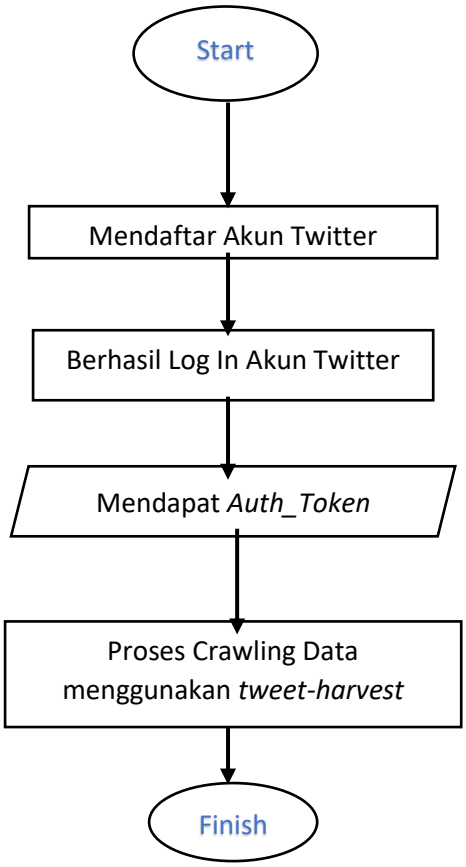
# BAB III
# METODOLOGI PENELITIAN

## 3.1 Metodologi SEMMA

Penelitian ini menggunakan metode SEMMA yang dibangun oleh SAS Institute, sebuah perusahaan perangkat lunak. SEMMA merupakan metode yang mudah dipahami dan digunakan sebagai referensi dalam suatu proyek data mining. Menurut akronimnya, SEMMA memiliki 5 langkah dalam penemuan data yaitu *Sample, Explore, Modify, Model dan Asses*. SEMMA akan memungkinkan pengguna untuk lebih mudah menerapkan teknik statistik dan visualisasi untuk mencari atau menemukan, memilah dan merubah variabel prediksi yang paling penting, kemudian variael dimodelkan untuk memperkirakan hasil yang berbeda dan memvalidasi kebenaran suatu model. Berikut gambaran yang akan dilalui pada metode SEMMA:

**a. *Sample***

Pada tahap ini pengumpulan data dilakukan berdasarkan topik penelitian terdahulu yang mendasari penelitian. Pengumpulan data dari Twitter juga dilakukan. Peneliti mengambil data berupa tweet yang berisi kata "*Marketplace* Guru" dan kemudian tweet tersebut disimpan dalam format file .csv. Proses pengumpulan data twitter atau crawling data di sini menggunakan *tweet-harvest*. Berikut *flowchart* alur pengambilan data pada Twitter:



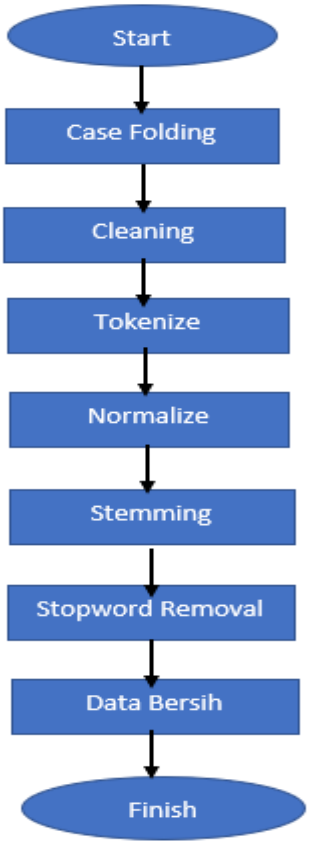

Gambar 3.1 *flowchart* alur *crawling* data

**b. *Explore***

Pada tahap ini peneliti mendeskripsikan data yang diperoleh dari hasil *crawling data* dengan menggunakan kata kunci "*Marketplace* Guru". Kemudian dilakukan langkah visualisasi data, yaitu data tweet yang diperoleh dibuat grafiknya dalam kurun waktu 67 hari untuk menampilkan informasi data tersebut secara visual.

**c. *Modify***

Dalam tahap ini langkah-langkah preprocessing akan dijalankan. Tahapan yang dilakukan seperti *case folding, cleaning, tokenize, normalize, stopword removal, dan stemming*.




