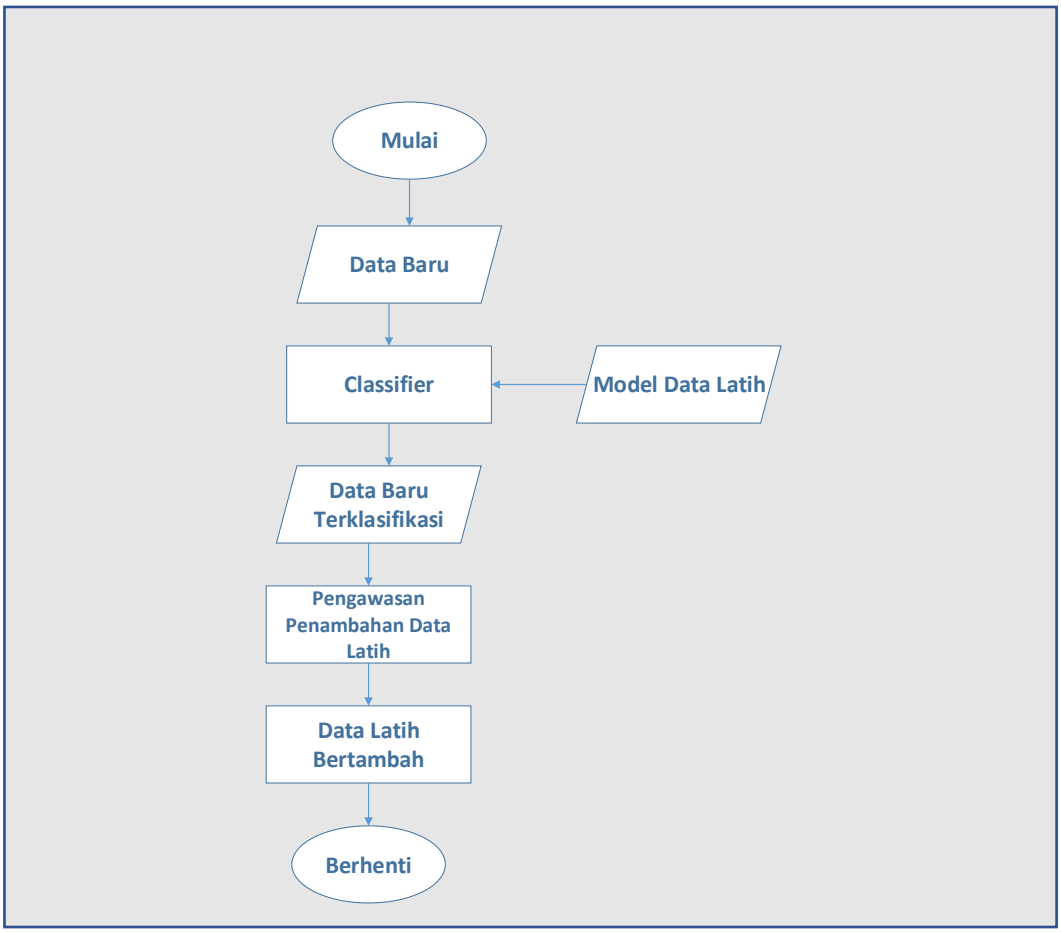
BAB 3
METODE PENELITIAN

3.1 SVM – TFIDFOR

Aplikasi analisis sentimen yang sudah dipakai untuk kebutuhan produksi suatu perusahaan pada tahap inisialisasinya dapat dipastikan sudah memiliki data latih agar dapat mengenali pola dokumen yang memiliki sentimen positif atau negatif, sehingga aplikasi tersebut dapat melakukan pengklasifikasian terhadap data baru. Bagi yang menggunakan sumber data yang dinamis seperti sosial media, akan menjadi masalah tersendiri apabila data latih yang digunakan tetap sama dari awal inisialisasi. Ini dikarenakan bahasa yang digunakan di sosial media terus mengalami perubahan sehingga diperlukannya suatu data baru agar aplikasi tetap dapat melakukan klasifikasi dengan akurasi yang tetap tinggi. Banyak skema yang bisa digunakan untuk menambah data latih, salah satunya adalah mengambil dari data baru yang telah diklasifikasi. Akan lebih valid apabila data baru tersebut sebelum ditambahkan sebagai data latih baru, diawasi terlebih dahulu terkait kesesuaian klasifikasinya. Berikut adalah skema penambahan data latih dari data baru yang diawasi.



