BAB II

LANDASAN TEORI

2.1 Belanja *Online*

Pada zaman *modern* dan era teknologi kini umumnya berbelanja dapat dilakukan dengan berbagai macam cara yang variatif serta pilihan yang beragam sesuai dengan kebutuhan yang ada, jika dirujukkan pada [7]**,** belanja *online* atau *E-Commerce* adalah proses transaksi yang dilakukan melalui media atau perantara yang berupa situs-situs jual beli *via online* dan menyediakan barang atau jasa yang diperjualbelikan.

Ketika berbelanja *online* yang dilakukan masyrakat umumnya adalah tingkat kemudahan yang diberikan dan tingkat efisiensi. Menurut sumber lain dari Jurnal yang dilakukan [8] menyebutkan bahwa belanja *online* dapat menghemat waktu bagi orang modern karena mereka begitu sibuk sehingga tidak memiliki banyak waktu atau bahkan tidak bisa untuk berbelanja.

Jika kembali pada jurnal [7], terdapat beberapa faktor yang bisa mempengaruhi keputusan seseorang untuk berbelanja *online*, yaitu:

1. Kenyamanan. Faktor ini dapat meminimalisir terjadinya interaksi tatap muka dan tidak perlu berdesakdesakan ketika ingin berbelanja.
2. Kelengkapan Informasi. Informasi barang dapat dengan mudah diakses melalui *internet*. Selain itu sudah ada fitur seperti rating dan review agar kita dengan mudah melihat ulasan tentang kualitas dan informasi produk, kemudian dapat memesan dimana saja.

3. Kepercayaan Konsumen. Para pelaku usaha dapat meminimalisasi efek penyesalan dan kekecewaan pembelian dari pembeli dengan mengevaluasi serta memberi keamanan lebih terhadap barang yang ingin dikirim.
4. Efisiensi Biaya dan Waktu. Calon pembeli dapat dengan mudah berbelanja selama 24 jam dimana saja dan kapan saja.

Dengan banyaknya faktor-faktor yang dapat mempengaruhi seseorang untuk berbelanja *online* , dalam situasi pandemi Covid-19 yang sedang terjadi saat ini adalah suatu hal yang berpengaruh dari dampak pandemi Covid-19.

2.2 Dampak Pandemi Covid-19

Pandemi adalah sebuah epidemi yang menyebar ke beberapa negara atau benua, dan dapat menjangkiti banyak kalangan masyrakat. [9] Pandemi Covid-19 adalah *coronavirus* jenis baru yang ditemukan pada manusia dan mengakibatkan kejadian luar biasa terjadi pada Desember tahun 2019, yang memulai penyebaran pertamanya di daerah Wuhan, China. Kemudian, *virus* tersebut menyebabkan penyakit *Coronavirus Disease-2019* **(Covid-19)**. Pada pertengahan bulan Maret 2020 Covid-19 sudah mulai menyebar di Indonesia, dan pemerintah memberlakukan kebijakan PSBB (Pembatasan Sosial Berskala Besar) agar dapat menahan laju pergerakan *virus* tersebut. Akibatnya masyarakat melakukan semua kegiatannya di rumah atau dikenal sebagai *Work from Home (WFH)*. Lalu pada bulan Juni, pemerintah mulai memberlakukan *New Normal* dengan tetap mentaati protokol kesehatan, seperti menggunakan masker 3 lapis, mencuci tangan setelah menyentuh properti umum, membawa hand sanitizer, serta jaga jarak minimal 1,5

meter. Efek dari pandemi covid-19 memiliki banyak pengaruh dalam berbagai aspek, termasuk tingkah laku masyarakat sebagai pelaku konsumen. Dalam jurnal yang disusun oleh [10] memaparkan bahwa kebijakan PSBB memaksa pelaku usaha untuk merubah model bisnis menuju *online* sehingga berpengaruh positif dan signifikan terhadap perubahan perilaku konsumen [3]dalam berbelanja *online* , pada penelitian ini dampak pandemi Covid-19 dilakukan melalui situs media sosial *twitter*.

2.3 Situs Media Sosial *(Twitter)*

Twitter dikutip dari situs *wikipedia.com* layanan jejaring sosial dan mikroblog daring *(microblogging)* yang memungkinkan untuk setiap penggunanya terhubung untuk mengirimkan dan membaca pesan berbasis teks hinga 140 karakter.

Umumnya para pengguna/*user* dari media sosial *twitter* ini menggunakan situs media sosial ini untuk menyampaikan pendapat mengenai sesuatu peristiwa serta topik-topik yang sedang menjadi tren , isi hati , serta melakukan kegiatan bersosialisasi yang mempunyai bergagam media seperti teks, foto , maupun video yang bersifat positif ataupun negatif tergantung dari topik yang bersangkutan ataupun dibahas.

Mengingat penggunaan dan pemanfaatan media sosial *twitter* yang berevolusi, *twitter* yang sebelumnya memuat karakter berbentuk teks sejumlah 140 karakter, Jack Dorsey sebagai pendiri dari media sosial *twitter* ini mengambil keputusan pada tanggal 7 November 2017 untuk menambahkan kapasitas teks karakter hingga menjadi 280 karakter yang dikenal dengan sebutan kicauan(*tweet).*

Tweet itu sendiri bisa dapat diisi dengan pesan berupa karakter teks, *web link, mention(@)* yang dapat menjangkau user lain dalam suatu kicauan yang bersangkutan atau *hashtag (#)* yang berarti suatu *tweet* kita dapat terhubung dengan suatu topik tertentu di *twitter*.

Twitter itu sendiri merupakan media sosial yang bersifat satu arah, yang cara kerjanya adalah dimana suatu *user* diharuskan untuk mem-*follow* akun *twitter* kita untuk melihat semua kicauan kita di *twitter*. Untuk mendapatkan data dari *twitter* ini disaat kita membutuhkan suatu informasi yang sedang tren tersebut kita dapat meng-klik *hashtag* yang dicantumkan oleh suatu *user* dalam kicauan-nya serta menggunakan tool #TAGS berdasarkan kata kunci yang diinginkan jika ingin mencari informasi terkait.

Melakukan penelitian yang berbasi media sosial *twitter* ini memerlukan proses tahapan yang dilakukan terlebih dauhulu yaitu melalui tahapan *pre-processing noisy text* untuk mengubah kata menjadi bentuk yang sesuai standar **KBBI** (Kamus Besar Bahasa Indonesia) seperti yang dilakukan oleh penelitian sebelumnya yang dilakukan oleh [11] [12] [13]. Dalam melakukan penelitian ini media proses analisis menggunakan bahasa pemrograman *python*.

