BAB II

TINJAUAN PUSTAKA

2.1. Penelitian Sebelumnya

Pada Penelitian kali ini, dilakukan studi literatur dengan permasalahan yang sedang dikaji baik berupa buku, jurnal, maupun dokumen yang relevan. Bab ini merupakan rangkuman dari penelitian – penelitian terdahulu yang membahas tentang topik yang berkaitan dengan penelitian yang akan dilakukan.

1. ***Aspect Based Sentiment Analysis Pada Review Produk Kecantikan Menggunakan Extreme Gradient Boosting (Jessie Gabriella Silalahi : 2021) [15]***

   Penelitian ini bertujuan untuk melakukan analisa serta melakukan implementasi algoritma Extreme Gradient Boosting dalam melakukan klasifikasi sentimen terhadap aspek pada *review* produk kecantikan. Penggunaan data pada penelitian ini menggunakan data kumpulan *review* produk kecantikan dengan metode crawling data bersumber dari website Female Daily. Data tersebut dikumpulkan dari beberapa kategori produk seperti sun protection, serum, toner, scrub, parfume dan lip product berjumlah 1500 data.

   Metode yang digunakan dalam penelitian ini adalah Extreme Gradient Boosting. Hasil pada penelitian ini menunjukan nilai yang cukup tinggi, dengan metode yang digunakan untuk melakukan Evaluasi model yaitu confusion matrix dengan output atau keluaran menjadi 4 hal yaitu precision, recall, f1-Score, dan accuracy. Hasil yang didapatkan pada nilai rata-rata accuracy dari empat aspek pada penelitian ini yaitu 90% [15].

2. ***Implementasi Algoritma Multinomial Naïve Bayes Untuk Analisis Sentimen Menggunakan TF-IDF dan N-Gram (Prisilia Ines : 2020) [16]***

   Penelitian ini bertujuan untuk melakukan implementasi algoritma Multinomial Naïve Bayes menggunakan TF-IDF dan N-Gram pada analisis

sentimen terhadap user *review* dari penelitian Endang Pamungkas sebanyak 554 data. Dataset nanti nya dibagi dengan rasio 80:20 dan terbagi menjadi 3 kategori yaitu positif, negatif, dan netral.

Hasil implementasi serta uji coba sistem, didapatkan performa terbaik nilai n yaitu dengan menggunakan unigram dengan ratio 80:20 pada *training* dan *testing* model yang dibuat. Kemudian model dilakukan evaluasi dengan menggunakan metode confusion matrix dengan nilai 84% untuk accuracy, precision, recall, dan f1-score [16].

3. ***Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes (Clarisa Hasya Yutika, Adiwijaya, Said Al Faraby : 2021) [17]***

Penelitian ini bertujuan untuk melakukan analisis sentimen berbasis aspek terhadap *review* pada website *femaledaily*. Pada penelitian ini juga dilakukan perbandingan antara model yang menggunakan tahapan *foreign word translation*, serta perbandingan antara data yang dilakukan *preprocessing stopword removal* dengan yang tidak menggunakan *stopword removal*. Dataset yang digunakan tentunya berasal dari website *femaledaily* dengan jumlah total data sebanyak 5054 *review* dengan kategori *serum, toner, scrub, sunscreen, dan exfoliator*. Kemudian aspek label yang ditentukan yaitu harga, kemasan, produk, serta aroma untuk mendapatkan kategori positif, negatif, dan netral.

Metode yang digunakan pada penelitian kali ini yaitu dengan *naïve bayes* classifier. Pemilihan metode dikarenakan *naïve bayes* classifier dikenal sebagai metode yang sederhana namun efisien, serta memiliki asumsi yang sangat kuat terhadap independensi dari masing-masing kondisi, terutama ketika dihadapkan dengan data train yang sedikit. Penelitian ini mendapatkan hasil performansi tertinggi oleh dataset yang diterjemahkan ke dalam Bahasa Inggris kemudian diterjemahkan ke Bahasa Indonesia dan tidak menggunakan stopword removal dengan parameter alpha atau smoothing sebesar 1, min_df sebesar 0,01, max_df sebesar 0,7, dan

max_features sebesar 2000 menghasilkan performansi terbaik dengan nilai F1-Score sebesar 62,81% [17]

4. ***Analisis Sentimen Pada Data Ulasan Twitter dengan Menggunakan Long Short Term Memory (Sharfina Febbi Handayani, Riszki Wijayatun Pratiwi, Mulyana Putriyani : 2021) [18]***

Penelitian ini bertujuan untuk melakukan analisis sentimen ulasan twitter dengan menggunakan metode Long Short Term Memory dan pengaplikasian word2vec untuk merepresentasikan kata yang terdistribusi. Analisis yang dilakukan kemudian akan dibagi menjadi tiga kelas emosi yaitu positif, negatif, dan netral. Data yang digunakan pada penelitian ini berjumlah 10806 tweet yang diperoleh dari penelitian *"Dataset Indonesia untuk Analisis Sentimen"* [19].