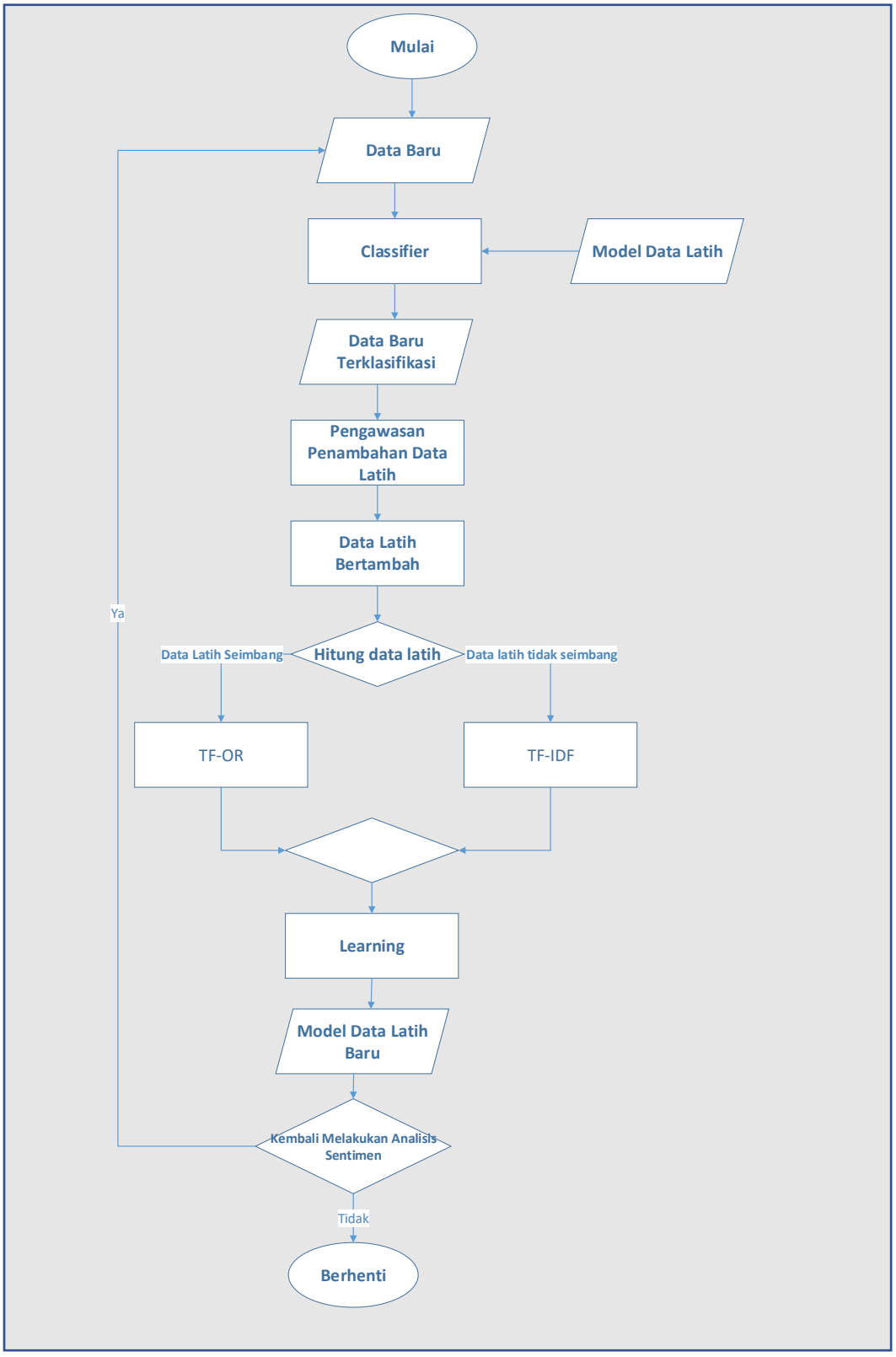
Gambar 3-1 Skema Penambahan Data Latih

Dengan skema tersebut data latih baru akan memberi pengetahuan baru kepada aplikasi analisis sentimen dan dapat mengikuti perkembangan bahasa serta bahasan sehingga akurasi tetap terjaga. Skema penambahan data latih tersebut bukan tanpa kekurangan. Data latih di awal akan didesain seseimbang mungkin komposisi data latihan yang berlabel positif dan negatif. Bertambahnya data baru tidak bisa diprediksi apakah akan seimbang antara yang bernilai positif atau negatif. Sangat memungkinkan apabila data baru yang akan dijadikan data latih tersebut memiliki komposisi yang salah satunya lebih dominan.