2.4 *Python*

Python salah satu bahasa pemograman yang diciptakan oleh seorang *programmer* berasal dari negara Belanda yang bernama Guido Van Rossum. Sedangkan menurut [14], *python* juga memiliki pendekatan yang cukup mudah tapi efektif untuk pemograman berbasis objek. *Python* memiliki keungulan terutama dalam perhitungan matematika. Selain dalam perhitungan matematika *Python*

mempunyai keunggulan yaitu bisa menggunakan variable tanpa harus dideklarasikan bahkan dalam menggunakan *python* tidak harus selali membuat *class* . *Python* saat ini sudah memiliki sampai pada versi 3 atau sering disebut *Python 3.* Dalam penerapan penelitian ini dilakukan dengan menggunakan modul-modul yang terdapat dalam bahasa pemogramana *python* untuk bahasa serta penggunaan *stopword* serta *corpus* atau kamus bahasa yang dimiliki modul *python* yang tergabung dalam *Natural Language ToolKit* yang akan dibahas pada sub-bab berikutnya.

2.5 *Natural Language ToolKit (NLTK)*

Natural Language Toolkit berdasarkan [15], pertama kali dibuat pada tahun 2001 di Universitas Pennsylvania. Modul-modul , dokumentasi serta informasi yang berkaitan dengan NLTK bisa mengunjungi pada *website* www.nltk.org

[16] berpendapat bahwa NLTK memiliki kumpulan tulisan (text corpora/corpus) yang sangat berguna dan digunakan secara luas oleh komunitas yang melakukan penelitian dengan NLP. 3 contoh corpora yang umum digunakan:

1. ***Brown Corpus*** Kumpulan tulisan dengan bahasa Inggris yang pertama kali digunakan. Kumpulan tulisan ini berisi kata-kata berbahasa Inggris Amerika dengan total 1 juta kata. 15 genre berbeda terdapat didalamnya, seperti fiksi, berita, keagamaan.
2. ***Gutenberg Corpus*** Kumpulan tulisan yang berisi 14 tulisan yang diambil dari Project Gutenberg (kumpulan e-book gratis terbesar). Kumpulan tulisan ini berisi kata-kata dengan total 1,7 juta kata.

3. ***Stopwords Corpus*** Corpus ini berisi kata-kata *stopwords* seperti yang sudah dijelaskan sebelumnya. Kumpulan tulisan ini berisi kata-kata dengan total 2400 *stopwords* yang terdiri dari 11 bahasa, salah satunya yaitu Inggris..

Dengan terdapatnya kumpulan bahasa kamus yang yang terdapat dalam modul NLTK tersebut sebagai modul pendukung dalam melakukan proses analisis berbasis teks tersebut atau yang akan dibahas dalam sub bab berikutnya yaitu *text classification*.

2.6 Sastrawi / *PySastrawi*

Sastrawi [17]adalah sebuah modul sederhana yang dimiliki oleh *library python* yang memungkinkan untuk melakukan pengurang kata-kata yang ter-infleksi dalam bahasa Indonesia ke bentuk baku-nya atau sesuai dengan standar kamus. Moduk *Python* ini merupakan bagian dari karya orisinal Sastrawi *project* yang dituliskan dalam bahasa pemrograman PHP. Penggunaan modul Sastrawi ini yang di *install* dalam *python* yang akan digunakan dalam proses *text classification* dan dapat dijelaskan dalam sub bab berikutnya *text classification.*

2.7 *Text Classification*

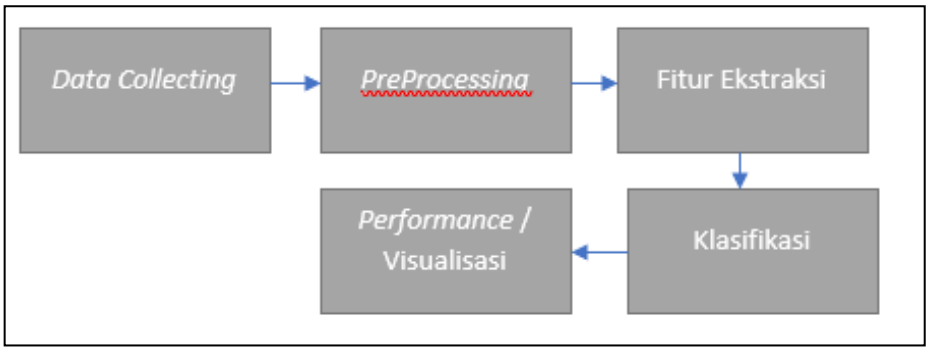
Semakin banyaknya jenis dokumen-dokumen elektronik dan informasi yang beragam berupa teks dari berbagai sumber, *Text Mining* merupakan salah satu cara yang berpotensi dilakukan untuk melakukan penelitian yang objeknya merupakan teks. Prinsip kerja *Text Mining* umumnya sama dengan cara kerja *Data Mining* hanya saja data yang diolah atau difokuskan adalah data berupa Teks. *Text Mining* menurut [18] , merupakan analisis teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakilkan isi dari dokumen sehingga dapat dilakukan Analisa keterhubungan ,

keterkaitan dan kelas antas dokumen. Dengan definisi lain *Text Mining* melingkupi sebuah proses ekstrasi informasi yang besar yang berasal dari sejumlah atau sekumpulan data yang terpola seperti dalam penelitian yang dilakukan dalam penelitian ini berupa kicauan *(tweet)* didalam media sosial *twitter*.

Text Classification adalah proses untuk mengelompokkan suatu dokumen yang besar ke dalam kategori atau kelas yang ditentukan. Sebagai contoh jika diberikan Da adalah sebuah dokumen dari sekumpulan dokumen (D) , dan {K1,K2,K3,K4,K5} merupakan kelas/kategori yang sudah ditentukan. Maka klasifikas teks adalah bertugas untuk mengkategorikan (Ca) dari masing-masing dokumen (Da) berdasarkan juga dari karakteristik yang dimiliki oleh setiap masing-masing kelas. Dalam penerapannya, Jika suatu setiap dokumen hanya dimasukkan kedalam salah satu kelas, maka klasifikasi tersebut dapat dinyatakan sebagai *single label.* Namun , jika suatu dokumen dimasukkan kedalam lebih dari satu kelas makan dapat dinyatakan sebagai *Multi Label.*

Dalam pemerosesannya menurut [19] klasifikasi teks memiliki proses yang sama dengan tahapan proses yang *data mining* dimana proses-proses tersebut mencakupi *document collecting , pre-processing ,* ekstrasi fitur , klasifikasi , dan validasi. *Document Collecting* merupakan sebuah proses tahapan pengumpulan data yang diperlukan , dalam penelitian ini adalah pengumpulan data kicauan *(tweets)*, tahapan yang dilakukan selanjutnya adalah *pre-processing* yang ditujukan untuk melakukan normalisasi data dan membuang karakter yang tidak diperlukan pada proses klasifikasi selanjutnya. Selanjutnya , adalah tahapan mengekstraksi data yang akan digunakan untuk mengkategorikan ciri-ciri data terkait untuk

dilakukan tahap berikutnya adalah *classifier* untuk mengukur tingkat akurasi klasifikasi dan model klasifikasi yang telah dihasilkan dalam memisahkan data kedalam masing-masing kelas serta tahapan selanjutnya adalah visualisasi untuk memastikan apakah adanya data yang menyimpang dalam klasifikasi. Tahapan proses yang dilakukan dapat dilihat melalui gambar sebagai berikut:



