BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian ini dibuat dengan mengambil informasi dari peneltian - penelitian sebelumnya yang berkaitan dan relevan terhadap penelitian yang dilakukan, buku dan penelitian - penelitian tersebut digunakan sebagai bahan referensi, serta sebagai bahan perbandingan agar penelitian yang dilakukan menjadi lebih baik lagi. Adapun penelitian terkait dapat dilihat pada Tabel 2.1.

Agus Sasmito Aribowo dan Siti Khomsah pada tahun 2021 melakukan implementasi *text mining* menggunakan metode *Lexicon* untuk mendeteksi emosi pengguna Twitter mengenai *Covid-19*. *Dataset* yang terkumpul untuk penelitian ini sebanyak 42.675 opini yang dikumpulkan menggunakan Twitter API. Hasil yang didapatkan adalah percakapan yang paling dominan pada topik tersebut adalah emosi takut yaitu sebesar 92% hingga 94% [10].

Mei Silviana Saputri, Rahmad Mahendra, dan Mirna Adriani pada tahun 2018 melakukan klasifikasi emosi menggunakan dataset dari Twitter di Indonesia dan menentukan fitur terbaik dalam klasifikasi emosi. Fitur yang digunakan dalam penelitian ini adalah *Lexicon Based*, *Bag-of-Words*, *word embedding* atau penyisipan kata, *orthography*, dan *Part-of-Speech* (*POS*). Berdasarkan hasil penelitian yang dilakukan, dapat disimpulkan bahwa fitur yang paling signifikan terbentuk berdasarkan kombinasi dari kamus emosi (*Emotion Lexicon*), *Bag-of-Words*, dan *FastText*. Penelitian tersebut juga membuat fitur kombinasi dasar dari *emotion lexicon*, *orthography*, dan *POS* yang dapat meningkatkan kinerja kumpulan data yang dibuat [11]

Penelitian yang dilakukan oleh Bilal Ghanem, Paolo Rosso, dan Francisco Range pada tahun 2020 melakukan analisis emosi terhadap informasi palsu seperti propaganda, hoax, clickbait dan satir dari situs berita online serta media sosial. Penelitian menggunakan EmoSenticNet, EmoLex, SentiSense, LIWC, dan Empath untuk meningkatkan cakupan kata – kata emosional dalam teks serta agar rentang emosi saat analisis lebih luas [12].

Rachmad Mahendrajaya, Ghulam Asrofi Buntoro, Moh. Bhanu Setyawan (2019) melakukan penelitian menggunakan opini dari pengguna Gopay di media sosial Twitter, data yang berhasil dikumpulkan terdiri dari 1210 opini menggunakan API Key Twitter dan dilanjutkan ke tahap *Prepocessing*. Menggunakan algoritma *Lexicon Based* didapatkan frekuensi sentimen positif sebanyak 923 opini, dan sentimen negatif sebanyak 287 opini. Hasil perbandingan kernel SVM menunjukan bahwa kernel linear memiliki akurasi 89,17%, presisi 91,5%, dan *recall* 97,83% sedangkan kernel polynomial menunjukan akurasi sebesar 84,38%, presisi sebesar 84,23%, dan *recall* sebesar 97,83% [13]

Siti Khomsah dan Agus Sasmito Aribowo dalam penelitiannya pada tahun 2020 membahas mengenai model *text preprocessing* pada komentar YouTube berbahasa Indonesia. Penelitian ini menggunakan fitur *N-Gram Word* untuk analisis dokumen teks, *Count-Vectorizer* untuk mengubah *Bag of Words* menjadi vector, dan TF-IDF untuk menghitung bobot kemunculan kata pada dokumen. Hasil dari penelitian tersebut menjelaskan bahwa model *preprocessing* yang baik pada data berbahasa Indonesia adalah menghapus *stopword*, konversi *slang word*, dan menghilangkan kata yang jenisnya objek atau subjek [14]

Amrita Mathur, Purnima Kubde, Sonali Vaidya melakukan analisis emosi menggunakan data dari Twitter selama situasi pandemi COVID-19. Data yang telah bersih kemudian diproses menggunakan NRC EmoLex. Analisis emosi yang dilakukan dalam penelitian ini digunakan untuk memahami kesehatan mental masyarakat dan dapat digunakan dalam pengambilan tindakan untuk memutuskan kebijakan saat memerangi virus COVID-19 yang dapat memengaruhi kesejahteraan masyarakat serta ekonomi seluruh dunia [15].

