,0		
	99,5		
	98,8		
	98,3		
	99,0		
	98,5		
Rata-rata	97,1		

Lampiran 9 Hasil Ketepatan Klasifikasi SVM Kernel RBF $\gamma = 1000$

<i>C</i>	<i>Accuracy (%)</i>	<i>C</i>	<i>Accuracy (%)</i>
0,001	49,9	10	92,8
	49,9		94,0
	49,9		93,5
	49,9		97,8
	52,5		99,0
	52,5		99,5
	55,0		98,8
	52,5		98,3
	51,7		99,0
	53,7		98,5
Rata-rata	51,7	Rata-rata	97,1
0,01	49,9	100	92,8
	49,9		94,0
	49,9		93,5
	49,9		97,8
	52,5		99,0
	52,5		99,5
	55,0		98,8
	52,5		98,3
	51,7		99,0
	53,7		98,5
Rata-rata	51,7	Rata-rata	97,1
0,1	50,6	1000	92,8
	53,6		94,0
	54,8		93,5
	54,8		97,8
	53,5		99,0
	53,5		99,5
	55,0		98,8
	53,0		98,3
	51,7		99,0
	53,7		98,5
Rata-rata	53,4	Rata-rata	97,1
1	92,8		
	94,0		
	93,5		
	97,8		
	99,0		
	99,5		
	98,8		
	98,3		
	99,0		
	98,5		
Rata-rata	97,1		

Lampiran 10 Syntax scrapping data menggunakan Python

```
!pip install snsrape
# Import libraries
import csv
import pandas as pd
import snsrape.modules.twitter as sntwitter
#Specify name and directory of csv file as the result of the scraping
csvFile = open('C:\\Users\\asus\\Desktop\\twitter baru.csv','a', encoding = 'utf-8') #creates a
file in which you want to store the data.
csvWriter = csv.writer(csvFile)
# Customize the scraping based on variables needed, keywords of tweets, the amount of
tweets, and the date of tweets
maxTweets = 3000 # the number of tweets you require
#2021-09-22
for i,tweet in enumerate(sntwitter.TwitterSearchScrapper('klaim OR bpjs OR BPJS OR Bpjs
OR claim OR Badan Penyelenggara Jaminan Sosial OR Klaim OR KLAIM' + 'since:2019-
01-01 until:2021-12-31').get_items()) :
    if i > maxTweets :
        break
    print(tweet.date, i)
    csvWriter.writerow([tweet.date, tweet.id, tweet.content,
tweet.user.username.encode('utf-8'), tweet.user.verified, tweet.user.followersCount,
tweet.user.friendsCount, tweet.mentionedUsers, tweet.replyCount, tweet.retweetCount,
tweet.likeCount, tweet.media, tweet.lang.encode('utf-8'), tweet.user.location])
    #If you need more information, just provide the attributes
header_name = ['date', 'tweet_id', 'text', 'username', 'verified', 'followers','following',
'mentioned users','retweet','like ', 'reply', 'media', 'language', 'location']
df = pd.read_csv("C:\\Users\\asus\\Desktop\\twitter baru.csv", names=header_name)
```

Lampiran 11 *Syntax preprocessing data menggunakan Python*

```
##preprocessing data
import pandas as pd
import re
import nltk
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
nltk.download('stopwords')
nltk.download('punkt')
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
import os
import tweepy
import string
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
from textblob import TextBlob
import preprocessor as p
from preprocessor.api import clean, tokenize, parse
import datetime
from datetime import timedelta
import emoji
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
import seaborn as sns
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
#remove mention and retweet
def remove_pattern(text, pattern_regex):
    r = re.findall(pattern_regex, text)
    for i in r:
        text = re.sub(i, "", text)
    return text
df['clean_tweet'] = np.vectorize(remove_pattern)(df['text'], "*RT* | *@[\w]*")
```