Gambar 2. 1 Tahapan Proses Pada Klasifikasi Teks

Sumber: (Vandana and C Namrata, 2013)

Pada tahapan klasifikasi ada 3 (tiga) teknik dasar yang umumnya digunakan dalam penerapannya, meliputi *unsupervised learning* , *supervised learning* dan *semi supervised learning*. Dalam *supervised learning* mempunyai ciri khas yang merupakan penentuan *output* kelas yang diharapkan sudah ditentukan di awal. Data yang digunakan juga dibagi menjadi dua kelas data yakni *Data Training* dan *Data Testing*. Sedangkan, pada Teknik *semi supervised learning* adalah *output* kelas yang diharapkan tidak ditentukan diawal, akan tetapi data yang digunakan untuk proses *training* adalah data berlabel, serta pada Teknik yang terakhir yaitu *unsupervised learning* ialah memiliki ciri *output* kelas yang tidak diketahui dan data yang digunakan pun tidak berlabel. Dalama melakukan *text classification* proses

yang dilakukan pertama dirangkum dalam sub bab berikutnya yaitu proses *sentiment analysis* yang mempunyai sekumpulan proses untuk mendapatkan hasil sentimen.

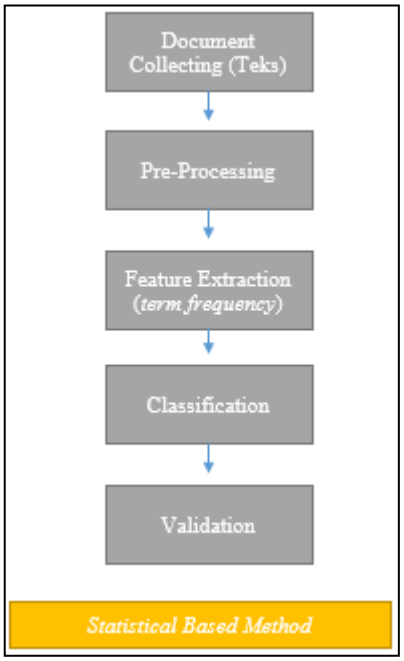
2.8 *Sentiment Analysis*

Perkembangan teknologi yang kian meningkat kini membuat *sentiment analysis* yang merupakan cabang dari *text mining* ini sedang hangat diperbincangkan. *Sentiment Analysis* pada penerapan umumnya sering digunakan untuk berbagai macam peristiwa yang terjadi pada masyarakat , sebagai contoh pada saat pilkada untuk pemimpin kepala daerah untuk mempelajari *sentiment* terhadap suatu pasangan calon pemimpin daerah ataupun bahkan untuk mempelajari opini masyarakat tentang suatu hal yang sedang hangat menjadi bahan pembicaraan di masyarakat , bahkan hasil *sentiment analysis* umumnya dapat dipakai untuk media pengambil keputusan terhadap topik yang diteliti itu sendiri.

Jika diartikan *sentiment analysis* itu sendiri mempunyai kata *"sentiment"* menurut KBBI (Kamus Besar Bahasa Indonesia) dapat diartikan sebagai pendapat atau pandangan yang didasarkan pada perasaan. Untuk arti yang lebih dalam jika merujuk pada pendapat para ahli, *Sentiment Analysis* adalah jenis *natural language* yaitu pengolahan kata untuk melacak *mood* masyarakat tentang produk atau topik tertentu [20]. Menurut pandangan ahli yang lain juga tentang *sentiment analysis* [5] , berpendapat bahwa tujuan *sentiment analysis* mengelompokkan teks (dokumen) yang mengandung opini sebagai *positive sentiment , negative sentiment ,* atau netral. Untuk menambahkan penjelasan tentang *sentiment analysis* mengenai apa yang diteliti dan menjadi dasar dari *sentiment analysis,* diambil dari jurnal yang

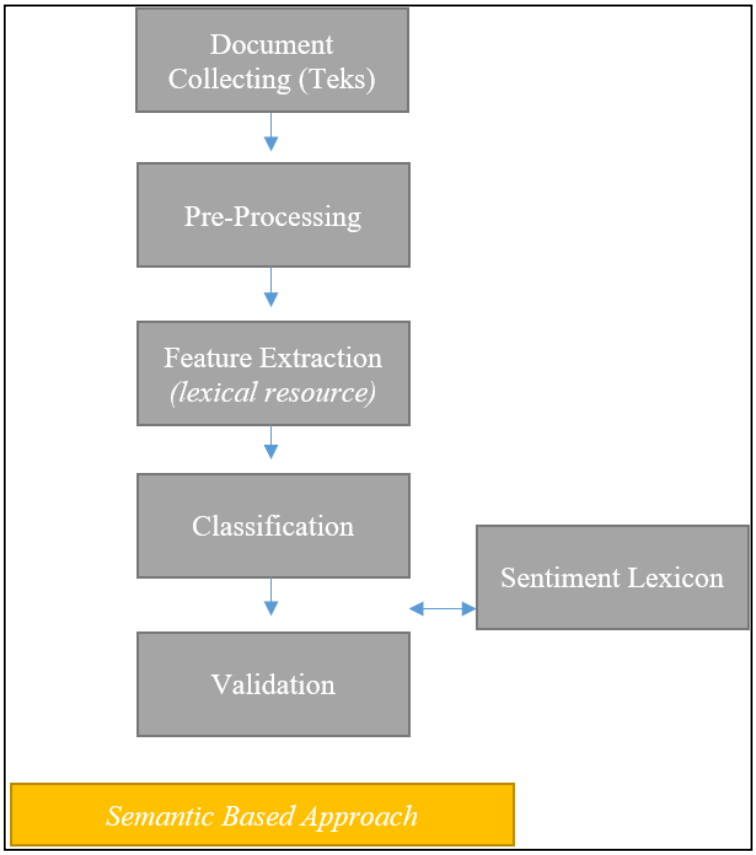
diteliti oleh [6], menyebutkan bahwa dalam *sentiment analysis* fitur yang dihasilkan dan algoritma klasifikasi menjadi poin utama. Dengan definisi lain , *sentiment analysis* adalah proses analisis yang dilakukan dari berbagai data (dokumen) yang umumnya berupa teks yang berisi pandangan atau opini sehingga dihasilkan kesimpulan dari berbagai opini yang telah dikumpulkan. Hasil dari *sentiment analysis* ini dapat dihasilkan berupa persentase *sentiment* positif , negatif ataupun netral. Beberapa pengguna dari proses *sentiment analysis* ini juga umumnya menginginkan keluaran yang dihasilkan berupa representasi visual dari data teks (opini) sehingga dihasilkan kesimpulan salah satu contoh yang umumnya dipakai yaitu *word plot.* Ada dua pendekatan (Algoritma) yang sering digunakan dalam penerapan *sentiment analysis* yakni , *Rule Based Method* dan *Statistical based method.* Pada umumnya pendekatan yang digunakan *rule based method* adalah pendekatan yang digunakan berbasis semantic (*semantic based)* , yang berbasis *lexical resource* untuk proses ekstrasi fitur. Sedangkan, *statistical based method* yang menggunakan perhitungan statistik secara otomatis. Pendekatan (Algoritma) statistik sudah umum banyak dan sering digunakan dalam penerapan *sentiment analysis* tetapi yang sedang hangat dilakukan dan dikembangkan adalah pendekatan semantik dengan harapan hasil keluaran akurasi yang tinggi lebih dari pendekatan statistik, tingginya akurasi pada pendekatan semantik dapat diperoleh dari makna kata untuk ekstraksi fitur, yang mana pada pendekatan (algoritma) ini *lexical resource* (*sentiment* leksikon) memegang peranan penting [21]**.** Sentimen leksikon mengambil peranan penting dalam mengetahui dan menganalisa serta untuk mendapatkan sebuah pola yang baik maupun pola yang buruk, semakin lengkap