Diperlukan suatu mekanisme agar data latih dapat leluasa terus bertambah tanpa harus mengkhawatirkan keseimbangan dari komposisi data latih. Seperti yang telah dibahas pada latar belakang penelitian ini, SVM adalah algoritma yang memiliki akurasi paling tinggi diantara algoritma lainnya yang bisa digunakan untuk melakukan analisis sentimen. Akurasi algoritma SVM sangat bergantung kepada penentuan *hyperplane* yang memisahkan dua kelompok data yang tersebar. Sebaran

data tersebut terjadi setelah dilakukan pembobotan kata pada masing-masing dokumen. Maka penentuan teknik pembobotan yang sesuai dengan kondisi data latih akan menentukan akurasi dari hasil klasifikasi.

SVM – TFIDFOR adalah algoritma dari hasil analisis dari permasalahan yang telah dibahas. Algoritma ini melakukan analisis sentiment yang terdiri dari algoritma SVM dan dua teknik pembobotan yaitu TF-IDF dan TF-OR. Algoritma TF-OR memiliki keunggulan dalam melakukan pembobotan lebih proporsional di saat kondisi data latih seimbang dikarenakan TF-OR mempertimbangkan kemunculan kata dalam dokumen yang berlabel positif atau negatif. Algoritma ini akan menghasilkan akurasi yang lebih tinggi ketika digunakan oleh algoritma SVM ketika kondisi data latih sedang seimbang. Di sisi lain, algoritma TF-IDF melakukan pembobotan tanpa mempertimbangkan kemunculan kata tersebut dalam dokumen positif atau negatif. Sehingga TF-IDF akan menghasilkan akurasi yang lebih tinggi ketika digunakan oleh algoritma SVM meskipun dalam kondisi data latih tidak seimbang. Berikut adalah skema yang diusulkan dalam penelitian ini untuk penambahan data latih dengan mempertimbangkan keseimbangan data latih.



Gambar 3-2Skema SVM - TFIDFOR

Penelitian ini berfokus pada pemilihan teknik pembobotan yang tepat untuk menghadapi kondisi data latih yang sedang digunakan. Seperti yang telah dibahas

pada subbab 3.1 bahwa dengan penambahan data latih yang terus menerus memberikan banyak kemungkinan komposisi data latih. Selanjutnya akan dibahas lebih lanjut mengenai data penelitian yang digunakan dan detil dari alur pengujian.

3.1.1 Data Penelitian

Analisis sentimen secara garis besar terbagi menjadi dua tahapan yaitu *training* dan *testing*. Dalam penelitian ini, dalam setiap tahapan dari analisis sentimennya diperlukan data yang berbeda. Penelitian ini berbeda dengan penelitian sebelumnya yang menggunakan *cross validation* sebagai metode evaluasinya di mana sebagian data latih dijadikan sebagai data uji, apabila *cross validation* digunakan maka akan terdapat kesulitan dalam mengatur komposisi data latihnya. Pada penelitian ini akan digunakan data uji yang sama dihadapkan kepada beberapa data latih dengan komposisi yang beragam.

Dataset yang digunakan dalam penelitian ini bersumber dari hasil ulasan film dari imdb. Pada dataset tersebut terdapat 25.000 data latih dan 25.000 data uji. Masing-masing data latih dan data uji telah diberi label kelas di mana komposisi dari keduanya seimbang, masing-masing di dalamnya terdapat 12.500 dokumen berlabel positif dan negatif. Secara keseluruhan dari 25.000 data latih terdapat 32.565.670 kata dan apabila dirata-ratakan maka terdapat 1.303 kata per dokumen. Jumlah terbanyak yang dimiliki dalam satu dokumen adalah 13.593 kata dan jumlah paling sedikit dalam satu dokumen adalah 52 kata. Bentuk asli dari dataset ini adalah berkas teks, kemudian ditransformasikan ke dalam basis data untuk memudahkan untuk pengolahan datanya. Berikut adalah struktur dan contoh isi dari dokumen dari dataset yang digunakan.

