BAB I

PENDAHULUAN

1.1 Latar Belakang

Data Mining merupakan kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan , pola dan hubungan dalam dataset berukuran besar [1]. Dalam data mining sendiri terdapat beberapa algoritma yang bisa digunakan untuk melakukan permodelan klasifikasi contohnya seperti C4.5, K-Mean, Support Vector Machine, dam Naïve bayess classfier. Algoritma tersebut sudah banyak di terapkan di berbagai domain dan sukses menghasilkan akurasi yang maksimal.

Dalam *data mining* kita juga mengenal *Text mining* yang mengacu pada proses pengambilan informasi berkualitas tinggi dari sebuah atau beberapa *sample text*. Text mining sendiri merupakan suatu proses penambangan data berupa teks yang di lakukan oleh komputer dimana data tersebut dapat memberikan informasi-informasi untuk dilakukan analisa keterhubungannya [2]. Pembelajaran sebuah pola statistik bisa digunakan untuk menemukan informasi berkualitas tinggi melalui peramalan pola dan kecendrungan sarana pada sebuah *text*. Proses umum pada sebuah *text mining* biasanya meliputi pengelompokan *text*, *text clustering*, pengambilan intisari konsep/entitas, produksi taksonomi granular,

sentiment analysis, penyimpulan dokumen, serta permodelan relasi entitas (pembelajan hubungan antara entitas berlabel).

Sentiment Analysis dalam *text mining* adalah bidang studi yang menganalisis pendapat seseorang, *sentiment* seseorang, evaluasi seseorang, sikap seseorang dan emosi seseorang ke dalam bahasa tertulis [3]. Studi *Analysis Sentiment* sendiri dapat digunakan untuk membantu kita mengambil keputusan paling tepat guna menyelesaikan satu atau lebih skenario yang kita temukan. Dalam skenario kali ini peneliti juga membagi *sentiment analysis* menjadi tiga *class attribute*. Dimana ada label positif, label negatif, dan label netral pada setiap classnya.

Beberapa penelitian yang telah di lakukan sebelumnya tentang sentimen analisis diantaranya adalah penelitian yang membahas tentang analisis sentimen pasar otomotif mobil pada *tweet* Twitter dengan metode *Naïve Bayes* dimana pada penelitian tersebut bertujuan untuk mengetahui merek mobil terlaris di Twitter. Hasil penelitian ini menunjukkan bahwa tingkat akurasi *Naïve Bayes* yaitu 93% [2]. Penelitian kedua yaitu sentimen analisis *tweet* berbahasa Indonesia dengan *Deep Belief Network* (DBN) dan hasil DBN dibandingkan dengan Naïve Bayes. Tujuan penelitian tersebut adalah untuk mengetahui hasil sentimen terhadap *tweet* berbahasa Indonesia di Twitter dan hasil akurasi DBN sebesar 93.31%, Naïve Bayes sebesar 79.10 [4].

Pada awal penelitian penulis mengambil data dari microbloging *Twitter*. Data yang di ambil merupakan hasil dari *crawling tweet* opini masyarakat tehadap

pembangunan infrastruktur di Indonesia selama rentang waktu 6 bulan ke belakang dari akhir tahun 2018 hingga pertengahan 2019. Untuk hak akses, data di ambil menggunakan API *Twitter* yang telah di ajukan sebelumnya oleh penulis guna mendapatkan *tokenize* dari *Twitter* itu sendiri.

Pembangunan infrastruktur sendiri merupakan kegiatan yang cocok di teliti mengingat setiap kegiatannya memiliki *output* yang patut di dinilai oleh masyarakat banyak. Pentingnya pembangunan infrastruktur juga dapat berdampak pada berbagai macam sektor ekonomi guna menunjang keberlangsungan hajat orang banyak. Tujuan akhir dari pembangunan infrastruktur ini seperti yang telah di sampaikan oleh pemerintah Indonesia sendiri adalah untuk memajukan segala aspek dalam menjalankan sebuah negara. Oleh karena itu setiap bentuk usaha pembangunan bisa di teliti dan di nilai sejauh mana kepentingan dari pembangunan tersebut.