sentiment lexicon yang digunakan maka hasil yang diperoleh akan semakin tepat. Sentimen Lexicon itu sendiri dapat berisi daftar kata-kata yang akan menentukan pola nya atau *polarity* , yakni *polarity* positif berisi seperti “baik”, “tenang” , “aman” ataupun jika *polarity* negatif seperti “lambat” , “buruk” , “Jelek”. Secara keseluruhan tahapan yang diterapkan pada pendekatan *sentiment analysis* ini memiliki tahapan proses yang hampir sama, jika pada pendekatan statistic *(Statistical based method)* tahapan proses yang dilakukan adalah *pre-processing* , *feature extraction* , *classification* ,dan *validation* , maka perbedaan yang terjadi pada pendekatan semantik *(Rule Based Method)* adalah pada tahapan proses *feature extraction.* Berikut adalah gambaran yang menyajikan tahapan proses pada masing-masing pendekatan:



Gambar 2. 2 Tahapan *Statistical Based Method*

Pada pendekatan statistic *(statistical based method)* yang menjadi peranan penting dan perbedaan adalah *feature extraction* yang menggunakan dan menfaatkan perhitungan matematis / statistic dalam hal ini seperti kemunculan kata (*term)* pada dokumen yang sering biasa disebut dengan istilah *term frequency* (TF) atau kemunculan kata *(term)* terhadap keseluruhan dokumen yang biasa disebut dengan istilah *term frequency – inverse document frequency* (TF-IDF). Sedangkan, pada pendekatan semantik *(semantic based approach)* yang digunakan jika menggunakan metode *Rule Based Method* dapat disajikan dengan gambar berikut:



Gambar 2. 3 Tahapan *Semantic Based Approach*

Pada pendekatan semantik yang menjadi ciri dan perbedaan dengan pendekatan statistikal adalah pada proses *feature extraction* yang memanfaatkan *lexical resource / sentiment lexicon* seperti yang dilihat pada gambar 2.3. Selanjutnya adalah bagian pertama dalam proses *sentiment analysis* yaitu dengan melakukan proses *preprocessing data* yang akan dijelaskan pada sub bab berikutnya.

2.8.1 *Preprocessing Data*

Pada proses tahapan *sentiment analysis* , proses *pre-processing* ialah tahapan proses yang harus dilalui sebelum melakukan langkah-langkah selanjutnya dalam menganalisa topik yang dianalisa , pada tahapan proses

ini merupakan proses dari ekstraksi teks, yang telah dilakukan dari *Document Collecting* yang bertujuan untuk mengubah data-data yang telah dikumpulkan terdapat data yang tidak ataupun belum terstruktur menjadi data yang lebih terstruktur agar bisa diolah menjadi lebih lanjut untuk proses klasifikasi. Tujuan dilakukannya hal ini adalah guna mempersiapkan dokumen yang ada berupa teks untuk menjadi data yang lebih baik dan mengurangi *noise* pada teks. Beberapa tahapan yang terdapat dalam *preprocessing* yakni, *tokenizing* (memotong kata), *formalization* (mengubah ke bentuk standar sesuak kamus), *translate, pos tagging, filtering* (membuat *stopword*), *stemming* (mengubah ke bentuk dasar) dengan selesainya tahapan proses *preprocessing* yang dilakukan tersebut maka data tersebut akan dapat dilanjutkkan untuk proses berikutnya yaitu klasifikasi.

2.8.2 *Tokenizing* (*Cleansing* Data)

Tahapan pada proses ini adalah proses *tokenizing* atau sering disebut dengan *cleansing data* , pada proses *cleansing data* ini dilakukan dengan tujuan untuk membersihkan *noise* atau data dari karakter-karakter yang tidak berpengaruh dalam pemerosesan klasifikasi dan membuang adanya *redundancy data* atau disebut dengan duplikasi data. Pada proses ini dilakukan dengan cara memenggalkan setiap kata yang Menyusun sebuah dokumen/kalimat yang diteliti , hal ini biasa disebut dengan *tokenizing.* Hasil dari proses *tokenizing* ini adalah daftar kata-kata yang berdiri sendiri yang Menyusun kalimat dalam sebuah dokumen.

Proses *tokenizing* ini sangat penting dilakukan didalam proses *preprocessing* dikarenakan data yang dimiliki oleh media sosial *twitter* banyak mengandung *noise* seperti jenis singkatan , pengunaan angka yang menggantikan huruf serta dapat juga kombinasi antara huruf dan angka yang menyerupai kalimat, *mention* (@) , *hashtag* (#) , kode html (http:/www) dan memposting ulang *tweet* orang lain dengan tujuan untuk membagikan kembali informasi yang terdapat dalam suatu kicauan (*tweet*) tersebut (*Retweet).* Mengingat dapat kemungkinan terjadinya duplikasi data pada saat melakukan klasifikasi maka diperlukan *retweet removal* sehingga data *tweet* yang akan digunakan untuk proses klasifikasi adalah data yang dibutuhkan tanpa ada informasi yang terduplikasi dan lebih akurat.

Setelah melakukan proses mehilangkan duplukasi *tweet* yang ada tahapan yang perlu dilakukan selanjutnya adalah untuk menghilangkan *noise* yaitu membuang *mention, hashtag,* dan *html* yang terdapat pada kicauan *(tweet).* Menghilangkan ketiga elemen tersebut diperlukan karena ketiga elemen tersebut tidak diperlukan dan tidak berpengaruh dalam proses klasifikasi, selanjutnya adalah menghapus angka yang terdapat pada data *tweet,* lalu hanya karakter huruf sajalah yang dipakai untuk proses selanjutnya, setelah dihilangkan angka langkah yang terakhir adalah mengubahnya menjadi karakter atau huruf kecil yang setiap kata umumnya dipisahkan oleh yang biasanya disebut *delimiter* seperti titik, atau tanda koma, setelah itu lakukan pengecekan persamaan kalimat yang tersimpan didalam *database,* jika ada maka data yang sama tersebut akan ditolak untuk

menghindari duplikasi data pada saat proses selanjutnya dilakukan tahapan *formalization* yang akan dijelaskan dalam sub bab berikutnya.