Gambar 3.2 *flowchart* tahapan *modify*

1. *Case Folding*

*Case Folding* adalah tahapan dalam *pre-procesing* yang tugasnya mengubah semua karakter menjadi *lower case* atau huruf kecil. Tahapan biasa digunakan pada data teks, karena dapat membantu mengurangi format data dengan mengurangi jumlah kata atau token unik. Pada *Case Folding* menggunakan *Transform Case* dari *RapidMiner*, tools ini merubah semua karakter dalam teks menjadi huruf kecil pada dokumen teks yang akan digunakan.

Tabel 3.1 Contoh *Case Folding*

| Before | After |
|---|---|
| @collegemenfess Susah nder, aplg ktnya skrg sistemnya mau di ubah kan ke ky marketplace itu. Padahal guru yg mendidik, kerja serius, gaji bercanda bgt | @collegemenfess susah nder, aplg ktnya skrg sistemnya mau di ubah kan ke ky marketplace itu. padahal guru yg mendidik, kerja serius, gaji bercanda bgt |

2. *Cleaning*

Cleanning membantu mengidentifikasi dan memperbaiki kesalahan yang timbul saat memproses data dari berbagai sumber. Dengan menghilangkan inkonsistensi dan ketidakakuratan, pembersihan data dapat menjamin keandalan kumpulan data. Dimana data Twitter sendiri pasti memiliki banyak data kotor seperti *hashtag*, angka, *username*, URL, dan teks retweet. Data yang digunakan memerlukan pembersihan data seperti simbol *link* URL, angka.

Tabel 3.2 Contoh *Cleaning*

| Before | After |
|---|---|
| @collegemenfess susah nder, aplg ktnya skrg sistemnya mau di ubah kan ke ky marketplace itu. padahal guru yg mendidik, kerja serius, gaji bercanda bgt | susah nder, aplg ktnya skrg sistemnya mau di ubah kan ke ky marketplace itu. padahal guru yg mendidik, kerja serius, gaji bercanda bgt |

3. *Tokenize*

Pada tahap ini data akan diperiksa keseluruhan kemudian melakukan proses pemenggalan kalimat menjadi kata per kata.

Tabel 3.3 Contoh *Tokenize*

| Before | After |
|---|---|
| susah nder, aplg ktnya skrg sistemnya mau di ubah kan ke ky marketplace itu. padahal guru yg mendidik, kerja serius, gaji bercanda bgt | ‘susah’ ‘nder,’ ‘aplg’ ‘ktnya’ ‘skrg’ ‘sistemnya’ ‘mau’ ‘di’ ‘ubah’ ‘kan’ ‘ke’ ‘ky’ ‘marketplace’ ‘itu.’ ‘padahal’ ‘guru’ ‘yg’ ‘mendidik,’ ‘kerja’ ‘serius,’ ‘gaji’ ‘bercanda’ ‘bgt’ |

4. *Normalize*

*Normalize* adalah proses membakukan kata-kata yang mempunyai makna yang sama dengan cara mengubah ejaan akronim dan tidak baku sehingga mempunyai kesatuan makna.

Contoh :

Aplg : apalagi

Ktnya : katanya

Skrg : sekarang

Bgt : banget

Tabel 3.4 Contoh *Normalize*

| Before | After |
|---|---|
| ‘susah’ ‘nder,’ ‘aplg’ ‘ktnya’ ‘skrg’ ‘sistemnya’ ‘mau’ ‘di’ ‘ubah’ ‘kan’ ‘ke’ ‘ky’ ‘marketplace’ ‘itu.’ ‘padahal’ ‘guru’ ‘yg’ ‘mendidik,’ ‘kerja’ ‘serius,’ ‘gaji’ ‘bercanda’ ‘bgt’ | ‘susah’ ‘nder,’ ‘apalagi’ ‘katanya’ ‘sekarang’ ‘sistemnya’ ‘mau’ ‘di’ ‘ubah’ ‘kan’ ‘ke’ ‘kayak’ ‘marketplace’ ‘itu.’ ‘padahal’ ‘guru’ ‘yg’ ‘mendidik,’ ‘kerja’ ‘serius,’ ‘gaji’ ‘bercanda’ ‘banget’ |

5. *Stopword Removal*

Langkah ini merupakan proses menyaring kata-kata umum seperti "dan", "atau" serta kata-kata lain yang tidak diperlukan saat mengolah data. Kata-kata ini dihapus dari tweet.