Siti Khomsah, Rima Dias Ramadhani, Sena Wijayanto pada tahun 2021, melakukan penelitian untuk mengetahui pengalaman wisatawan pada objek wisata yang ada di Purwokerto dengan menerapkan analisis *big data* berdasarkan opini wisatawan pada *platform* Google Map *review*. Hasil dari penelitian tersebut menunjukan bahwa sentimen wisatawan paling rendah yaitu pada objek wisata Telaga Sunyi sebesar 56%, dan yang paling tinggi didapatkan oleh objek wisata

The Village sebesar 88%, secara umum didapatkan bahwa presentase sentimen negatif wisatawan cukup rendah karena tidak mencapai 30% [9].

Laura Serrano , Antonio Ariza-Montes , Martín Nader , Antonio Sianes & Rob Law pada tahun 2020 melakukan eksplorasi preferensi dan sikap berkelanjutan melalui ulasan pengguna Airbnb menggunakan pendekatan *text mining*. Analisis data dilakukan menggunakan 10.488 komentar pengguna, yang dimana komentar tersebut telah melalui tahap pembersihan dan pemilahan data sesuai dengan kata kunci yang ditentukan. Hasilnya, platform akomodasi digital dapat merancang dan menerapkan sistem rekomendasi otomatis untuk calon pengguna Airbnb berdasarkan preferensi yang diidentifikasi dalam penelitian ini, yaitu dengan menggabungkan kategori pencarian yang ditetapkan (misalnya, harga, lokasi, pengalaman, dan fasilitas) dengan informasi yang terkandung dalam komentar dan peringkat yang diberikan oleh pengguna, sehingga, keputusan yang lebih cepat dan tepat serta andal dapat dibuat oleh pengguna di masa mendatang [16].

Huy Quan Vu, Gang Li, Rob Law, dan Yanchun Zhang melakukan penelitian mengenai preferensi tempat makan wisatawan berdasarkan ulasan pada *restaurant* di tahun 2019 menganalisis komentar pengunjung *restaurant* di Australia melalui situs TripAdvisor menunjukan hasil bahwa informasi berdasarkan ulasan wisatawan dapat dieksplorasi menggunakan analisis sentimen, dan sangat efektif dalan menilai kepuasan wisatawan serta mengidentifikasi kekurangan untuk dilakukan perbaikan di masa mendatang. Metode *text processing* dan analisis sentimen dapat diterapkan tidak hanya untuk komentar pada situs web perjalanan, tetapi juga teks singkat yang membahas tentang *restaurant* di berbagai platform media sosial sepeerti Facebook ataupun Twitter [17].

Penelitian yang dilakukan oleh Samer Muthana Sarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, Waleed Alnumay, Andrew Paul Smith pada 2021 mengambil *dataset* dari Twitter mengenai pesan bunuh diri, dengan menggunakan *NRC Affect Intensity Lexicon* dan *SentiStrength technique*, penelitian ini menemukan bahwa tweet yang terkait dengan konten bunuh diri meiliki keterkaitan dengan emosi takut (*fear*), sedih (*sad*), dan sentimen negatif [18].

Vimala Balakrishnan dan Wandeep Kaur pada tahun 2019 melakukan penelitian dengan *dataset* yang diambil dari grup resmi pendukung diabetes di Facebook, penelitian ini dilakukan untuk melakukan perbandingan antara *NRC Emotion Lexicon* dengan *Multinomial Naïve Bayes algorithm* berbasis *string vector*. Hasil yang didapatkan adalah metode *MNB string vector* lebih unggul. Selain itu, sebagian besar unggahan pada Facebook yang ditemukan bersifat gembira, diikuti oleh unggahan dengan kategori takut dan sedih [19].

Berdasarkan penelitian – penelitian terdahulu, dapat diketahui bahwa terdapat beberapa perbedaan dengan penelitian yang akan dilakukan. *Dataset* yang digunakan untuk penelitian ini menggunakan data komentar dari Google Map, terdapat beberapa penelitian yang menggunakan *dataset* dari sumber yang sama namun dengan subjek yang berbeda. *Emotion lexicon* yang digunakan pada penelitian ini adalah *NRC Emolex* Saif Mohammad yang terdiri dari delapan kategori emosi. Beberapa penelitian menggunakan *emotion lexicon* yang lain seperti *EmoSenticNet* [12] dan *NRC Affect Intensity Lexicon* [18].