2.8.3 *Formalization*

Pada tahapan ini adalah proses *formalization* yang mempunyai tujuan untuk mengubah kata yang tidak standar atau diluar ketentuan yang tertera secara resmi ke bentuk yang lebih formal atau sesuai dengan standar struktur KBBI (Kamus Besar Bahasa Indonesia). Dalam media sosial *twitter* sering dijumpai dalam kicauan *(tweet)* yang suarakan pengguna *twitter* penulisan yang tidak sesuai struktur KBBI dengan penggunaan singkatan yang tidak baku , menggunakan Bahasa lokal ataupun daerah masing-masing. Mengganti huruf dengan angka serta kombinasi lainnya , seperti contoh (“ber2 denganmu”,”bapak2”,”hati2”) serta juga menambahkan karakter huruf untuk suatu kata tertentu yang umumnya untuk memberikan ekspresi emosi yang lebih dalam (misal: “belanja di shopee untuuuuung banget”). Dikarenakan hal tersebut diperlukannya penentuan daftar kamus yang ditentukan untuk mengetahui dan mengubahnya sesuai kaidah struktur bentuk standar kamus yang pada dalam hal ini yaitu menggunakan kamus yang terdapat dalam *nltk.corpus* Sastrawi, setelah mendapatkan hasil dari *formalization* selanjutnya akan dilakukan tahap *filtering* yang akan dijelaskan pada sub bab berikutnya.

2.8.4 *Filtering*

Tahap selanjutnya yang dilakukan setelah melakukan proses pemenggalan kata sehingga dihasilkanlah daftar kata penyusun dokumen, dilakukanlah tahapan *filtering* yang bertujuan untuk proses pengambilan kata-kata yang dianggap penting untuk digunakan pada proses klasifikasi.

Pada tahap proses *filtering* ini digunakan algoritma *stoplist* atau yang sering disebut sebagai *stopword,* yaitu mengambil kata-kata yang penting dan membuang kata yang tidak mendukung proses selanjutnya klasifikasi yang menjadi penciri khas suatu dokumen. *Stopword* adalah sekumpulan kata-kata yang tidak deskriptif dan tidak memiliki pengaruh jika dihilangkan , seperti kata “saya” , “yang” , “seringkali” , “tersebut” , kata-kata tersebut dapat dihilangkan karena mengandung kata sambung, partikel dan preposisi. Seringkali ditemukan dalam sebuah dokumen terkait yang dilakukan dalam penelitian kata-kata *stopword* ini lebih sering muncul dibandingkan dengan kata penciri khas dokumen. Sehingga, diperlukan pembuangan kata-kata *stopword* ini agar tidak keluar sebagai kata yang mewakili dokumen dalam tahap proses klasifikasi. Berikut adalah contoh beberapa kata *stopword* yang diambil dari jurnal [22]**:**

Tabel 2. 1 *Stopword* Fadilla Z Tala

Stopword Tala				
Yang	**Di**	**Dan**	**Itu**	**Dengan**
Untuk	**Tidak**	**Ini**	**Dari**	**Dalam**
Akan	**Pada**	**Juga**	**Saya**	**Ke**
Karena	**Bisa**	**Tersebut**	**Ada**	**Mereka**
Lebih	**Kata**	**Tahun**	**Sudah**	**Atau**
Saat	**Oleh**	**Menjadi**	**Orang**	**Dia**

Telah	Adalah	Sebagai	Seperti	bahwa

2.8.5 *Stemming*

Tahapan proses ini juga perlu dilakukan karena untuk mempersempit kembali kata-kata sehingga dihasilkan kata dasar yang akan mempunyai pengaruh pada proses klasifikasi. Proses *stemming* ini adalah tahapan yang berguna dan bertujuan untuk mengubah kata yang terdapat dalam dokumen ke bentuk dasarnya sesuai kamus besar Bahasa Indonesia (KBBI) dengan cara membuang imbuhan yang terdapat didalam kata. Tujuan utama dari *stemming* ini adalah mengurangi kata-kata yang bermakna jamak dan kata tunggal. Karena kata berimbuhan dikembalikan ke bentuk dasarnya dan dianggap memiliki makna yang sama, maka proses *stemming* ini bermanfaat untuk mengurangi dimensi dari data teks. Jika merujuk kepada algoritma yang sering digunakan untuk proses *stemming* untuk teks Bahasa Indonesia adalah algoritma **Nazief dan Andriani** dengan umumnya kata dasar Bahasa Indonesia dapat dilihat dari kombinasi ini;

Prefiks 1 + Prefiks 2 + Kata Dasar + Sufiks 3 + Sufiks 2 + Sufiks 1

Keterangan:

Prefiks = Awalan
Sufiks = Akhiran
Infiks = Sisipan
Confiks = Awalan dan Akhiran

Menurut algoritma **Nazief dan Andrani** juga menyebutkan ada beberapa daftar awalan dan akhiran yang tidak diijinkan untuk digunakan Bersama-sama. Pada penelitian ini menggunakan fasilitas *stemming* yang terdapat dalam kamus yang tergabung dalam *nltk.corpus* [16] yang terdapat dalam modul *python* dimana modul *python* Sastrawi yang berperan dalam proses *stemming* atau sebagai *stemmer* dengan memberikan *input.* proses selanjutnya akan dilakukan proses ekstraksi fitur yang akan dijelaskan pada sub bab berikutnya.

2.8.6 *Feature Extraction* / Proses Ekstraksi Fitur

Pada proses tahapan ini adalah proses yang dilakukan setelah dilakukannya proses *preprocessing* yang bertujuan untuk mengekstraksi fitur yang diperlukan yaitu pengambilan ciri khas dari sebuah objek yang akan digunakan untuk proses selanjutnya yaitu klasifikasi. Dalam pemerosesan ekstraksi fitur ini sangat penting terutama dalam *text classification* karena dari proses ini akan diambil fitur yang tepat untuk kasus yang bersangkutan dan relevan dengan topik klasifikasinya. Misalnya, pada kasus klasifikasi tentang *website* yang mengandung konten memasak , fitur yang diekstraksi dan dianggap relevan adalah kata (*term*) yang mengandung makna konotasi memasak adalah seperti “asin” , “matang” , “enak”, sedangkan jika pada kasus klasifikasi *sentiment* pada teks terhadap *website* belanja *online* fitur yang relevan dan dapat diambil adalah seperti “bagus”, “murah” , “buruk” kata-kata tersebut mengandung sentimen/opini.