Tabel 3.5 *Stopword Removal*

| Before | After |
|---|---|
| 'susah' 'nder,' 'apalagi' 'katanya' 'sekarang' 'sistemnya' 'mau' 'di' 'ubah' 'kan' 'ke' 'kayak' 'marketplace' 'itu.' 'padahal' 'guru' 'yg' 'mendidik,' 'kerja' 'serius,' 'gaji' 'bercanda' 'banget' | 'susah' 'apalagi' 'katanya' 'sekarang' 'sistemnya' 'mau' 'ubah' 'kayak' 'marketplace' 'padahal' 'guru' 'mendidik,' 'kerja' 'serius,' 'gaji' 'bercanda' 'banget' |

6. *Stemming*

*Stemming* adalah proses yang digunakan untuk menelusuri asal kata. Hal ini juga bertujuan untuk membersihkan kata yang ejaannya kurang tepat. *Stemming* dilakukan untuk mendapatkan kata dasar dengan cara menghilangkan imbuhan yang terdapat pada kombinasi awalan, sisipan, akhiran, dan imbuhan suatu kata (P. Arsi, 2021). Pada tahapan stemming ini melibatkan algoritma porter dimana proses stemming yang dilakukan terdapat lima langkah dengan mensimulasikan *infleksi* dan *derivasi* dari kata. Pada setiap langkahnya, sebuah akhiran tertentu dihapus dengan cara subtitusi (Arie Atwa Magriyanti, n.d.). Lima langkah dalam algoritma porter yaitu Menghilangkan akhiran kepemilikan, Mengilangkan imbuhan pada awalan pertama. Jika tidak ada, langsung melanjutkan menghilangkan awalan, jika ada, maka dicari kemudian lanjutkan ke langkah menghilangkan akhiran, jika tidak ditemukan maka, kata tersebut diasumsikan sebagai kata dasar. Jika ditemukan maka lanjutkan menghapus awalan ke dua, Kemudian kata akhir diasumsikan sebagai kata dasar (Agusta et al., 2009)

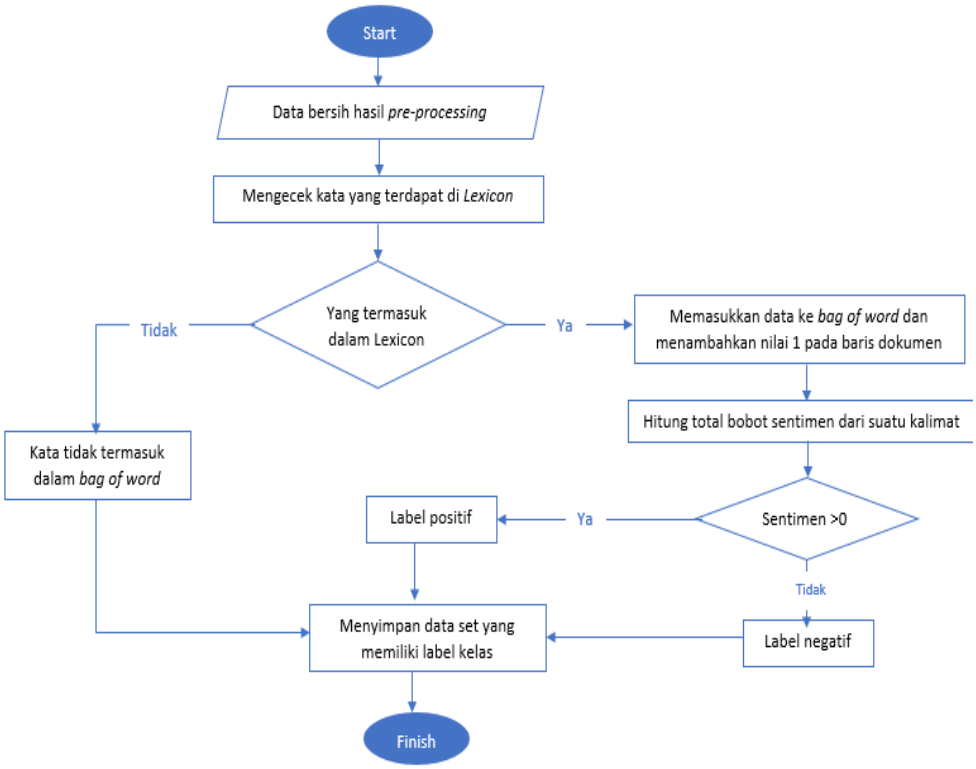
Setelah melewati tahapan diatas maka akan didapatkan kata yang telah bersih dari imbuhan awalan dan akhiran sebagai berikut:

Tabel 3.6 *Stemming*

| Before | After |
|---|---|
| ‘susah’ ‘apalagi’ ‘katanya’ ‘sekarang’ ‘sistemnya’ ‘mau’ ‘ubah’ ‘kayak’ ‘marketplace’ ‘padahal’ ‘guru’ ‘mendidik,’ ‘kerja’ ‘serius,’ ‘gaji’ ‘bercanda’ ‘banget’ | ‘susah’ ‘apalagi’ ‘kata’ ‘sekarang’ ‘sistem’ ‘mau’ ‘ubah’ ‘kayak’ ‘marketplace’ ‘padahal’ ‘guru’ ‘didik,’ ‘kerja’ ‘serius,’ ‘gaji’ ‘canda’ ‘banget’ |

**d. *Model***

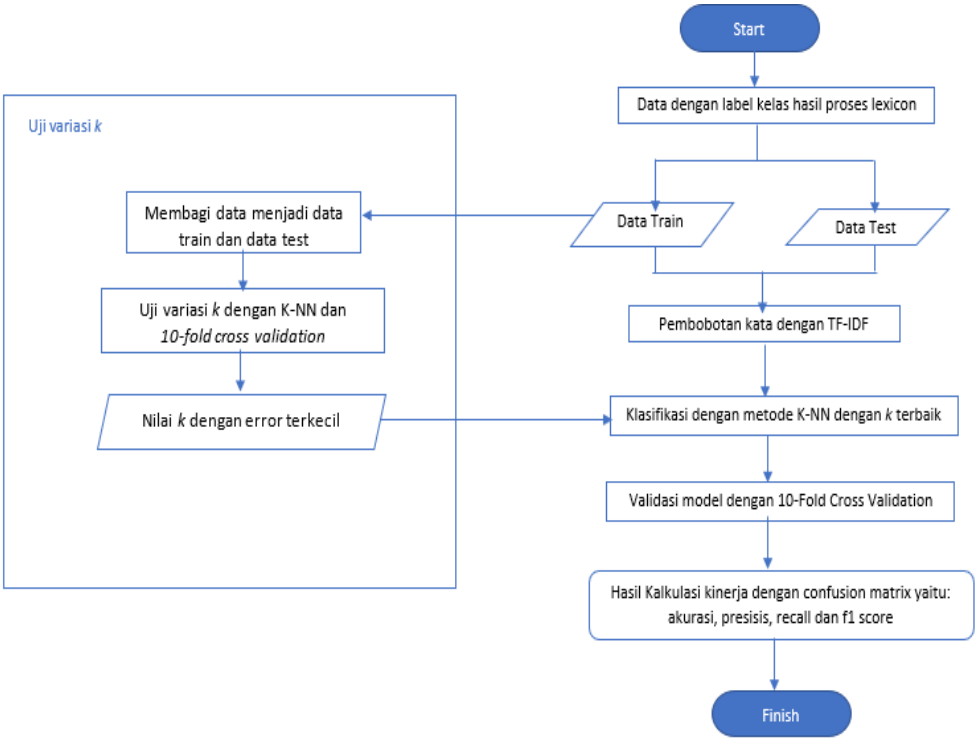
Pada titik ini, kumpulan data yang difilter dibagi menjadi kelas positif, netral dan negatif dengan metode *lexicon*. Kumpulan data ini kemudian ditampilkan sebagai wordcloud berdasarkan setiap lapisan. Kemudian untuk pengolahan datanya, dataset akan dibagi menjadi dua yakni data latih dan data uji. Data ini kemudian dibobotkan per kata menggunakan TF-IDF. Kemudian dilakukan pengolahan data untuk setiap metode algoritmik. Untuk metode *K-Nearest Neighbor*, hal pertama yang dilakukan adalah mencari varian terbaik dari nilai k dengan cara menguji beberapa nilai k, kemudian memilih nilai k dengan nilai error terkecil, dan menguji algoritma. Jadi adanya metode lexicon dalam penelitian ini ialah untuk membantu metode K-NN dalam melabelkan data menjadi kelas positif, netral dan negatif.




Gambar 3.3 *flowchart* tahapan *model*

**e. *Asses***

Merupakan tahapan terakhir dalam metode SEMMA. Asses merupakan tahapan menilai data dengan mengevaluasi kegunaan dan keandalan hasil proses data mining dan mengevaluasi seberapa baik kerjanya. Pada langkah ini, evaluasi metode *K-Nearest Neighbor* menggunakan 10 *folds cross validation* untuk mengetahui akurasi, presisi, *recall* dan *f1-score* dari model yang digunakan.