Tabel 2. 1 Penelitian Terdahulu

No	Judul	Penulis	Dataset	Metode Penelitian	Hasil Penelitian	Perbedaan
1	*Implementation Of Text Mining For Emotion Detection Using The Lexicon Method (Case Study: Tweets About Covid-19)*	Agus Sasmito Aribowo, Siti Khomsah (2021)	*Dataset* bersumber dari Twitter.	*Emotion Lexicon*	Hasil yang didapatkan adalah percakapan yang paling dominan pada topik tersebut adalah emosi takut yaitu sebesar 92% hingga 94%.	Sumber data yang digunakan pada penelitian terdahulu berasal dari Twitter.
2	*Emotion Classification on Indonesian Twitter Dataset*	Mei Silviana Saputri, Rahmad Mahendra, dan Mirna Adriani (2018)	Data yang digunakan sebagai analisis berasal dari komentar netizen pada media sosial Twitter.	*Lexicon Based*, *Bag-of-Words*, *word embedding* atau penyisipan kata, *orthography*, dan *Part-of-Speech* (POS).	Hasil penelitian disimpulkan bahwa fitur yang paling signifikan adalah kombinasi dari kamus emosi (*Emotion Lexicon*), *Bag-of-Words*, dan *FastText*	Sumber data pada penelitian terdahulu berasal dari Twitter. Metode yang digunakan pada penelitian terdahulu merupakan campuran dari beberapa metode.
3	*An Emotional Analysis of False Information in Social Media*	Bilal Ghanem, Paolo Rosso, dan Francisco Range (2020)	Sumber data yang digunakan berasal dari artikel berita dan Twitter.	*EmoSenticNet, EmoLex, SentiSense, LIWC,* dan *Empath*	Analisis menunjukkan bahwa emosi yang digunakan pada sebuah kalimat memiliki perbedaan yang	Sumber data berasal dari Twitter dan portal berita.

No	Judul	Penulis	Dataset	Metode Penelitian	Hasil Penelitian	Perbedaan
	and News Articles				signifikan antara berita nyata dan berita bohong.	
4	Analisis Sentimen Pengguna Gopay Menggunakan Metode *Lexicon Based* dan *Support Vector Machine*	Rachmad Mahendrajaya , Ghulam Asrofi Buntoro, Moh. Bhanu Setyawan (2019)	Penelitian ini menggunakan data komentar netizen pada media sosial Twitter mengenai aplikasi Go-Pay.	*Lexicon Based* untuk pelabelan data, klasifikasi menggunakan metode *Support Vector Machine* atau SVM, dan TF-IDF untuk proses pembobotan kata. *Text preprocessing :* normalisasi kalimat, *cleansing*, menghilangkan angka, *emoticon* dan tanda baca, *case folding*, *filtering* dan tokenizing.	Hasil pelabelan menggunakan *Lexicon Based* terdiri dari 923 komentar positif dan 287 komentar negatif, serta menghasilkan akurasi sebesar 89,17% (kernel Linear) dan 84,3% (kernel Polynomial) dengan klasifikasi SVM.	Sumber data yang digunakan berasal dari media sosial Twitter. Metode klasifikasi pada penelitian ini menggunakan SVM.
5	Model Text Preprocessing Komentar Youtube Dalam Bahasa Indonesia	Siti Khomsah dan Agus Sasmito Aribowo (2020)	*Dataset* yang digunakan adalah komentar berbahasa Indonesia pada Youtube	*N-Gram Word, Count-Vectorizer,* TF-IDF	Model *preprocessing* yang baik pada data berbahasa Indonesia adalah menghapus *stopword*, konversi *slang word*, dan menghilangkan kata objek atau subjek	Sumber data, metode penelitian