Lampiran 11 *Syntax preprocessing* data menggunakan *Python* (Lanjutan)

```
#remove mention and retweet
def remove_pattern(text, pattern_regex):
    r = re.findall(pattern_regex, text)
    for i in r:
        text = re.sub(i, "", text)
    return text
df['clean_tweet'] = np.vectorize(remove_pattern)(df['text'], " *RT* | *@[\w]*")
df.head()

# remove simbol
def remove(text):
    text = ''.join(re.sub("(@[A-Za-z0-9]+)|(^0-9A-Za-z \t)|(\w+:\w+\S+)", "", text).split())
    return text
df['remove_http'] = df['clean_tweet'].apply(lambda x: remove(x))
df.head()

def remov(tweet):
    #remove $GE
    tweet = re.sub(r'\$w*', '', tweet)
    tweet = re.sub(r'^RT[\s]+', '', tweet)
    tweet = re.sub(r'#', '', tweet)
    tweet = re.sub('[0-9]+', '', tweet)
    return tweet
df['remove_hashtag'] = df['remove_http'].apply(lambda x: remov(x))

#remove duplicate
df.drop_duplicates(subset="remove_hashtag", keep='first', inplace=True)
df.head()

#import stopwords
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stopwords_indonesia = stopwords.words('indonesian')
from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory, StopWordRemover, ArrayDictionary
stop_factory = StopWordRemoverFactory().get_stop_words()
```


Lampiran 11 *Syntax preprocessing* data menggunakan *Python* (Lanjutan 2)

```
more_stopwords = [  
    'yg','utk','cuman','deh','Btw','tapi','gua','gue','lo','lu','kalo','trs',  
    'jd','nih','ntar','nya','lg','gk','dpt','dr','kpn','kok','kyk','dong',  
    'donk','yah','u','ya','ga','km','eh','sih','bang','br','rp','Rp','jt','kan',  
    'gpp','sm','usah','mas','sob','thx','ato','jg','gw','wkwk','mak','haha','iy',  
    'k','tp','dg','dr','dri','duh','ye','wkwkwk','syg','btw','gaes','guys','moga',  
    'smg','kmrn','nemu','yukkk','yuk','klas','kls','lho','sbnry','org','gt','gitu',  
    'gtu','bwt','klrg','lbh','cpt','ku','mba','mas','sdh','spt','dlm','bs','jgn',  
    'kmn','tdk','tuh','dah','kek','pls','pd','cm','kpd','byr','byar','bpjs','badan','penyelenggara',  
    'jaminan','sosial','klaim'  
]  
data = stop_factory + more_stopwords  
  
dictionary = ArrayDictionary(data)  
str = StopWordRemover(dictionary)  
print(data)  
#import sastrawi  
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory  
factory = StemmerFactory()  
stemmer = factory.create_stemmer()  
#tokenize  
from nltk.tokenize import TweetTokenizer  
stem_word = stemmer.stem(word) #stemming word  
    tweets_clean.append(stem_word)  
  
return tweets_clean  
df["Tweet"] = df['remove_hashtag'].apply(lambda x: clean_tweets(x))  
df.drop_duplicates(subset='remove_hashtag', keep='first', inplace=True)  
df.head()
```

Lampiran 12 Syntax Pelabelan, *Oversampling* data dan TF-IDF Menggunakan Python

```
#pelabelan
df = pd.read_excel("data labeling.xlsx")
#konversi label ke polaritas
def convert(polarity):
    if polarity == 'Positif':
        return 1
    else:
        return -1
df['Label'] = df['Sentiments'].apply(convert)
df['Sentiments'].value_counts()

from numpy import mean
from imblearn.over_sampling import RandomOverSampler
#oversampling
ros = RandomOverSampler()
train_x, train_y = ros.fit_resample(np.array(df['stop removal']).reshape(-1, 1),
np.array(df['label']).reshape(-1, 1));
train_os = pd.DataFrame(list(zip([x[0] for x in train_x], train_y)), columns = ['stop removal',
'label']);
train_os['label'].value_counts()
X = train_os['stop removal'].values
y = train_os['label'].values
#tfidf
clf = CountVectorizer()
X_cv = clf.fit_transform(X)
tf_transformer = TfidfTransformer(use_idf=True).fit(X_cv)
X_tf = tf_transformer.transform(X_cv)
```