Tabel 3-1 Struktur dan contoh isi dokumen

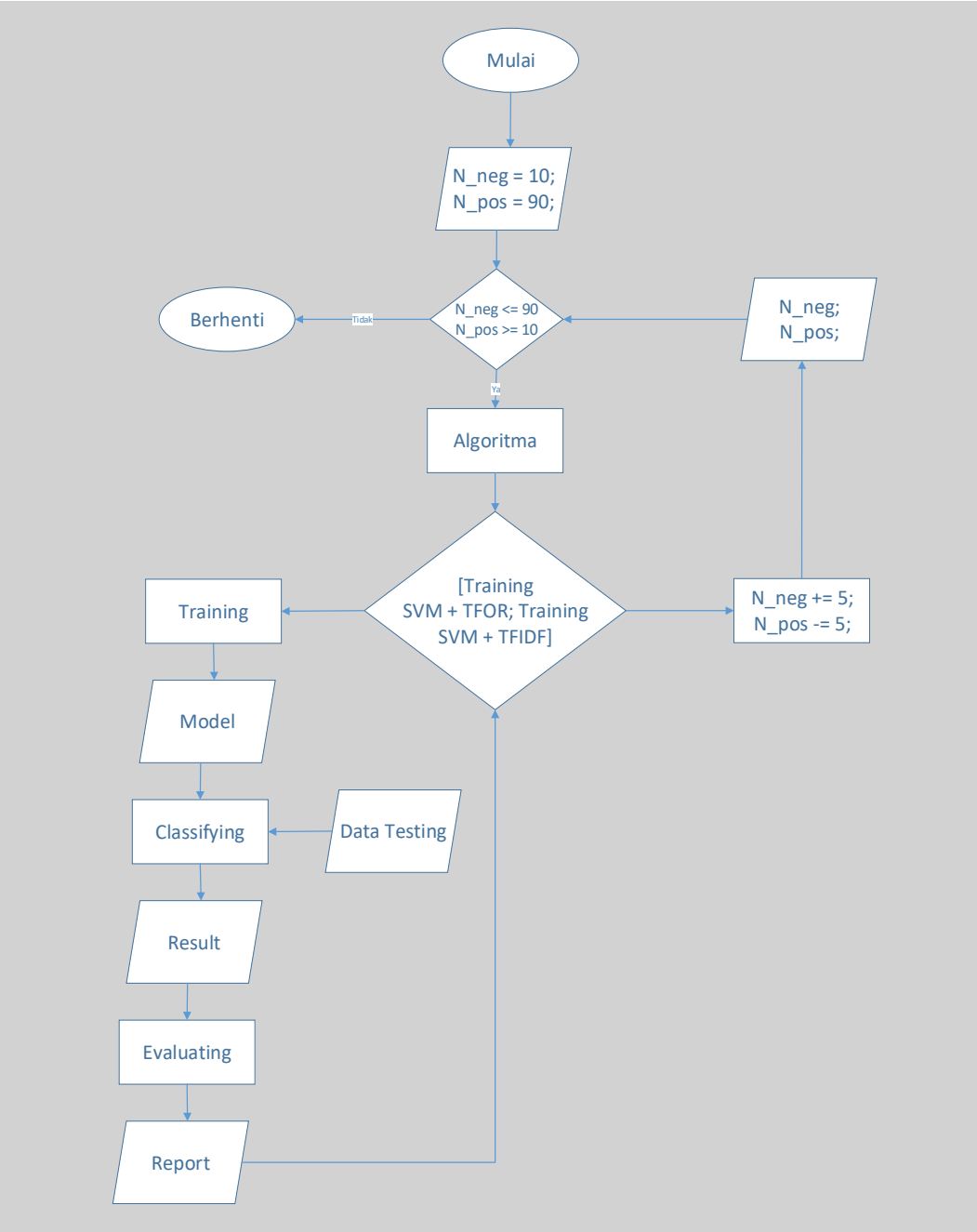
id	Isi dokumen	label
75014	…Saya merasa heran saat monton film ini di mana adegan argumennya begitu datar dan tanpa emosi, film ini sangat membosankan meski diperankan oleh aktor yang terkenal….	0
87510	…Karakterfilm yang diperankan oleh Robin Williams akhir-akhir ini begitu fantastis, terutama *insomnia*dan *one hour photo*.Pada film *the night listener*, meski ulasannya biasa-biasa saja dan *release* DVD yang cepat, ini adalah salah satu karya terbaiknya…	1