Metode yang digunakan pada penelitian ini yaitu dengan menggunakan Long Short Term Memory yang merupakan varian dari *Recurrent Neural Network*. Pemilihan metode dikarenakan dapat mengingat informasi jangka panjang dengan menggunakan sel LSTM untuk menyimpan informasi terlebih dahulu. Sebelum dilakukan *modelling*, data akan melalui proses ekstraksi dengan menggunakan Word2vec dan kemudian hasil *modelling* akan di validasi dengan menggunakan confusion matrix. Dari penelitian yang dilakukan didapatkan hasil terbaik dengan parameter word2vec 57.15% menggunakan CBOW, jumlah neuron 150 dengan waktu 2 menit 50 detik dan akurasi 57.35%, jumlah epoch 30 dengan waktu 3 menit 20 detik dan akurasi 57.40%, serta menggunakan fungsi aktivasi softmax dengan waktu 2 menit 55 detik dan akurasi 57.35% [18].

5. ***Analisis Sentimen Twitter Bahasa Indonesia Menggunakan Algoritma Convolutional Neural Network (Sartini : 2020) [20]***

Penelitian ini bertujuan untuk melakukan melakukan analisis sentimen dengan pendekatan Deep Learning lebih spesifiknya menggunakan metode Convolution Neural Network. Penggunaan dataset menggunakan 10806 tweet yang berasal dari *Indonesian-Sentimen-Analysis-Dataset*

dengan bahasa Indonesia. Untuk melakukan representasi kata dalam bentuk vector, digunakan model Word2vec dengan bahasa Indonesia.

Metode yang digunakan pada penelitian ini yaitu CNN atau Convolution Neural Network dengan bantuan model word2vec untuk representasi kata dalam bentuk vector. Pada penelitian perbandingan antara nilai train dan test yang digunakan adalah 80:20. Hasil yang didapatkan pada penelitian kali ini yaitu pada variasi CNN 12 dengan nilai akurasi 81.4% sementara nilai paling buruk didapatkan pada variasi ke 1 dengan akurasi 70.5% [20].

6. ***Analisis Sentimen Review Customer Terhadap Produk Indihome Dan First Media Menggunakan Algoritma Convolutional Neural Network (Saleh Hasan Badjrie, Oktariani Nurul Pratiwi, Hilman Dwi Anggana : 2021) [21]***

Penelitian ini bertujuan untuk melakukan analisis sentimen *review* customer di twitter terhadap produk indihome dan first media menggunakan algoritma Convolution Neural Network. Dataset yang digunakan pada penelitian berasal dari twitter dengan menggunakan web crawler berdasar dengan metadata, keywords, dan lain sebagainya. Jumlah data yang didapatkan menggunakan web crawler yaitu 13689 dengan total data provider *review* terhadap indihome sebanyak 7256 serta terhadap first media sebanyak 6433.

Metode yang digunakan pada penelitian ini yaitu dengan algoritma Convolutional Neural Network dikarenakan model CNN dianggap mendominasi model secara luas untuk melakukan klasifikasi teks. Hasil yang didapatkan setelah melakukan modeling dengan algoritma CNN, yaitu 98% tingkat akurasi untuk provider indihome dan 91% tingkat akurasi untuk provider first media. Dari hasil yang didapatkan dapat disimpulkan bahwa metode CNN merupakan algoritma yang sangat baik dalam melakukan proses klasifikasi data [21].

7. ***Analisis Sentimen Tweet Menggunakan Backpropagation Neural Network (Maulana Aziz Assuja, Saniati : 2016) [22]***

Penelitian ini bertujuan untuk melakukan analisis sentimen tweet pada twitter dengan menggunakan Backpropagation Neural Network. Klasifikasi teks membutuhkan beberapa fitur untuk menjadi ciri pengelompokan teks, pada penelitian kali ini yang akan digunakan adalah perangkingan term terbaik dengan menggunakan TF-IDF. Hal tersebut tentunya harus didukung dengan data yang baik, untuk mencapai hal tersebut perlunya dilakukan *preprocessing*. Tahap *preprocessing* pada penelitian ini yaitu menggunakan transformasi kata/term gaul, cleaning dan normalization, stopword removal, stemming, soundex, dan yang terakhir tokenisasi.

Dataset yang digunakan pada penelitian ini yaitu sebanyak 944 tweet yang terdiri dari 3 kelas yaitu netral (500 tweet), positif (254 tweet), dan negatif (190 tweet). Kemudian tweet akan dibagi menjadi data training sebesar 2/3 dan data test sebesar 1/3. Setelah data dibagi kemudian akan dipecah kembali menjadi 5 grup dengan kriteria tertentu untuk mendapatkan hasil dengan 3 jenis ukuran yaitu Correct Classified (C), Precision (P), dan Recall (R). Hasil akhir yang didapatkan dengan nilai terbaik jatuh pada group ke 4 dengan kriteria data melalui *preprocessing* transformasi, kata gaul, cleaning & normalization, stopword removal, stemming, dan tokenisasi yaitu 76.5% untuk nilai C, 80.93% untuk nilai P, dan 70.92% untuk nilai R [22].