No	Judul	Penulis	Dataset	Metode Penelitian	Hasil Penelitian	Perbedaan
6	*Emotional Analysis using Twitter Data during Pandemic Situation: COVID-19*	Amrita Mathur, Purnima Kubde, Sonali Vaidya (2020)	*Dataset* menggunakan data komentar netizen pada media sosial Twitter selama situasi pandemi COVID-19.	Pemrosesan data menggunakan *NRC Emotion Lexicon*	Pada penelitian ini, data Twitter dianalisis untuk memahami kesehatan mental masyarakat selama situasi pandemi COVID 19 melalui analisis emosi dan mengklasifikasikannya ke dalam emosi dasar.	Sumber data yang digunakan berasal dari media sosial Twitter.
7	Big Data Analytic Untuk Mengetahui Pengalaman Wisatawan (Studi Kasus: Objek Wisata di Purwokerto)	Siti Khomsah, Rima Dias Ramadhani, Sena Wijayanto (2021)	Dataset berupa komentar wisatawan pada *platform* Google Map review sebanyak 8 objek wisata.	*Big Data Analytic* *Preprocessing Data* : penghapusan karakter, konversi *lowercase*, stopword *removing*, konversi kata slang, *stemming*	Sentimen yang dihasilkan secara umum merupakan sentimen positif, pada setiap objek wisata didapatkan sentimen positif pada kisaran 55% hingga sentimen tertinggi 92% yang didapatkan oleh objek wisata The Village.	Jumlah objek wisata yang ada pada penelitian sebanyak 8 objek wisata.
8	*Exploring preferences and sustainable attitudes of Airbnb green users in the review*	Laura Serrano, Antonio Ariza - Montes , Martín Nader, Antonio	Data didapatkan dari situs Inside Airbnb dan menghasilkan sebanyak 13.181.297 komentar	*Text mining,* sentimen analisis *Text preprocessing :* pengumpulan data, *data cleaning*, stopword *removing*	Hasilnya, platform akomodasi digital dapat merancang dan menerapkan sistem rekomendasi otomatis untuk calon pengguna Airbnb berdasarkan	Dataset penelitian berasal dari situs Inside Airbnb

No	Judul	Penulis	Dataset	Metode Penelitian	Hasil Penelitian	Perbedaan
	comments and ratings: a text mining approach	Sianes & Rob Law (2020)			preferensi pengunjung sebelumnya.	
9	*Exploring Tourist Dining Preferences Based on Restaurant Reviews*	Huy Quan Vu, Gang Li, Rob Law, dan Yanchun Zhang (2019)	Dataset berasal dari komentar pengunjung *restaurant* di platform TripAdvisor	*Text mining,* sentimen analisis *Text preprocessing* : *Crawling data*, tokenization, *filtering*, *stemming*	Analisis sentimen efektif digunakan untuk menilai kepuasan pengguna, khususnya dalam dunia kuliner sehingga preferensi setiap pengunjung dapat didefinisikan.	Topik yang dibahas dikhususkan pada industri kuliner (*restaurant*), dan data diambil dari situs TripAdvisor
10	*A lexicon-based approach to detecting suicide-related messages on Twitter*	Samer Muthana Sarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, Waleed Alnumay, Andrew Paul Smith (2021)	Dataset yang digunakan adalah *tweet* keinginan bunuh diri pada aplikasi Twitter	*NRC Affect Intensity Lexicon* dan *SentiStrength techniques*	Hasil penelitian menunjukkan bahwa tweet yang terkait dengan konten bunuh diri meiliki keterkaitan dengan emosi takut (*fear*), sedih (*sad*), dan sentimen negatif.	Topik yang dibahas dikhususkan pada *tweet* bunuh diri dan data diambil dari aplikasi Twitter
11	*String-based Multinomial Naïve Bayes for*	Vimala Balakrishnan,	Data berasal dari postingan resmi grup pendukung	Klasifikasi emosi menggunakan *NRC Emotion Lexicon* dan	Sebagian besar unggahan ditemukan memiliki emosi bahagia, diikuti	Data berasal dari Facebook khusus dari