3.1.2 Alur Pengujian

Seperti yang telah dibahas pada subbab 2.1 bahwa analisis sentimen ini secara garis besar terbagi ke dalam dua tahapan, yaitu *training* dan *testing*. Maka dalam penelitian ini, tahapan *training* akan menjadi variabel *independent* dan tahap *testing* akan menjadi variable *dependent*. Ukuran yang digunakan dalam pengujian ini adalah akurasi. Pengujian yang dilakukan dalam penelitian ini terdiri dari tiga jenis pengujian yang meliputi aspek komposisi data latih, rasio data latih terhadap data uji, dan kuantitas dari data latih dan data uji. Ketiga pengujian tersebut saling berkaitan, dikarenakan hasil pengujian komposisi data latih akan digunakan pada pengujian rasio data latih terhadap data uji. Begitu juga hasil dari pengujian rasio data latih terhadap data uji akan digunakan dalam pengujian kuantitas dari data latih dan data uji.

3.1.2.1 Pengujian Komposisi Data Latih

Pengujian komposisi data latih ini berupa perubahan komposisi data latih yang akan digunakan dalam proses *training*. Komposisi ideal yang biasa digunakan dalam penelitian-penelitian sebelumnya adalah 50% data latih berlabel positif dan 50% data latih berlabel negatif. Pada penelitian ini akan dilakukan perubahan progresif pada data latih di mulai dengan komposisi 10% data latih berlabel positif dan 90% data latih berlabel negatif hingga komposisi data latih sebaliknya. Tujuan dari pengujian ini adalah untuk mengetahui teknik pembobotan mana di antara TF-IDF dan TF-OR yang akan membantu SVM memperoleh akurasi terbaiknya pada komposisi data latih yang telah ditentukan. Berikut adalah alur dari pengujian komposisi data latih.



Gambar 3-3 Alur Pengujian Komposisi Data Latih

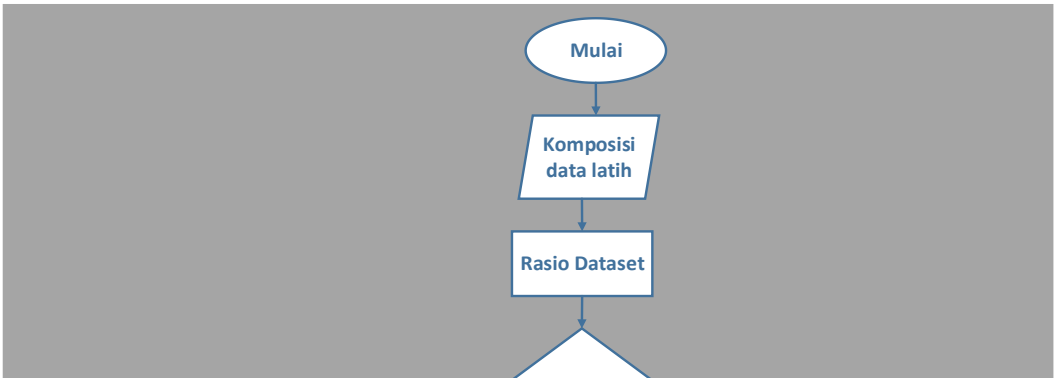
Gambar di atas menunjukkan bahwa komposisi data latih akan dimulai dengan 90% data latih positif dan 10% data latih negatif. Setidaknya akan dilakukan sebanyak 17 kali iterasi pengujian dengan perubahan komposisi sebesar 5%. Tentunya data latih berlabel negatif akan terus berkurang setiap iterasinya sedangkan data latih positif akan terus bertambah hingga batas yang telah ditentukan. Disetiap iterasi akan diperoleh masing-masing evaluasi menggunakan *confusion matrix* berdasarkan hasil klasifikasi dari masing-masing kombinasi. Untuk menguji tingkat

konsistensi dari hasil yang di dapat, maka 17 iterasi tersebut akan diulang sebanyak 10 kali dengan catatan akan diambil 1.000 dari 25.000 data latih yang tersedia, perkomposisinya akan diambil secara acak dan dari komposisi yang dibuat dari 17 iterasi tersebut, pada akhirnya akan dibuat nilai rata-ratanya.

3.1.2.2 Pengujian Rasio Dataset

Pengujian rasio dataset merupakan kelanjutan dari pengujian data latih. Pada pengujian ini akan diambil dua komposisi data latih. Komposisi yang diambil adalah komposisi yang menjadi nilai akurasi yang tertinggi dari dua teknik pembobotan yaitu TF-IDF dan TF-OR. Tujuan dari pengujian dataset ini adalah untuk mengetahui pengaruh perbandingan jumlah data latih terhadap jumlah data uji yang digunakan. Kondisi ideal yang diharapkan untuk mengetahui modal minimal (jumlah dataset) dalam mendapatkan akurasi di atas adalah 80%.