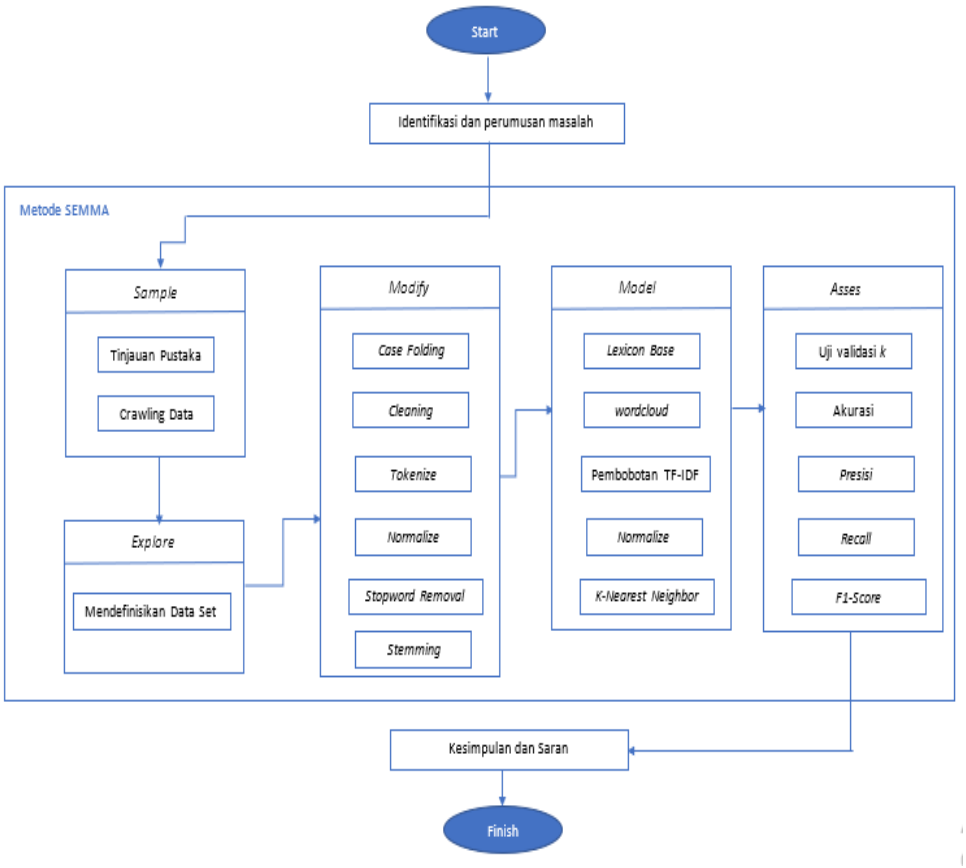



Gambar 3.4 *flowchart* tahapan *asses*

## 3.2 Tahapan Penelitian

Tahapan penelitian ini, penulis mengawali penelitian dengan mengidentifikasi dan merumuskan permasalahan yang ada, kemudian dilanjutkan dengan pemilihan metode penelitian yaitu pemilihan metode SEMMA yang mempunyai 5 tahapan yaitu *sample, explore, modify, model dan asses*. Tahap *sample* merupakan tahap dimana penulis melakukan tinjauan pustaka dan mengindeks data di Twitter. Tahap *explore* adalah fase dimana data terindeks yang digunakan dalam dataset ditentukan dan dibuat grafik tweet yang diurutkan berdasarkan waktu. Didalam tahap *modify* atau *preprocessing* data terdapat langkah-langkah preprocessing seperti yang sudah dijelaskan sebelumnya yaitu *case folding, cleansing, tokenizing, normalize, stopword removal, dan stemming*, dilakukan agar dataset lebih terstruktur dan mudah diolah untuk tahap selanjutnya.. Tahap pemodelan, dimana data diberi pengenal kelas menggunakan metode berbasis kosakata

dan pembobotan TF-IDF, kemudian diolah menggunakan metode pengklasifikasi *K-nearest neighbour*. Dan langkah evaluasi model atau evaluasi kinerja terdiri dari akurasi, presisi, *recall* dan *f1-score*. Pada tahap terakhir dilakukan pembahasan mengenai hasil yang diperoleh dan kesimpulan serta usulan yang dibuat dari penelitian. Berikut gambaran proses penelitian yang akan dilakukan:




Gambar 3.5 Tahapan Penelitian