Dari berbagai macam algoritma yang di gunakan di dalam sentimen analisis belum di temukan model paling tepat untuk menganalisa sentimen terhadap objek pembangunan infrastruktur di Indonesia. Maka dari itu peneliti melakukan komparasi terhadap dua buah algoritma klasifikasi yaitu support vector machine dan naïve bayes classifier untuk mendapatkan nilai akurasi terbaik dari analisa objek tersebut.

Metode yang digunakan dalam penelitian kali ini yaitu *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) dimana akan di lakukan perbandingan hasil terhadap kedua algoritma tersebut pada obyek yang akan di

teliti. Hasil akhir yang di pusatkan pada penelitian kali ini adalah untuk memperoleh metode mana yang lebih efektif untuk mendapatkan hasil akurasi pengolahan data. Di harapkan dari penelitian ini dapat menghasilkan suatu informasi yang berkualitas untuk pengambilan keputusan dalam suatu skenario.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah di kemukakan pada judul "Perbandingan Algoritma *Naïve Bayes Classifier* dan *Support Vector Machine* pada *Sentiment Analysis* opini pembangunan infrastruktur di media *social twitter*." Maka permasalahan yang akan di bahas pada penelitian kali ini adalah:

1. Bagaimana Algoritma *Naïve Bayes Classifier* dan *Support Vector Machine* dapat mengklasifikasikan tweet dengan baik?
2. Berapa nilai akurasi yang di hasilkan oleh Algoritma *Naïve Bayes Classifier* dan *Support Vector Machine*?
3. Apakah Algoritma *Naïve Bayes Classfier* lebih baik dari *Support Vector Machine* dalam hal nilai akurasi yang di hasilkan?
4. Berapa selisih nilai akurasi dari kedua model yang telah di bandingkan?

1.3 Batasan Masalah

Pada penelitian ini, akan di berikan batasan – batasan masalah untuk permasalahan yang ada guna mengarahkan penelitian sesuai dengan tujuannya. Berikut batasan masalah pada penelitian kali ini :

1. *Tweet* yang diambil dan digunakan merupakan *tweet* yang berbahasa Indonesia..
2. Penelitian mengambil objek berupa opini masyarakat tentang pembangunan infrastruktur yang telah dilakukan di Negara Indonesia.
3. Penelitian dilakukan pada bulan Desember 2018 sampai Juni 2019 dengan data *tweet* berjumlah 437 *tweet* yang di ambil menggunakan metode *crawling* pada microbloging *Twitter*.
4. *Tweet* yang di gunakan berupa *text* dan tidak mengandung gambar.
5. Algoritma yang di gunakan untuk mengukur akurasi pada hasil pengolahan data kali ini adalah *Naïve Bayes Classifier* dan *Support Vector Machine*.

1.4 Tujuan

Tujuan yang di kedepankan dalam penelitian kali ini adalah mengetahui proses dan alur dari memperoleh data pada microbloging *Twitter* untuk *Sentiment*

Analysis hingga peneliti dapat mengolah data tersbut guna mendapatkan nilai akurasi yang terbaik. Nilai akurasi di ukur melalui dua algoritma yaitu *Naïve Bayes Classifier* dan *Support Vector Machine*. Setelah mendapatkan nilai akurasi akan dilakukan perbandingan dan *analysis* terhadap nilai akurasi yang di peroleh.

1.5 Manfaat Penelitian

Manfaat yang di berikan dari penelitian ini adalah :

1. Mampu memberikan pengetahuan tentang bagaimana melakukan *sentiment analysis* serta cara mendapatkan informasi yang berkualitas dari suatu sumber melalui pengolahan data opini masyarakat.

2. Diharapkan dari hasil penelitian ini dapat menambah pengetahuan dan informasi mengenai opini yang berkembang di masyarakat melalui media sosial.