8. ***Analisis Sentimen Hashtag "Dirumahaja" Saat Pandemi Covid-19 Di Indonesia Menggunakan NLP (Annisa Raudya Wibowo, Nuke Nidya, Aisyah Firdausi Rahma, Agussalim : 2020) [23]***

Penelitian ini bertujuan untuk melakukan analisis sentimen tweet pada twitter dengan keyword hashtag "dirumahaja" saat pandemic covid-19 di Indonesia dengan menggunakan NLP lebih tepatnya dengan menggunakan metode LDA atau Latent Dirichlet Allocation. Data yang digunakan pada penelitian kali ini berasal dari twitter menggunakan teknik Web Scrapping

dengan library “GetOldTweets3”. Data yang diambil berasal dari 10 daerah yang ada di Indonesia dari tanggal 03 Mei 2020 hingga 03 Juni 2020.

Hasil akhir yang didapatkan kemudian akan divisualisasikan agar hasil dapat diamati dengan baik. Pemodelan topik yang dilakukan dengan menggunakan implementasi dari Mallet mendapatkan hasil lima kata terpopuler yaitu Masker, Niat, Virtual, Repost, dan Promo untuk dijadikan sebagai motivasi masyarakat pada masa pandemi. Dari data yang didapatkan, sentimen yang didapatkan yaitu 88.11% menyatakan bahwa komentar masyarakat bersifat netral, 1.56% menyatakan bahwa komentar masyarakat bersifat negatif, dan 10.33% menyatakan bahwa komentar masyarakat bersifat positif [23].

9. ***Analisis Sentimen Tweet Vaksinasi Covid-19 Menggunakan Rnn Dengan Metode Tf-Idf Dan Word2vec (Nadia Ristya Dewi , Eva Yulia Puspaningrum , Hendra Maulana : 2022) [24]***

Penelitian ini bertujuan untuk melakukan analisis sentimen tweet vaksinasi covid-19 menggunakan algoritma RNN dengan metode word embedding TF-IDF dan Word2vec. Data yang digunakan pada penelitian berasal dari beberapa sumber yaitu lama keggle dengan judul “Indonesian Vaccination Tweet” dan crawling data pada twitter dengan keyword hashtag yang terkait yaitu #vaksin dan #vaksinasi. Hasil data yang didapatkan dari kedua sumber yaitu sebanyak 6490 data tweet.

Metode yang digunakan pada penelitian kali ini yaitu RNN dengan pendekatan metode word embedding TF-IDF dan Word2vec dengan perbandingan yang digunakan yaitu 7:3. Dari percobaan yang dilakukan, didapatkan bahwa penggunaan metode word embedding word2vec lebih baik dibandingkan TF-IDF dengan nilai akurasi 53%, presisi 56%, dan recall 78%. Tidak hanya dari nilai akurasi namun didapatkan juga word2vec lebih baik daripada TF-IDF pada time estimation ketika melakukan *modelling* [24].

10. ***Sentiment Analysis Twitter Bahasa Indonesia Berbasis Word2vec Menggunakan Deep Convolutional Neural Network (Hans Juwiantho, Esther Irawati Setiawan, Joan Santoso, Mauridhi Hery Purnomo : 2018) [25]***

Penelitian ini bertujuan untuk melakukan sentimen analisis pada twitter dengan bahas Indonesia berbasis word2vec menggunakan Deep learning lebih spesifiknya Convolutional Neural Network. Pemilihan metode dikarenakan dianggap memiliki kelebihan dalam pengenalan sentimen pada teks dengan melakukan perhitungan sequence kata pada kalimat sehingga mempedulikan urutan kata. Data yang digunakan berasal dari tweet masyarakat dengan bahasa Indonesia yang berjumlah 999 tweet.

Penggunaan metode CNN akan didukung dengan penggunaan Word2vec bahasa Indonesia sebagai inisialisasi kata menjadi vector sehingga tidak butuh pelatiihan word vector serta pencarian fitur secara manual. Setelah didapatkan akurasi nantinya akan dibandingkan hasilnya dengan beberapa metode lain sebagai pembanding seperti NB (Naïve Beyes) dan SVM (Support Vector Machine) serta dengan beberapa metode yang dapat mendiking model yaitu TF-IDF dan Word2vec. Hasil akhir dari percobaan yang telah dilakukan berupa tingkat akurasi memiliki nilai akurasi tertinggi dengan nilai 76.4% menggunakan metode CNN, 50 varian node buffer, serta 2 sentimen [25].