Penerapan klasifikasi teks terdapat dua pendekatan yang umumnya dan dapat digunakan untuk ekstraksi fitur yakni , pertama adalah *statiscial based method* merupakan pendekatan berbasis statistik, yang pembobotan fiturnya menggunakan perhitungan matematis dan statistik. Proses ini menggunakan cara kerja yaitu melakukan penghitungan frekuensi kemunculan kata (*term*) yang biasanya disebut *term frequency* (TF) yang dianggap sebagai ciri ataupun fitur yang mewakili suatu dokumen dan *term frequency inverse document frequency* dimana frekuensi kemunculan kata (TF) pada sebuah dokumen juga dibandingkan dengan kemunculan kata di keseluruhan dokumen yang berarti jika banyak kemunculan suatu kata didokumen tetapi muncul juga banyak di keseluruhan dokumen maka kata tersebut tidak dijadikan sebagai kata penciri khas nya atau dengan arti lain, kata tersebut merupakan kata umum dan tidak bisa mewakili sebuah dokumen.

Pada pendekatan kedua adalah pendekatan yang berbasis semantik (*semantic based approach*) , dimana proses ekstraksi fitur dimanfaatkan dengan makna dari kata penyusun kalimat. Pendekatan ini dalam *sentiment analysis* yang menjadi fitur untuk diekstraksi adalan kalimat yang mengandung sentimen/opini baik yang mengandung positif ataupun negatif , *sentiment* leksikon menjadi sangat penting untuk menemukan kata (*term*) yang menjadi fitur dari kelas yang telah ditentukan. Menurut [5], sebuah kata dapat dikatakan ataupun mengandung sentimen , jika kata tersebut

mengekspresikan perasaan yang diinginkan (*sentiment* positif) maupun tidak diinginkan (sentimen negatif).

2.8.7 Klasifikasi / Metode

Pada tahapan proses ini adalah melakukan klasifikasi yang merupakan proses untuk menemukan model yang tepat untuk membedakan kelas yang satu dengan kelas yang lain, dengan harapan model yang dipakai untuk dapat memprediksi kelas dari objek yang belum diketahui kelasnya. Dalam menyelesaikan problem *sentiment analysis* atau juga sering disebut *opinion mining* metode klasifikasi yang dipakai adalah *Rule Based Method* yang berbasiskan pada pendekatan semantik dan *Statistical based Method* yang berbasiskan pada statistik dan memanfaatkan perhitungan matematis atau sering disebut juga *machine learning* ataupun mengombinasikan keduanya. metode statistikal sering juga dipakai dalam klasifikasi teks , begitu pula dalam *sentiment analysis* , beberapa metode statistikal yang sering dipakai dalam menyelesaikan klasifikasi yakni naïve bayes, decision tree, support vector machine (SVM), maximum enthropy, dan lain sebagainya tergantung peneliti ingin mengunakan metode yang tepat untuk topiknya dikarenakan setiap metode mempunyai karakteristik dan teknik yang berbeda-beda dalam memisahkan data sesuai kelasnya.

Sedangkan, jika dibandingkan dengan *rule based method* yang menggunakan pendekatan semantik menggunakan aturan yang sudah ditentukan dengan yang sesuai dengan label kelas seperti “if … then … dalam proses klasifikasi datanya. Pendekatan *rule based method* sangat

cocok digunakan jika data diaplikasikan dalam klasifikasi sederhana dan data yang tidak kompleks dan bervariasi.

2.8.8 *Lexical Resource*

Pendekatan dalam metode klasifikasi yang berbasis semantik (*rule based approach*) ini dimanfaatkan untuk mengumpulkan *sentiment* leksikon yang mempunyai beberapa metode yakni, *manual approach* , *dictionary based approach* dan *corpus based approach* . pendekatan *dictionary based approach* adalah pendekatan yang berbasiskan kamus dengan memanfaatkan relasi antar kata (*synset)* seperti sinonim , antonim , hipernim dan hiponim yang digunakan di *wordnet* untuk memperoleh kata yang mengandung opini *opinionword* lainnya. Sedangkan pada pendekatan *corpus based approach* memanfaatkan kumpulan *opinionword* sebagai benih dan pola sintaksis dari benih kata untuk menambang *opinionword* pada corpus yang besar. Pada penelitian ini digunakan *corpus based approach* yang juga bisa dilakukan dengan cara menerapkan *rule base* pada metode klasifikasi-nya.

2.9 Sentimen Leksikon Indonesia

Sentimen leksikon umumnya adalah yang paling penting dalam *sentiment analysis* yang menggunakan pendekatan semantik. Sentimen leksikon berisi daftar kata-kata dengan kecenderungan yang mengandung sentimen positif atau negatif. Kata -kata yang mengandung Seperti kata “baik”, “bagus”, “aman” , “mantap” memiliki kecenderungan kata yang mengandung sentimen positif, sedangkan kata

“buruk”, “lambat”, dan “ceroboh” memiliki kecenderungan kata yang mengandung sentimen negatif.

Perkembangan *sentiment* leksikon Indonesia sangat berbeda dengan sentimen leksikon Inggris yang sudah beredar luas secara bebas dan bersifat publik (*free*) secara *online*. Sentimen leksikon Indonesia masih sangat terbatas jumlahnya masih dalam perkembangan dan bahkan ada yang belum beredar secara bebas publik. Peneliti terdahulu [11] membangun sentimen leksikon Indonesia menggunakan pendekatan manual (*manual approach*), sedangkan Vania dkk [21] membangun *sentiment* leksikon Indonesia menggunakan pendekatan corpus (*corpus based approach*). Vania dkk [21] mengekstraksi kumpulan *opinion word* sebagai benih kata (*feed seed*) lalu pergi menggunakan hasil *translate synset* dari *sentiwordnet* yang memiliki *sentiment score* lebih besar dari 0.7 dan sentimen leksikon dari *opinion finder* yang memiliki *subjectivity score* tinggi. Selanjutnya benih (feed seed) digunakan untuk menemukan pola (pattern) dan struktur kalimat yang mengandung kata benih tersebut dan memiliki sentimen yang sama dengan sentimen benih. Selanjutnya, Vania menggunakan kata benih dan pola dari kalimat untuk mendapatkan *opinion word* lainnya di corpus dalam ukuran besar dengan topik tertentu. Data yang telah dilakukan *preprocessing data* akan kemudian dilakukan analisis dalam konteks penelitian ini dilakukan proses analisis dengan metode *Rule Based Method* yang akan dijelaskan pada sub bab berikutnya.

2.10 *Rule Based Method*

Rule based method adalah metode klasifikasi yang memanfaatkan aturan - aturan (*rule*) untuk membedakan kelas yang saatu dengan kelas yang lain. *Rule*

dibuat berdasarkan karakter dari masing-masing kelas dan dinotasikan dalam bentuk “IF …(kondisi)… THEN …(solusi)…”. Dimana “IF” merupakan kondisi prasyarat (*rule antecedant*) yang terdiri dari satu atau lebih atribut tes, dimana tesnya bersifat logika. Sedangkan “THEN” merupakan konsekuen (*rule consequent*) yang berisi hasil prediksi kelas. Misalnya untuk memisahkan data ke dalam kelas positif, negatif dan netral *rule* yang bisa digunakan adalah :

– Jika jumlah frekuensi kemunculan kata bersentimen positif lebih banyak dari kata bersentimen negatif, maka data digolongkan sebagai kelas positif.

