BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Twitter sering digunakan oleh peneliti sebagai objek riset, seperti riset pengklasifikasian pengguna dan pengklasifikasian sentimen. Dalam risetnya, peneliti memanfaatkan kemudahan dalam pengaksesan twitter karena twitter telah menyediakan *Application Programming Interface (API)* untuk mengambil data kicauan yang dibutuhkan peneliti. Klasifikasi sentimen biasa dilakukan dengan menggunakan algoritma pembelajaran mesin *supervised* yaitu menggunakan data berlabel dalam proses pembelajaran.

Berbagai penelitian klasifikasi sentimen telah dilakukan dengan pengklasifikasi pembelajaran mesin *supervised* antara lain *Support Vector Machine, Naïve Bayes, Linear Regression, Decision Tree, Multinomial Naive Bayes* dan sebagainya. Penelitian yang melakukan perbandingan terhadap pengklasifikasi-pengklasifikasi tersebut juga sudah banyak dilakukan oleh peneliti untuk mencari pengklasifikasi yang terbaik. Sebagaimana yang dilakukan oleh Ang Yang, Jun Zhang, *et al*. (2015) membandingkan kinerja enam pengklasifikasi pembelajaran mesin *supervised* yang berbeda untuk menguji skema kombinasi fitur baru yang dikembangkan yaitu *unigrams* dan *sentiment lexicon*, menunjukkan bahwa

Multinomial Naive Bayes memiliki akurasi paling tinggi dibanding pengklasifikasi lainnya yaitu sebesar 84,60 % sehingga menjadikan *Naive Bayes Multinomial Classifier* sebagai pilihan terbaik untuk melakukan klasifikasi sentimen pada twitter dengan fitur kombinasi tersebut. Kemudian penelitian lain yang dilakukan oleh Bal Krishna Bal, *et al* pada tahun 2016 untuk klasifikasi sentimen bahasa Nepal juga menunjukkan bahwa *Multinomial Naïve Bayes* merupakan pengklasifikasi paling sesuai untuk klasifikasi sentimen, dalam penelitian tersebut membandingkan empat fitur ekstraksi yaitu *Bag-of-Words, Bag-of-Words (stop removed), TF-IDF* dan *TF-IDF (stop removed)*. Maka berdasarkan dari dua penelitian tersebut dapat disimpulkan bahwa *Multinomial Naïve Bayes* tampil lebih baik bahkan dengan menggunakan fitur apapun dibandingkan dengan pengklasifikasi pembelajaran mesin *supervised* lainnya.

Seperti yang telah diketahui dengan baik bahwa untuk mencapai tingkat akurasi yang tinggi dalam algoritma pembelajaran mesin *supervised* diperlukan sejumlah data berlabel yang cukup besar, namun untuk mendapatkan data yang berlabel merupakan pekerjaan yang tidak mudah karena karakteristiknya yang mahal dan perlu usaha manual dalam pemberian labelnya, untuk klasifikasi sentimen terbagi menjadi label positif dan negatif. Sedangkan sebaliknya data yang tidak berlabel mudah didapatkan dan banyak tersedia, untuk klasifikasi sentimen data tidak berlabel dapat diambil dari twitter. Oleh karena itu perlu pendekatan *semi-supervised* agar pemberian label dapat dilakukan secara otomatis untuk seluruh

data yang tidak berlabel dengan menggunakan data berlabel kecil yang ada sebagai tolak ukur untuk membandingkannya, dengan begitu pekerjaan untuk mengumpulkan data berlabel pun dapat dikurangi.

Algoritma *Co-Training* merupakan salah satu algoritma dari pembelajaran *semi-supervised* yang dapat diterapkan dalam klasifikasi sentimen. Seperti dalam penelitian Shenghua Lie, *et al* pada tahun 2014 yang menerapkan model multiclass SVM dalam skema *Co-Training* untuk melakukan klasifikasi sentimen lintas domain pada kicauan twitter, dan mereka mengusulkan sebuah topik semi-sentimentil *Topic-Adaptive Sentiment Classification* (TASC). TASC-t dirancang untuk menyesuaikan diri dengan *timeline* kicauan yang dinamis. Algoritma yang diterapkan mencapai peningkatan yang menjanjikan dalam akurasi rata-rata pada 6 topik dari korpus kicauan publik.