Tabel 2.1 Penelitian Sebelumnya

No.	Judul	Penulis	Metode	Hasil
1.	Aspect Based Sentimen Analysis Pada *Review* Produk Kecantikan Menggunakan *Extreme Gradient Boosting*	Jessie Gabriella S 2021	*Extreme Gradient Boosting*	Metode yang digunakan dalam penelitian ini adalah Extreme Gradient Boosting. Hasil pada penelitian ini menunjukan nilai yang cukup tinggi, dengan metode yang digunakan untuk melakukan Evaluasi model yaitu confusion matrix dengan output atau keluaran menjadi 4 hal yaitu precision, recall, f1-Score, dan accuracy. Hasil yang didapatkan pada nilai rata-rata accuracy dari empat aspek pada penelitian ini yaitu 90%.

No.	Judul	Penulis	Metode	Hasil
2.	Implementasi Algoritma Multinomial Naïve Bayes Untuk Analisis Sentimen Menggunakan Tf-Idf Dan N-Gram.	Prisilia Ines 2020	*Multinomial Naïve Bayes*	Hasil implementasi serta uji coba sistem, didapatkan performa terbaik nilai n yaitu dengan menggunakan unigram dengan ratio 80:20 pada training dan testing model yang dibuat. Kemudian model dilakukan evaluasi dengan menggunakan metode confusion matrix dengan nilai 84% untuk accuracy, precision, recall, dan f1-score.
3.	Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes	Clarisa Hasya Yustika, Adiwijaya, Said Al Faraby 2021	*Naïve Bayes Classifier*	Didapat performansi tertinggi oleh dataset diterjemahkan ke dalam Bahasa Inggris kemudian diterjemahkan ke Bahasa Indonesia dan tidak menggunakan stopword removal dengan parameter alpha atau smoothing sebesar 1, min_df sebesar 0,01, max_df sebesar 0,7, dan max_features sebesar 2000 menghasilkan performansi terbaik dengan nilai F1-Score sebesar 62,81%.
4.	Analisis Sentimen Pada Data Ulasan Twitter dengan Menggunakan Long Short Term Memory	Sharfina Febbi Handayani, Riszki Wijayatun Pratiwi, Mulyana Putriyani 2021	*Long Short Term Memory*	Dari Pengujian didapatkan hasil terbaik dengan parameter word2vec 57.15% menggunakan CBOW, jumlah neuron 150 dengan waktu 2 menit 50 detik dan akurasi 57.35%, jumlah epoch 30 dengan waktu 3 menit 20 detik dan akurasi 57.40%, serta menggunakan fungsi aktivasi softmax dengan waktu 2 menit 55 detik dan akurasi 57.35%.

No.	Judul	Penulis	Metode	Hasil
5.	Analisis Sentimen Twitter Bahasa Indonesia Menggunakan Algoritma Convolutional Neural Network	Sartini 2020	*Convolution Neural Network*	Pada penelitian perbandingan antara nilai train dan test yang digunakan adalah 80:20. Hasil yang didapatkan pada penelitian kali ini yaitu pada variasi CNN 12 dengan nilai akurasi 81.4% sementara nilai paling buruk didapatkan pada variasi ke 1 dengan akurasi 70.5%.
6.	Analisis Sentimen *Review* Customer Terhadap Produk Indihome Dan First Media Menggunakan Algoritma Convolutional Neural Network	Saleh Hasan Badjrie, Oktariani Nurul Pratiwi, Hilman Dwi Anggana 2021	*Convolutional neural network*	Hasil yang didapatkan setelah melakukan modeling dengan algoritma CNN, yaitu 98% tingkat akurasi untuk provider indihome dan 91% tingkat akurasi untuk provider first media. Dari hasil yang didapatkan dapat disimpulkan bahwa metode CNN merupakan algoritma yang sangat baik dalam melakukan proses klasifikasi data.
7.	Analisis Sentimen Tweet Menggunakan Backpropagation Neural Network	Maulana Aziz Assuja, Saniati 2016	*Backpropagation Neural Network*	Hasil akhir yang didapatkan dengan nilai terbaik jatuh pada group ke 4 dengan kriteria data melalui *preprocessing* transformasi, kata gaul, cleaning & normalization, stopword removal, stemming, dan tokenisasi yaitu 76.5% untuk nilai C, 80.93% untuk nilai P, dan 70.92% untuk nilai R.
8.	Analisis Sentimen Hashtag "Dirumahaja" Saat Pandemi Covid-19 Di Indonesia Menggunakan NLP	Annisa Raudya Wibowo, Nuke Nidya, Aisyah Firdausi Rahma, Agussalim 2020	*Latent Dirichlet Allocation*	Hasil akhir yang didapatkan kemudian akan divisualisasikan agar hasil dapat diamati dengan baik. Pemodelan topik yang dilakukan dengan menggunakan