– Jika jumlah frekuensi kemunculan kata bersentimen negatif lebih banyak dari kata bersentimen positif, maka data digolongkan sebagai kelas negatif.

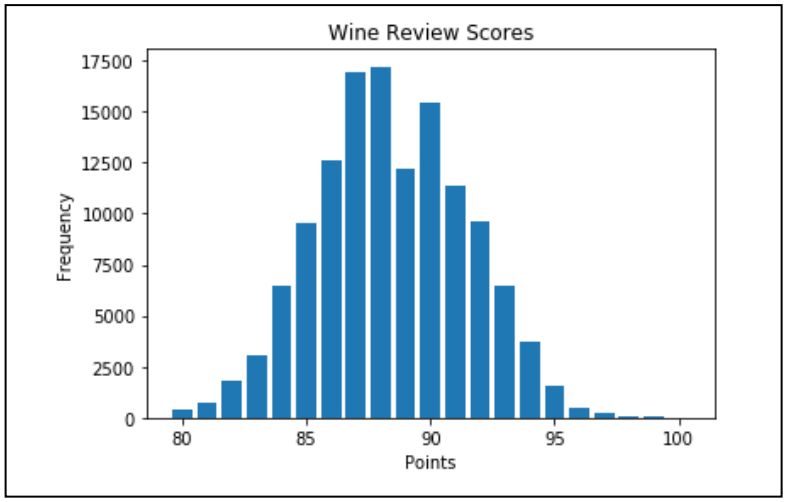
– Jika jumlah frekuensi kemunculan kata bersentimen positif sama dengan kata bersentimen negatif, maka data digolongkan sebagai kelas netral.

– Jika tidak ditemukan kata bersentimen positif maupun kata bersentimen negatif pada data, maka digolongkan sebagai kelas netral.

Aturan *(rule*) yang digunakan bisa bersifat *mutually exclusive* atau *exhaustive*. *Mutually exclusive* berarti *classifier* mengandung aturan-aturan yang bersifat independen satu sama lain, sedangkan *exhaustive* berarti classifier mengandung aturan-aturan yang mencatat setiap kemungkinan kombinasi nilai atribut. Dimana setiap record hanya boleh dilingkupi paling banyak satu aturan saja. Hasil analisis menggunakan *Rule based method* akan dilakukan visualisasi untuk mengetahui penggambaran dengan beragam visualisasi.

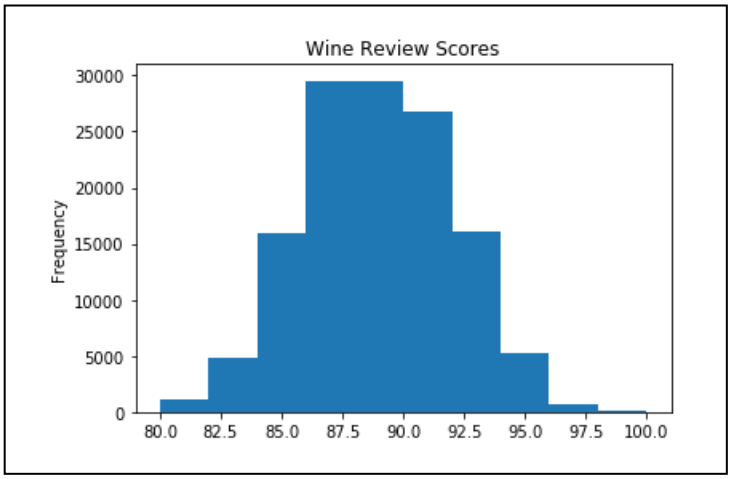
2.11 Visualisasi

Tahapan ini merupakan tahap terakhir yang cukup penting bertujuan untuk menganalisa hasil *sentiment analysis* berapakah persentase dari hasil *sentiment analysis* serta menyimpulkan bagaimana untuk menjelaskan hasil *sentiment analysis* menggunakan visualisasi dari data hasil *sentiment analysis*. Visualisasi dapat mempunyai beragam jenis [23]berupa *bar chart , histogram , scatter plot* dan macam banyak jenis lainnya sesuai kebutuhan. Berikut adalah contoh jenis-jenis visualisasi dirujuk secara ber-urutan dengan gambar:



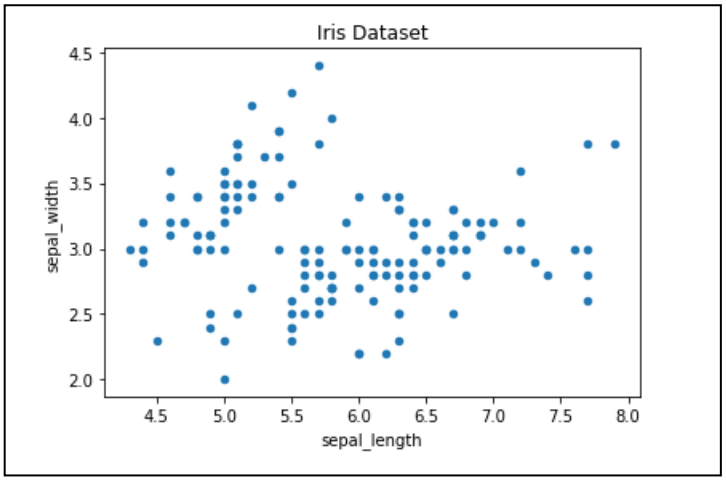
Gambar 2. 4 *Bar Chart*

Sumber: (*Towards Data Science, 2019*)



Gambar 2. 5 *Histogram*

Sumber: (*Towards Data Science, 2019*)



Gambar 2. 6 *Scatter Plot*

Sumber: (*Towards Data Science, 2019*)

2.12 Penelitian Terdahulu

Tabel 2. 2 Penelitian Terdahulu

No.	Nama Peneliti	Judul Penelitian	Nama Jurnal	Hasil Penelitian
1.	Feby Tri Saputra, Yani Nurhadryani, Sony Hartono Wijaya, Defina	Analisis Sentimen Bahasa Indonesia pada *Twitter* Menggunakan Struktur *Tree Berbasis* Leksikon	Jurnal Teknologi Informasi dan Ilmu Komputer, [S.l.], v. 8, n. 1, p. 135-146, peb. 2021. ISSN 2528-6579.	Twitter tersebar luas sehingga tidak mungkin membaca semua opini untuk mendapatkan seluruh sentimen. pendekatan dalam analisis sentimen ini adalah berbasis leksikon yang termasuk *rule-base approach*Metode berbasis *tree* diujikan pada data dengan lintas topik seperti data twit Pilgub Jabar 2018, Pilpres 2019, dan pandemik COVID-19. Ketiga data uji memiliki proporsi kelas yang tidak seimbang, dengan kelas terbanyak merupakan kelas positif. Metode berbasis *tree* menghasilkan akurasi sebesar 64,97% (meningkat 1,26%) pada data Pilgub Jabar 2018, 64,33% (meningkat 11,41%) pada data Pilpres 2019, dan 66,24% (meningkat 7,61%) pada data pandemik COVID-19. Metode berbasis *tree* dapat menghasilkan akurasi yang stabil pada beberapa lintas topik dibuktikan dengan standar deviasi akurasi yang kecil (0,97%) bahkan lebih kecil dari metode tanpa *tree* (5,4%). Metode berbasis *tree* dapat meningkatkan *weighted f1-measure* pada data Pilpres 2019 sebesar 10,45% dan