Lampiran 13 Naïve Bayes Classifier menggunakan Python

```
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics
from sklearn.pipeline import Pipeline
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
#define the K-Fold
kfold = StratifiedKFold(n_splits = 10)
Multinomial Naïve Bayes
#classifier data
from sklearn.model_selection import cross_val_predict
nb = MultinomialNB()
pred = cross_val_predict(nb, X_tf, y, cv=kfold)
prob = cross_val_predict(nb, X_tf, y, cv=kfold, method='predict_proba')
prob
scores_accuracy = cross_val_score(nb, X_tf, y, scoring='accuracy', cv=kfold)
scores_accuracy.mean()
scores_recall = cross_val_score(nb, X_tf, y, scoring='recall', cv=kfold)
scores_recall.mean()
#confusion matrix
cm =confusion_matrix(y, pred)
```

Lampiran 14 Support Vector Machine menggunakan Python

```
from sklearn import svm
from sklearn.svm import LinearSVC
from sklearn.svm import SVC
from sklearn.model_selection import GridSearchCV

SVM Linear
## linear C = 100
svmlin6 = svm.SVC(C=100, kernel='linear')
y_predsvm6 = cross_val_predict(svmlin6, X_tf, y, cv=kfold)
accuracy6 = cross_val_score(svmlin6, X_tf, y, cv=kfold, scoring='accuracy')
accuracy6.mean()
recall6 = cross_val_score(svmlin6, X_tf, y, cv=kfold, scoring='recall')
recall6.mean()
#confusion matrix
cm6 = confusion_matrix(y, y_predsvm6)
cm6

SVM RBF
## rbf c = 1 gamma = 10
svc = SVC(C=1, gamma= 10, kernel='rbf')
y_predsvc = cross_val_predict(svc, X_tf, y, cv=kfold)
accuracy = cross_val_score(svc, X_tf, y, cv=kfold, scoring='accuracy')
accuracy.mean()
recall = cross_val_score(svc, X_tf, y, cv=kfold, scoring='recall')
recall.mean()
#confusion matrix
cm = confusion_matrix(y, y_predsvc)
cm
```

Lampiran 15 Syntax Word cloud

```
stopword = nltk.corpus.stopwords.words('indonesian')
text=df['tweet']
stopword=open("stopwords.txt").read()
stopword=set(stopword.split())
not_stopword={ }
new_stopword=set([word for word in stopword if not word in not_stopword])
data_stop=[]
for line in text:
    word_token=nltk.word_tokenize(line)
    word_token=[word for word in word_token if not word in stopword]
    data_stop.append(" ".join(word_token))
#mengubah data_stop menjadi string
a=str(data_stop)
word=re.sub(r"\"", "",a)
#Wordcloud
wordcloud = WordCloud(width=2000, height=1000, collocations =
False,background_color='white',stopwords=stopword, max_words=150,
max_font_size=700,random_state=1).generate(word)
fig = plt.figure(1)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
#Menyimpan Wordcloud
fig.savefig("wordcloud_bpjs.png", dpi=1480)
```

(Halaman ini sengaja dikosongkan)

BIODATA PENULIS



Penulis bernama Diva Durrotun Nada yang biasa disapa Diva merupakan anak sulung dari dua bersaudara yang lahir di Lamongan pada 21 April 2000. Penulis mulai menempuh Pendidikan formal di TK Panca Kumara, SDN 29 Dangin Puri, SMPN 4 Denpasar, SMAN 5 Denpasar, dan pada tahun 2018 Penulis berkesempatan melanjutkan pendidikan di Departemen Aktuaria Fakultas Sains dan Analitika Data Institut Teknologi Sepuluh Nopember (ITS). Penulis juga berkesempatan menjadi bagian dari Himpunan Mahasiswa Aktuaria (HIMASAKTA) ITS. Selain menjadi bagian himpunan mahasiswa, Penulis juga berkesempatan mengikuti beberapa kepanitian seperti panitia kegiatan OKKBK Aktuaria 2019, Gerigi ITS 2019, dan Latihan Keterampilan Manajemen Mahasiswa Tingkat Dasar (LKMM-TD) I Himasakta 2020. Selain menjadi panitia, Penulis juga pernah mengikuti beberapa pelatihan yaitu LKMW, dan LKMM-Pra TD. Selain memiliki pengalaman dibidang organisasi dan kepanitian, Penulis juga pernah mengikuti program *internship* di Koperasi Simpan Pinjam dan Pembiayaan Syariah BMT BIM, dan Bank Mandiri. Jika ada keperluan atau ingin berdiskusi dengan penulis, dapat menghubungi email : divadnada@gmail.com.