No.	Judul	Penulis	Metode	Hasil
				implementasi dari Mallet mendapatkan hasil lima kata terpopuler yaitu Masker, Niat, Virtual, Repost, dan Promo untuk dijadikan sebagai motivasi masyarakat pada masa pandemi. Dari data yang didapatkan, sentimen yang didapatkan yaitu 88.11% menyatakan bahwa komentar masyarakat bersifat netral, 1.56% menyatakan bahwa komentar masyarakat bersifat negatif, dan 10.33% menyatakan bahwa komentar masyarakat bersifat positif.
9.	Analisis Sentimen Tweet Vaksinasi Covid-19 Menggunakan Rnn Dengan Metode Tf-Idf Dan Word2vec	Nadia Ristya Dewi, Eva Yulia Puspaningrum, Hendra Maulana 2022	*Recurrent Neural Network*	didapatkan bahwa penggunaan metode word embedding word2vec lebih baik dibandingkan TF-IDF dengan nilai akurasi 53%, presisi 56%, dan recall 78%. Tidak hanya dari nilai akurasi namun didapatkan juga word2vec lebih baik daripada TF-IDF pada time estimation ketika melakukan *modelling*.
10.	Sentiment Analysis Twitter Bahasa Indonesia Berbasis Word2vec Menggunakan Deep Convolutional Neural Network	Hans Juwiantho, Esther Irawati Setiawan, Joan Santoso, Mauridhi Hery Purnomo 2018	*Convolutional Neural Network*	Setelah didapatkan akurasi nantinya akan dibandingkan hasilnya dengan beberapa metode lain sebagai pembanding seperti NB (Naïve Beyes) dan SVM (Support Vector Machine) serta dengan beberapa metode yang dapat menduking model yaitu TF-IDF dan Word2vec. Hasil akhir

No.	Judul	Penulis	Metode	Hasil
				dari percobaan yang telah dilakukan berupa tingkat akurasi memiliki nilai akurasi tertinggi dengan nilai 76.4% menggunakan metode CNN, 50 varian node buffer, serta 2 sentimen.

Tabel 2.1 merupakan rangkuman dari data jurnal yang telah dijelaskan di atas sebagai referensi penulis. Informasi tersebut disajikan dalam bentuk tabel agar dapat mempermudah penulis dalam membaca dan menggali informasi lebih dalam mengenai penelitian ini.

2.2. Landasan Teori

Setelah membuat tinjauan pustaka dengan melihat penelitian sebelumnya, kemudian masuk ke bab landasan teori dimana pada bab ini akan dijelaskan beberapa teori yang berkaitan dengan penelitian yang dilakukan.

2.2.1. Text Mining

Text Mining sebetulnya merupakan pecahan dari data mining dimana proses utamanya adalah untuk melakukan suatu proses penambangan data berupa teks dengan mengekstrak secara langsung dan otomatis dari beberapa sumber dengan tujuan untuk merubah data yang tidak terstruktur menjadi data yang lebih terstruktur agar mudah untuk melakukan analisis[26]. Pengumpulan data biasanya dapat terlihat dari pola pola yang dibentuk oleh suatu teks dengan menggunakan mesin yang sudah dilatih sebelumnya[27].

Pada penerapannya, text mining dapat digunakan melakukan klasterisasi, klasifikasi, information retrieval, dan information extraction. Data yang telah diekstrak kemudian dapat digunakan untuk menganalisis opini, sentimen, penilaian, dan emosi seseorang dengan hal yang berkenaan dengan suatu topik yang sedang dibahas, individu, organisasi, maupun kegiatan yang sedang terjadi[28].

2.2.2. Analisis Sentimen

Analisis Sentimen merupakan suatu proses dalam menentukan maupun menemukan opini terhadap hal atau entitas yang sedang dibahas. Opini yang didapatkan, dapat digali lebih dalam untuk menentukan arti sebenarnya yaitu negatif, positif, atau netral, maupun perasaan yang diluapkan dalam sebuah data teks pada opini yang diberikan. Hal ini dapat digunakan pada berbagai bidang seperti kedokteran, politik, hingga bidang penerapan ilmu komputer yang dapat memudahkan pengambilan suatu keputusan dalam mengatur strategi bisnis kedepannya[29][30].

Analisis sentimen memiliki berbagai macam langkah yang dapat ditempuh, namun pada hakikatnya analisis sentimen memiliki langkah dasar yang selalu digunakan yaitu pengambilan data, *preprocessing* atau pembersihan data, melakukan *modelling* algoritma, serta evaluasi model untuk mendapatkan parameter penilai performa model seperti akurasi, presisi, f1-score, recall, dan error[31].

2.2.3. Web Scraping

Web scraping merupakan sebuah teknik yang biasa digunakan untuk melakukan pengambilan informasi dari website secara otomatis untuk mendapatkan data yang nantinya akan diolah kembali menjadi data yang siap dijadikan data olah dalam suatu analisis. Fokus yang dilakukan pada teknik ini yaitu dengan cara mengekstraksi suatu informasi sehingga memudahkan dalam melakukan suatu pencarian. Pengambilan data atau text mining sangat baik digunakan untuk mengambil data dalam jumlah yang banyak seperti pengambilan data sentimen masyarakat pada media sosial, pengambilan data pada suatu website e-commerce, dan masih banyak jenis data lainnya yang dapat diambil dengan teknik ini[32]. Langkah yang dilakukan untuk mengimplementasi teknik web scraping sebagai berikut.