				data pandemik COVID-19 sebesar 8,1%, sedangkan hasil pada data Pilgub 2018 tidak berbeda secara signifikan. Hasil akurasi dan *weighted f1-measure* memiliki selisih yang kecil sehingga pengukuran akurasi valid dan tidak bias terhadap data tidak seimbang
2.	Sigit Suryono, Ema Utami, Emha Taufiq Luthfi	Klasifikasi Sentimen Pada *Twitter* Dengan *Naive Bayer Classifier*	Jurnal Ilmiah Bidang Teknologi, ANGKASA Volume 10, No. 1, Mei 2018	Dalam tulisan ini, tweet yang berhubungan dengan kata kunci yang dicari dihimpun dengan menggunakan tools yaitu API Twitter. Data yang didapat dari proses penghimpunan akan diolah dengan menggunakan Natural Language Toolkit yang berjalan diatas bahasa pemrograman Python. dengan menggunakan Naive Bayes untuk melihat sentimen yang dihasilkan. Dari proses klasifikasi yang telah dilakukan akan diukur tingkat akurasi. Dari hasil uji coba sebanyak 3 kali, didapatkan tingkat akurasi pada percobaan pertama 64.95%, kedua

				66.36% dan ketiga 66.79% Hasil lain yang didapatkan dari proses klasifikasi yaitu Sentimen positif 28% sentimen negatif 20% dan sentimen netral 52%
No.	**Nama Peneliti**	**Judul Penelitian**	**Nama Jurnal**	**Hasil Penelitian**
3.	I Made Artha Agastya	Pengaruh *stemmer* Bahasa Indonesia Terhadap Performa Analisis Sentimen Terjemahan Ulasan Film	Jurnal TEKNO KOMPAK, Vol. 12, No. 1, 2018, 18-23. ISSN 1412-9663	Untuk mendapatkan pengaruh dari stemming terhadap analisis sentimen maka dilakukan percobaan dengan dataset ulasan film yang sudah diterjemahkan ke Bahasa Indonesia. Stemmer Sastrawi sebagai algoritma stemming terbaru digunakan pada penelitian ini. Dataset dibagi menjadi 5 (lima) kategori yang mana 100 data, 250 data, 500 data, 750 data, dan 1000 data. Hasil yang diperoleh menunjukan bahwa stemmer tidak memberikan peningkatan

				akurasi yang stabil.
No.	**Nama Peneliti**	**Judul Penelitian**	**Nama Jurnal**	**Hasil Penelitian**
4.	Calvin, Johan Setiawan	*Using Text Mining to Analyze Mobile Phone Provider Service Quality*	*International Journal of Machine Learning and Computing, Vol. 4, No. 1, February 2014* *106*	Penelitian yang ini bertujuan untuk mengukur tingkat kepuasan pelanggan terhadap penyedia layanan *operator* dari situs media sosial *twitter* dengan objek penelitian akun media sosial resmi masing- masing *provier* PT. XL Axiata, PT Telkomsel, PT Indosat menggunakan *text mining* yang berbasiskan metode *Naïve Bayes model* dengan hasil keluaran analisa PT. XL Axiata mendapatkan skor 29, PT. Telkomsel mendapatkan skor -70, PT. Indosat mendapatkan skor -100. Dari hasil yang didapat menunjukkan bahwa perusahaan penyedia jasa telepon pasti sudah memiliki banyak pengguna, namun mungkin masih belum

				mengetahui kualitas yang mereka berikan kepada pelanggannya. Biasanya setiap pendapat yang diajukan oleh pengguna diabaikan oleh perusahaan. Dengan mengolah opini yang telah disampaikan menggunakan text mining, tulisan ini telah menunjukkan kualitas pelayanan dari masing-masing perusahaan.
No.	**Nama Peneliti**	**Judul Penelitian**	**Nama Jurnal**	**Hasil Penelitian**
5.	Joviano Siahaan, Wella, Ririn Ikana Desanti	Apakah Youtuber Indonesia Kena Bully Netizen?	Ultima InfoSys : Jurnal Ilmu Sistem Informasi, 11(2), 130-134.	Penelitian yang dilakukan bertujuan untuk mempelajari tentang *cyberbullying* yang terjadi dalam objek penelitian ini adalah 10 *Youtuber Instagram* Indonesia pada post mereka, penelitian ini adalah penelitian *text mining* yang menggunakan pendekata *Support Vector Machine* (SVM) dengan hasil analisis

				menggunakan model SVM dengan akurasi 81,2% adalah 49,524% dari komentar di Bagian komentar Youtuber Indonesia dipertimbangkan sebagai cyberbullying.

Dalam penelitian Ini [24]menggunakan metode *rule-based approach* untuk referensi pengerjaan dalam topik minat belanja online pada masa pandemi COVID-19 Referensi pada jurnal [25]dalam penggunaan *twitter API* dalam *data collection* dengan *natural language toolkit NLTK* pada penelitian mengenai topik minat belanja *online* dalam penerapannnya dengan menggunakan algoritma bahasa pemrograman *Python* dengan media aplikasi *Anaconda Navigator* dengan tipe pemilihan *Jupyter Notebook* dimana *natural language toolkit* dapat diterapkan dalam penerapan *sentiment analysis* dengan harapan seperti apa hasil sentimen berupa positif, netral ataupun negatif. Dalam penerapannya *nltk* tersebut juga dibantu oleh modul yang dimiliki bahasa pemrograman *python* yaitu yang dipakai dalam penelitian ini adalah sebuah modul yang memiliki *lexical resources* atau kamus yang berisi kata-kata dalam bahasa indonesia serta *stopwords* dalam bahasa indonesia ialah Modul sastrawi.Penerapan Modul *Sastrawi* [26] pada penelitian terdahulu ini meneliti tentang bagaimana pengaruh modul *stemmer* Sastrawi dalam *dataset* ulasan film yang menghasilkan sentimen hal ini sehubungan sebagai

referensi dalam penelitian ini yang objek penelitiaanya adalah data *tweet* pengguna situs media sosial *twitter* dengan menggunakan modul sastrawi yang digabungkan dengan metode *Rule-Based Approach* mengenai topik minat belanja *online* pada masa pandemic COVID19 Referensi *Text Mining* dalam penelitian ini [27]dengan kontekstual penelitian hasil sentimen positif negatif ataupun netral. Penggunaan referensi [28]menggunakan situs media sosial *sentiment analysis* pada *youtuber*, pada konteks penelitian ini adalah penelitian dilakukan melalui situs media sosial *twitter* dengan meneliti data *tweet* minat mengenai minat belanja *online* pada masa pandemi COVID-19.