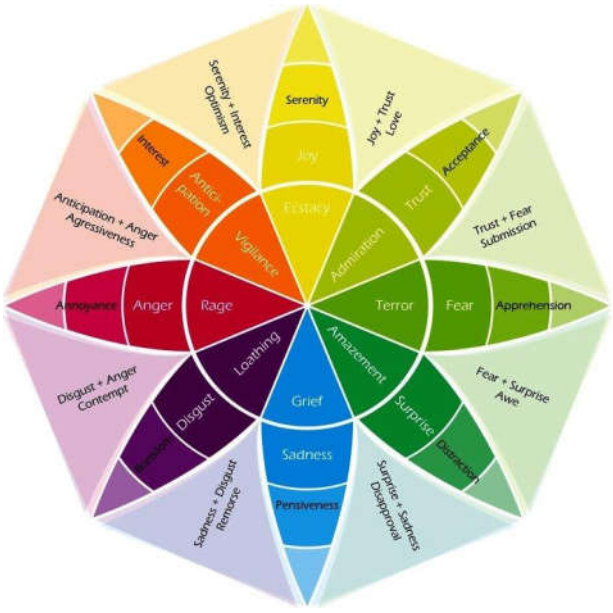
No	Judul	Penulis	Dataset	Metode Penelitian	Hasil Penelitian	Perbedaan
	Emotion Detection among Facebook Diabete Community	Wandeep Kaur (2019)	diabetes di Facebook	*Multinomial Naïve Bayes algorithm*	oleh emosi takut dan sedih. Temuan penelitian menunjukkan bahwa deteksi emosi menggunakan vector berbasis string lebih baik daripada hanya bergantung pada vector numerik	grup pendukung diabetes resmi

2.2 Dasar Teori

2.2.1. *NRC Emotion Lexicon*

Pendekatan berbasis leksikon (*Lexicon Based Approach*) adalah kamus berisi kata – kata yang memiliki sentimen positif dan negative yang digunakan untuk menentukan apakah *review* pengguna masuk dalam kategori kata negatif atau kata positif [20]. *NRC Emotion Lexicon* atau *NRC EmoLex* mengklasifikasikan kata ke dalam bentuk biner (ya/tidak) untuk kelas sentiment (positif dan negatif) dan untuk kelas emosi (gembira, yakin, takut, terkejut, sedih, muak, marah, serta antisipatif) [21]. Ketika menganalisis teks, secara otomatis emosi – emosi tersebut dideteksi, hal ini dapat berguna untuk sejumlah tujuan seperti mengidentifikasi komentar yang mengekspresikan emosi tertentu. *Lexicon* dapat menjadi metode yang berguna untuk mengidentifikasi emosi yang ditimbulkan pada sebuah kata secara otomatis [22]. *NRC Emotion Lexicon* dibuat dengan mengidentifikasi daftar kata sesuai dengan anotasi manusia, Saif Mohammad dan Peter Turney pada 2010 menggunakan *Macquarie Thesaurus* sebagai sumber anotasi, dan kemudian dilakukan validasi terhadap anotasi tersebut secara otomatis [23].

2.2.2. *Emotion Analysis*



Gambar 2. 1 *Plutchik's Wheel of Emotions* [24]

Ekspresi emosi merupakan bentuk komunikasi yang penting dalam hubungan interpersonal [25], yang dapat diekspresikan menjadi emosi positif, negatif, ataupun netral [26]. Terdapat delapan emosi primer yaitu *joy* (gembira), *trust* (yakin), *fear* (takut), *surprise* (terkejut), *sadness* (sedih), *disgust* (muak), *anger* (marah), dan *anticipation* (antisipatif) [24]. *Joy* dapat didefinisikan sebagai ketenangan, bahagia, dan candu. *Trust* didefinisikan sebagai penerimaan dan kekaguman. *Fear* adalah bentuk rasa takut dan terror. *Surprise* dapat didefinisikan sebagai bentuk ketidakpastian serta takjub. *Sadness* adalah sebuah kesuraman atau duka. *Disgust* adalah sebuah ketidaksukaan dan benci. *Anger* didefinisikan sebagai rasa jengkel dan geram. *Anticipation* adalah minat ataupun waspada [27]. Emosi - emosi tersebut diekspresikan dalam berbagai *platform* media sosial seperti YouTube, Twitter, Facebook, dan lainnya, yang kemudian dapat menjadi sumber informasi tekstual agar dapat bermanfaat untuk berbagai macam keperluan.

2.2.3. *Pariwisata*

Industri pariwisata merupakan suatu jenis usaha yang menyediakan jasa wisata seperti penginapan, transportasi, agen perjalanan, *restaurant*, klub malam, fasilitas hiburan dan olahraga, toko suvenir, dan sebagainya [28]. Kegiatan wisata dilakukan untuk mencari kesenangan atau kebahagiaan sesuai dengan keinginan wisatawan seperti wisata budaya, wisata alam, atau wisata pendidikan. Untuk dapat dikatakan sebagai kegiatan wisata, syarat yang harus dipenuhi adalah sifatnya sementara, sukarela atau tanpa paksaan, dan tidak melakukan sebuah pekerjaan yang menghasilkan upah [29].