1. Create Scraping Template = Pembuat sistem mempelajari dokumen HTML dari suatu website untuk diambil data atau informasi didalamnya sesuai dengan kebutuhan analisis.

2. Explore Site Navigation = Pembuat sistem mempelajari teknik navigasi website yang akan diambil data atau informasinya untuk dapat ditirukan ke sistem yang akan dibuat.
3. Automate Navigation and Extraction = Berdasarkan data atau informasi yang didapatkan pada step 1 dan 2, sistem akan dibuat untuk melakukan otomatisasi pengambilan data atau informasi dari website yang sudah ditentukan
4. Extracted Data and Package History = Setelah dilakukan langkah ke 3 kemudian data atau informasi disimpan pada tabel database[33].

2.2.4. Preprocessing

Preprocessing merupakan salah satu tahap dalam melakukan analisis sentimen untuk merubah data yang belum sempurna serta untuk mempersiapkan data agar lebih terstruktur. Tahap ini merupakan tahap yang paling penting pada proses analisis sentimen. Pada tahap ini dilakukan pembersihan data yang telah diambil atau diakuisisi dengan menggunakan crawling, scraping, survey maupun form. Data yang akan digunakan mengandung banyak noise seperti contohnya komentar pada sebuah website terkadang pengguna dapat menggunakan emoticon, hashtag, simbol, serta singkatan yang terkadang membuat sistem sulit memahami pola pada data yang akan digunakan [34]. Secara umum, *preprocessing* merupakan tahap untuk menghilangkan data yang tidak dibutuhkan dalam melakukan analisis dan merubah semua data menjadi data yang seragam [35]. Tahapan *preprocessing* secara teknis terdiri dari empat langkah yaitu :

1. Case Folding

Merupakan tahap untuk merubah karakter dari sebuah kalimat menjadi bentuk yang standar seperti contohnya akan diubah menjadi lowercase atau dapat juga diubah menjadi uppercase.

2. Normalization

Merupakan tahap untuk mengubah kata slang atau yang tidak baku menjadi kata baku yang ada pada KBBI serta merubah kata singkatan menjadi kata aslinya. Contohnya yaitu kalimat “gue kmrn pergi ke jakarta selatan” dilakukan normalisasi menjadi “saya kemarin ke jakarta selatan”.

3. Stopword Removal

Merupakan tahap untuk menghilangkan satu atau beberapa kata pada suatu kalimat yang tidak memiliki arti atau makna serta tidak memberikan informasi yang dibutuhkan dalam melakukan analisis sentimen. Kata yang termasuk dalam daftar kata stopword kemudian akan dihapus dengan tujuan agar kata tersebut tidak mempengaruhi tingkat error pada sistem. Stopword memiliki list kata yang digunakan, list tersebut disediakan oleh library sastrawi yang kemudian dapat digunakan oleh peneliti.

4. Tokenizing

Merupakan tahap untuk memotong kalimat menjadi potongan-potongan kata yang sesuai dengan kebutuhan analisis. Pemotongan kalimat dilakukan dengan melihat spasi yang ada pada tiap kalimat sehingga terpisah masing-masing katanya untuk dapat digunakan dalam melakukan analisis sentimen.

5. Stemming

Merupakan tahap dalam mengubah kata yang memiliki imbuhan menjadi kata dasar yang sesuai dengan kamus KBBI. Perubahan yang terjadi pada proses ini dengan menghapus awalan atau prefix, akhiran atau sufiks, serta kata imbuhan gabungan antara awalan dan akhiran yaitu konflik [36][37].

2.2.5. Pembobotan TF-IDF

Pembobotan TF-IDF merupakan sebuah teknik yang menggabungkan dua konsep dalam melakukan perhitungan bobot, konsep yang digabungkan dalam TF-IDF yaitu konsep TF *(Term Frequency)* dan IDF *(Invers Document Frequency)*. TF atau term frequency merupakan frekuensi dari banyaknya kemunculan kata dalam sebuah dokumen. Sementara IDF atau invers document frequency merupakan merupakan frekuensi kemunculan term pada keseluruhan dokumen. Term yang jarang muncul pada keseluruhan dokumen dapat dinyatakan memiliki nilai IDF lebih besar dibandingkan dengan term yang sering muncul pada dokumen[38][39]. Pembobotan TF-IDF bertujuan untuk mengerti seberapa penting sebuah kata dalam dokumen yang dianalisis. Adapun rumus dari pembobotan TF-IDF sebagai berikut.

$$W_{t,d} = tf_{t,d} \text{ x } idf_{t} \quad (2.1)$$

Dengan idf_{t} diperoleh dari

$$idf_{\mathrm{t}} = \log \frac{N}{df\mathrm{t}} \tag{2.2}$$

Maka dari itu rumus TF-IDF secara keseluruhan adalah sebagai berikut[40][41].

$$W_{t,d} = tf_{t,d} \text{ x } \log \frac{N}{df\mathrm{t}} \tag{2.3}$$

Keterangan :