Algoritma *Co-Training* menggunakan dua pengklasifikasi dan dua pandangan yang independen. Sebagaimana yang disampaikan oleh Ragini Gokhale *and* Maria Fasli (2017) dalam tulisannya bahwa gagasan di balik *Co-Training* adalah satu pengklasifikasi dapat menetapkan label yang benar ke beberapa contoh tertentu namun mungkin sulit bagi pengklasifikasi lainnya untuk melakukannya. Oleh karena itu, setiap pengklasifikasi bersama dapat meningkatkan kumpulan pelatihan dengan label yang sangat informatif untuk pengklasifikasi lainnya.

Berdasarkan penjelasan diatas dan penelitian Shenghua Lie, *et al* (2014) dan Ang Yang, Jun Zhang, *et al*. (2015), penelitian klasifikasi sentimen pada twitter

kali ini akan menerapkan *Multinomial Naïve Bayes* sebagai dasar pengklasifikasi kedalam skema algoritma *Co-Training*. Dimana fitur diekstrak dan dipecah menjadi dua pandangan, yaitu fitur *text* dari kata-kata sentimen dan *non-text* dari *emoticon* dan tanda baca. Sehingga diharapkan dapat menghasilkan akurasi yang lebih baik dibandingkan dengan pembelajaran mesin *supervised* yang hanya melibatkan data berlabel untuk klasifikasi sentimen pada twitter.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang dikemukakan diatas, rumusan masalah dalam penelitian ini adalah bagaimana menerapkan *Multinomial Naïve Bayes Classifier* sebagai dasar pengklasifikasi dalam algoritma *Co-Training*.

Untuk menyelesaikan permasalahan yang disebutkan diatas, maka dalam penelitian ini dibagi dalam beberapa *research question* antara lain:

1. Bagaimana cara kerja *Multinomial Naïve Bayes Classifier* sebagai pembelajaran mesin *supervised?*
2. Bagaimana cara kerja algoritma *Co-Training* ?
3. Pada bagian mana *Multinomial Naïve Bayes Classifier* diimplementasikan dalam skema algoritma *Co-Training* ?

4. Apa yang menjadi parameter untuk perhitungan kedua pengklasifikasi pada algoritma *Co-Training* terhadap dua pandangan yang dipisah *text* dan *non-text* ?
5. Bagaimana hasil akurasi algoritma *Co-Training* dengan pengklasifikasi *Multinomial Naïve Bayes* terhadap klasifikasi sentimen pada twitter ?

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah :

1. Mengetahui cara kerja *Multinomial Naïve Bayes Classifier* sebagai pembelajaran mesin *supervised.*
2. Mengetahui cara kerja algoritma *Co-Training*.
3. Mengetahui di bagian mana *Multinomial Naïve Bayes Classifier* diimplementasikan dalam skema algoritma *Co-Training.*
4. Mengetahui parameter untuk perhitungan kedua pengklasifikasi pada algoritma *Co-Training* terhadap dua pandangan yang dipisah *text* dan *non-text.*
5. Mengetahui hasil akurasi algoritma *Co-Training* dengan pengklasifikasi *Multinomial Naïve Bayes* terhadap klasifikasi sentiment pada twitter.

1.4 Manfaat Penelitian

Manfaat penelitian ini adalah mendapatkan hasil klasifikasi sentimen yang lebih baik dengan usaha yang minimal untuk membantu seseorang mengetahui kecenderungan suatu topik pada twitter apakah positif atau negatif.

1.5 Batasan Masalah

Batasan masalah dalam penelitian ini adalah :

1. Kicauan yang diambil berupa kicauan berbahasa Indonesia, karena kaidah pada praproses diatur dalam tata Bahasa Indonesia.
2. Metode klasifikasi yang digunakan adalah *Multinomial Naive Bayes Classifier*.
3. Algoritma pembelajaran *semi-supervised* yang digunakan adalah *Co-Training*.
4. Dua pandangan yang digunakan dalam algoritma *Co-Training* yaitu *text* dan *non-text.*