2.2.4. *Google Maps Review*

Google Maps *Review* merupakan salah satu fitur yang dimiliki oleh aplikasi Google Maps. Google Maps sendiri adalah sebuah aplikasi web yang digunakan untuk menghitung rute antar lokasi dan menawarkan peta serta foto yang dapat digulir oleh penggunanya [30]. Salah satu pencarian yang dapat dilakukan di aplikasi Google Maps adalah pencarian destinasi wisata, disini pengguna dapat

melihat ulasan, memberikan ulasan, melihat foto terkait, dan melihat rute menuju destinasi tersebut.

2.2.5. *Scraping*

Web Scraping atau yang dikenal dengan ekstraksi atau *web harvesting* adalah teknik yang digunakan untuk mengekstrak data dari WWW atau *World Wide Web* dan menyimpannya ke dalam sistem file atau database, *web scraping* dilakukan untuk pengambilan data yang digunakan sebagai bahan analisis nantinya. *Scraping* dapat dilakukan baik secara manual maupun secara otomatis menggunakan *web crawler* [31]. *Web crawler* merupakan program komputer berupa mesin pencari teks yang digunakan untuk melakukan penelusuran pada WWW (*World Wide Web*) dengan cara otomatis dan teratur. Pengguna dapat menemukan data yang mereka cari dengan menggunakan tautan *hypertext* yang berbeda dan mengekstraknya menjadi data yang bermanfaat [32].

2.2.6. WebHarvy

WebHarvy adalah salah satu alat *scrape* yang cukup ringan, dan dapat dikuasai secara singkat dalam kebutuhan mengekstrak data. WebHarvy dapat digunakan untuk melakukan *scrape* data tanpa memerlukan kode apapun. Data yang telah diekstrak kemudian dapat disimpan dalam format umum seperti CSV, TSV, dan XML [33]. WebHarvy merupakan *software* komersial yang dapat mengambil ulasan yang dihasilkan oleh pengguna dari situs web. Kelebihan yang dimiliki WebHarvy adalah aplikasi ini dapat secara otomatis melakukan *scrape* sebuah teks dari situs web, dan menyimpan data tersebut kedalam berbagai format [34].

2.2.7. *Preprocessing*

Preprocessing merupakan tahap penting dalam penambangan teks. Tahap ini dilakukan dengan mereduksi beberapa bentuk kata menjadi satu bentuk khusus.

Preprocessing dapat mengurangi waktu dan mempercepat proses yang dibutuhkan pada klasifikasi [35]. Tahapan yang dilakukan pada *preprocessing* adalah sebagai berikut [8] :

1. *Case Folding* merupakan tahapan untuk mengubah seluruh kalimat menjadi *lowercase* atau huruf kecil. Misalnya "Pergi" menjadi "pergi"
2. *Stopword Removing* dilakukan dengan menghapus kata kata dalam kalimat yang tidak berperan penting dalam memberi informasi, misalnya "apa", "bagaimana", dan lain sebagainya, serta menghilangkan karakter dalam kalimat yang dianggap tidak valid seperti tanda baca
3. *Tokenizing* merupakan tahapan yang dilakukan dengan memotong kalimat menjadi kata per kata yang disebut token, misalnya kalimat "saya suka hujan" menjadi "saya", "suka", dan "hujan"
4. *Stemming* adalah merubah kata yang memiliki imbuhan menjadi kata dasarnya. Contoh kata "berlari" menjadi "lari", "menyenangkan" menjadi "senang".

2.2.8. RapidMiner

RapidMiner adalah aplikasi *open source* yang digunakan untuk melakukan analisis mengenai *data mining*, *text mining*, dan analisis prediksi. Pemrosesan data yang dapat dilakukan pada RapidMiner antara lain fungsi *input, output*, *preprocessing data*, hingga visuasisasi data [36]. RapidMiner digunakan dengan melakukan *import* data kedalam aplikasi, kemudian apabila *dataset* telah siap, dapat memulai penggunaan dengan menyambungkan beberapa operator yang dibutuhkan [37].