$W_{t,d}$ = Bobot TF-IDF

$tf_{t,d}$ = Jumlah frekuensi kata

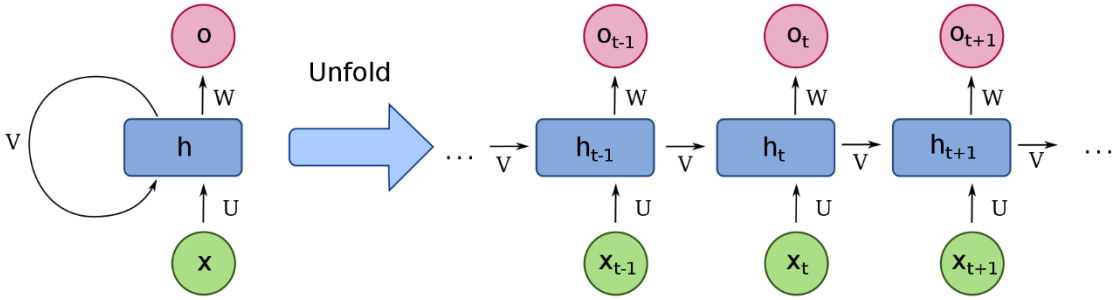
idf_{t} = Jumlah invers frekuensi dokumen tiap kata

N = Jumlah total dokumen

$df\mathrm{t}$ = jumlah dokumen yang mengandung kata yang sedang dihitung

2.2.6. Recurrent Neural Network (RNN)

Recurrent Neural network atau RNN merupakan salah satu algoritma deep learning dengan arsitektur yang prosesnya dipanggil secara berulang-ulang untuk dapat memproses sebuah data sekuensial atau data yang bersambung. Logika yang digunakan dalam algoritma RNN didapatkan dari pola pikir manusia yang tidak menentukan sebuah keputusan secara langsung namun mempertimbangkannya dengan informasi lampau atau lalu yang disimpan dalam menentukan keputusan. RNN memiliki sel memori yang mampu menyimpan informasi tentang urutan yang terbentuk dikarenakan RNN melalui proses pengulangan dalam arsitekturnya yang menyebabkan informasi yang lalu dapat tersimpan.



Gambar 2. 1 Arsitektur *Recurrent Neural Network*

Dari gambar 2.1 diatas, dapat dilihat arsitektur RNN dimana x_t merupakan input dalam waktu t atau masukan, h_t merupakan lapisan tersembunyi, dan o_t merupakan lapisan output. Proses diatas juga menunjukan adanya perulangan yang memungkinkan terjadinya informasi yang melewati satu jaringan ke jaringan yang lain. RNN bukanlah algoritma yang sempurna, RNN tetap memiliki permasalahan terutama pada short-term memory. Ketika sebuah data sekuensial yang lebih panjang, maka RNN akan mengalami kendala dalam membawa informasi dari step awal ke step berikutnya. Hal terjadi terjadi ketika muncul permasalahan vanishing gradient problem yang merupakan masalah ketika nilai gradient menyusut seiring proses backpropagation, ketika nilai menjadi sangat kecil maka proses learning akan berhenti dan menyebabkan RNN melupakan informasi pada sekuensial yang lebih panjang[42][43].

2.2.7. Long Short Term Memory (LSTM)

Jaringan Long Short Term Memory atau LSTM merupakan salah satu varian yang dimiliki oleh RNN yang dibuat untuk dapat mengatasi permasalahan utama dari algoritma RNN yaitu permasalahan vanishing gradient problem. Cara kerja LSTM yaitu dengan mengubah lapisan tersembunyi RNN menjadi blok memory yang memiliki gate system, hal tersebut membuat LSTM dapat menjaga memori janga panjang dengan prinsip melatih gate weight secara tepat dan telah terbukti dapat memecahkan masalah yang terjadi.

Model yang digunakan tersusun dari beberapa lapisan diantaranya yaitu pada lapisan pertama merupakan lapisan embedding. Pada lapisan kedua lapisan

LSTM dengan neuron yang dibutuhkan. Kemudian lapisan ketiga yaitu dense sebagai lapisan penghubung sebagai lapisan yang dapat memetakan lapisan output LSTM ke ukuran output yang diinginkan, dan diberikan aktivasi sigmoid untuk merubah semua nilai output yang dihasilkan menjadi nilai antara 0 dan 1[44].

2.2.8. Word Embedding

Word Embedding merupakan sebuah teknik dalam melakukan representasi makna kata yang dimungkinkan memiliki makna yang sama ataupun representasi yang serupa. Pada proses word embedding, data yang digunakan umumnya memiliki karakteristik seperti berdimensi tinggi, terdapat noise, serta terdapat struktur teks yang kurang baik. Data yang berasal dari komentar pada era ini banyak menggunakan bahasa yang tidak baku/slang serta terdapat kesalahan dalam ejaan katanya [45]. Word embedding yang digunakan pada penelitian kali ini yaitu word embedding yang bekerja untuk dapat merubah setiap data alphanumerical menjadi vector dengan panjang tertentu. Vektor yang dihasilkan merupakan vektor yang padat dengan nilai yang riil tidak hanya 1 dan 0.