Rasio yang digunakan dalam pengujian ini terdiri dari lima kondisi. Rasio perbandingan data latih terhadap data uji dimulai dari 1:100, 1:10, 1:1, 10:1, dan 100:1. Pada pengujian ini dibatasi rasio terbesar dengan margin 1:100 atau sebaliknya karena keterbatasan data, dan jumlah minimal data latih atau data uji yang akan digunakan dalam pengujian ini adalah 100 dokumen. Berikut adalah gambaran mengenai alur pengujian rasio dataset.

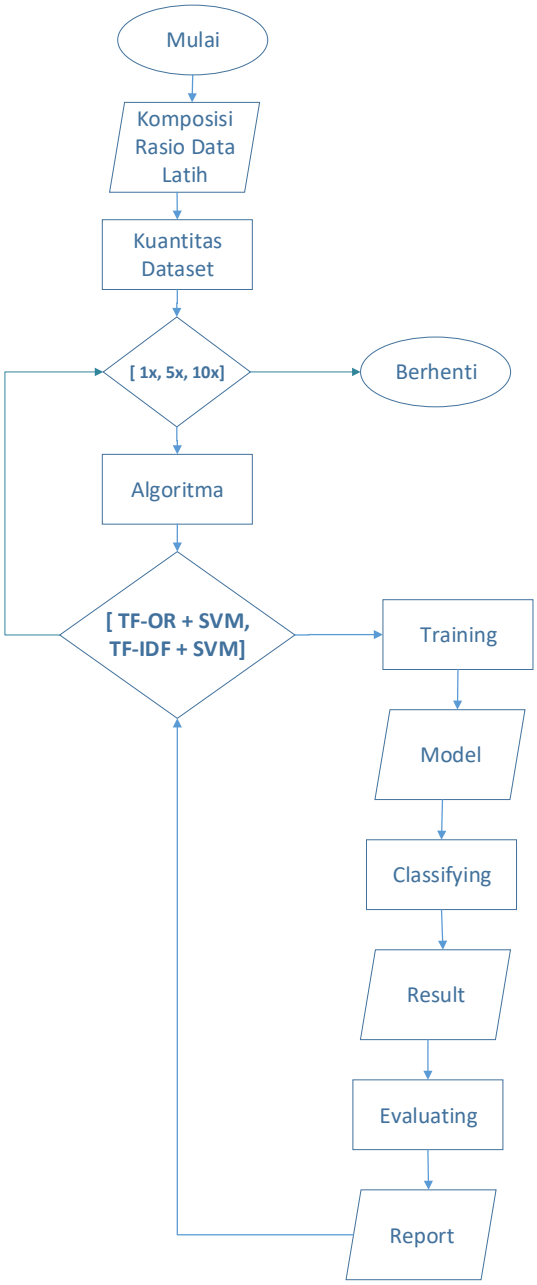


Gambar di atas menunjukan bahwa setiap komposisi data latih yang terpilih akan dilakukan proses pengujian dengan kondisi rasio dataset. Selanjutnya akan dipilih rasio yang paling minimal dari jumlah data latih yang digunakan untuk mencapai tingkat akurasi 80% pada pengujian selanjutnya.

3.1.2.3 Pengujian Kuantitas Dataset

Pengujian kuantitas dataset ini merupakan pengujian terakhir yang dilakukan. Pengujian ini menggunakan rasio dataset yang sudah terpilih untuk selanjutnya

dilakukan peningkatan kuantitas dataset sebanyak tiga kondisi yaitu 1 kali, 5 kali, dan 10 kali. Berikut adalah alur dari pengujian kuantitas dataset.



Gambar di atas menunjukkan bahwa dua dari rasio komposisi data latih yang telah terpilih dari pengujian sebelumnya akan dilakukan pengujian sebanyak tiga kondisi yaitu dengan penaikan data latih dan data uji sebesar 1 kali, 5 kali, dan 10 kali. Tujuan dari pengujian ini adalah untuk melihat apakah akan didapatkan pengaruh dari rasio yang telah terpilih dengan jumlah data yang paling minimal hingga

diperbanyak sebesar 10 kali terhadap tingkat akurasi. Setiap kondisi peningkatan jumlah dataset, dilakukan percobaan sebanyak 10 kali untuk melihat tingkat konsistensi dari akurasi dikarenakan data diambil secara acak.