2.2.9. Convolutional neural network (CNN)

Convolutional neural network atau CNN merupakan sebuah algoritma dari deep learning yang dominan untuk digunakan dalam CV *(Computer Vision)* seperti mengklasifikasikan gambar, identifikasi wajah, penggolongan penyakit melalui gambar, serta berbagai aspek dengan data visual. Namun fakta di lapangan bahwa metode CNN tidak hanya dapat digunakan pada CV saja namun dapat digunakan juga untuk NLP (Natural Language Processing). Penggunaan CNN dianggap efektif dalam melakukan klasifikasi dikarenakan memiliki lapisan yang dapat mempelajari fitur pada data yang dimiliki. Algoritma CNN tersusun atas tiga lapisan yang saling terhubung yaitu lapisan convolution, lapisan pooling, dan lapisan yang menghubungkan sepenuhnya (*Fully Connect*). Kedua lapisan pertama yaitu convolution dan pooling merupakan lapisan untuk melakukan ekstrasi fitur. Sedangkan lapisan yang ketiga yaitu lapisan *fully connected* merupakan lapisan dengan tugas untuk merubah fitur yang sudah terekstraksi menjadi output yang diinginkan seperti klasifikasi.

Konvolusi merupakan proses yang merupakan sebuah tipe operasi linier khusus yang digunakan untuk mengekstraksi fitur, dimana kernel diimplementasikan pada inputan yang berupa array dan disebut tensor. Matriks kernel dan masukan tensor dijumlahkan untuk mendapatkan output yang disebut feature map. Proses ini dilakukan secara berulang hingga menghasilkan feature map yang merepresentasikan karakteristik dari masukan tensor. Kernel yang digunakan umumnya menggunakan ukuran ganjil seperti 3x3, 5x5, atau 7x7. Penentuan ukuran kernel dapat mempengaruhi hasil ekstraksi yang dilakukan. Proses konvolusi membutuhkan beberapa hyperparameter yaitu stride dan padding untuk meningkatkan akurasi. Stride merupakan hyperparameter untuk menentukan jumlah pergeseran varian node buffer. Sedangkan padding merupakan jumlah pixel berisikan 0 pada input dengan tujuan memanipulasi dimensi output pada feature map [46][47].

Pooling dapat disebut dengan downsampling atau penurunan sampling untuk mengurangi dimensionalitas bidang dengan tetap menjaga informasi yang dianggap penting. Pooling memiliki beberapa jenis salah satu yang sering digunakan adalah max pooling dimana terjadi proses ekstraksi dari varian node buffer masukan menjadi feature maps dengan memilih nilai tertinggi dari masing-masing varian node buffer kemudian menghapus data varian node buffer yang tidak dibutuhkan. Varian node buffer dengan ukuran 2x2 sering digunakan dalam proses max pooling.

Lapisan *Fully Connected* atau FC merupakan lapisan neuron yang memiliki hubungan dengan seluruh aktivasi pada lapisan sebelumnya. Lapisan FC umumnya digunakan sebelum hasil klasifikasi didapatkan. Hasil output dari lapisan pooling biasanya flattened yang kemudian dihubungkan dengan lapisan FC untuk terhubung ke setiap output yang memiliki bobot. Hasil ekstraksi fitur dari lapisan konvolusi dan penurunan sampel pada lapisan pooling, kemudian hasilnya akan dipetakan oleh lapisan FC menjadi output akhir jaringan yang memiliki jumlah node sama dengan kelasnya [48][49].

2.2.10. Confusion Matrix

Merupakan suatu metode dalam proses analisis sentimen dengan tujuan untuk melakukan evaluasi terhadap klasifikasi suatu algoritma. Proses yang dilakukan pada metode ini adalah dengan mengukur nilai performansi algoritma yang digunakan[50]. Confusion matrix sering digunakan dalam proses klasifikasi yang melibatkan lebih dari dua kelas atau multiple classifiers. Hasil output yang diharapkan dari proses ini yaitu akurasi dengan menggunakan rumus persamaan[51].

$$Akurasi = \frac{TP+TN}{TP+FP+TN+\text{FN}} \quad (2.4)$$

Keterangan :

1. True Positif (TP), ketika kelas yang diprediksi positif dan faktanya positif.
2. False Positif (FP), ketika kelas yang diprediksi positif dan faktanya negatif.
3. True Negatif (TN), ketika kelas yang diprediksi negatif, dan faktanya negatif.
4. False Negatif (FN), ketika kelas yang diprediksi negatif dan faktanya positif [52].

Perhitungan akurasi pada metode ini ditampilkan dengan menggunakan tabel sebagai berikut.

Tabel 2.2 Perhitungan Performance Confusin Matrix

		True Class	
		Positif	Negatif
Predicted Class	Positif	True Positif (TP)	False Positif (FP)
	Negatif	False Negatif (FN)	True Negatif (TN)