3. Mampu memberikan sumbangan ide dalam melakukan penelitian *sentiment analysis* dan perbandingan antara dua algoritma yaitu *Naïve Bayes Classifier* dan *Support Vector Machine*.

4. Membantu menganalisis *sentiment* pada *Twitter* dengan algoritma *Naïve Bayes Classifier* dan *Support Vector Machine*.

5. Mampu mangambil kepetusuan dalam kelanjutan perencanaan pembangunan infrastruktur di Indonesia.

1.6 Metodologi Penelitian

Metodologi yang dikerjakan oleh penulis dalam penelitian kali ini adalah metode pengumpulan data dan *Analysis Sentiment* menggunkan *Naïve Bayes Classifier* serta *Support Vector Machine* yang di laksanakan dengan tahapan-tahapan sebagai berikut:

1.6.1 Metode Pengumpulan Data

Dalam proses ini pengumpulan data dilakukan oleh penulis dengan cara sebagai berikut :

1. Metode Deskriptis

   Dalam metode ini penulis mencoba mendapatkan data dari microbloging *Twitter* dengan cara crawling data menggunakan API *Twitter* dari akun-akun acak berdasarkan *keyword* 'pembangunan infrastruktur'.

2. Metode Studi Pustaka

   Sebelum melakukan penelitian perbandingan antara *Naïve Bayes Classifier* dan *Support Vector Machine* penulis melakukan studi pustaka terlebih dahulu dengan cara mengumpulkan referensi baik dari buku, artikel, *paper*, jurnal, makalah, maupun situs internet. Nantinya setiap sumber yang telah di kumpulkan tersebut akan menunjang penelitian demi mendapatkan hasil yang optimal.

1.6.2 Metode Analisis

Dalam tahapan analisis, penulis melakukan sebagai berikut :

1. Penulis melakukan analisis tahapan demi tahapan secara sistematis dari proses *Sentiment Analysis* sehingga penulis memperoleh hasil akhir yang optimal dari penelitian ini.

2. Penelitian ini menganalisis setiap data yang di peroleh melalui cara kerja algoritma *Naïve Bayes Classifier* dan *Support Vector Machine* dalam mengklasifikasikan setiap data *tweet* yang di peroleh dari *crawling data* itu sendiri.

1.6.3 Perancangan

Pada tahap perancangan terdiri dari beberapa bagian tahapan yang memiliki masing-masing metode untuk *Analysis Sentiment*. Hal pertama yang di lakukan adalah melakukan *crawling* data dari *Twitter* menggunakan API yang telah disediakan oleh *Twitter* sendiri. Selanjutnya data tersebut di simpan dalam bentuk *file document*, lalu data yang telah di proses di bagi menjadi dua tipe data yaitu *data training* dan *data testing*. Berikut tahapan selanjutnya dalam data pengolahan *data training* dan *data testing* :

1. Pelabelan Manual

Data latih atau data training yang telah dikumpulkan dibagi menjadi dua kelompok negatif dan positif dengan melabelkan setiap data sesuai dengan karakter *Tweet* yang di peroleh. Setiap data yang tidak masuk pada *term negative* atau *term positive* akan dianggap netral dari data yang dikumpulkan.

2. Tokenizing

Tahapan *tokenizing* adalah pemotongan *document* atau *Twitter* setiap *user* menjadi potongan-potongan kecil yang disebut *token*. Pada tahapan ini juga setiap tanda baca yang terdapat dalam *Tweet* di hilangkan guna mendapatkan data yang bisa di terjemahkan oleh algoritma yang diimplementasikan.

3. Stopword Removal

Stopword removal merupakan tahapan dimana data dokumen yang telah di *tokenizing* akan di *filter* lagi menggunakan *stopword removal*. Fungsi dari *stopword removal* yaitu melakukan penghapusan kata-kata yang tidak dibutuhkan sehingga tidak mengganggu proses pencarian fitur ataupun pengklasifikasian.

4. Stemming

Steming adalah proses yang memungkinkan sebuah kata dihilangkan imbuhannya sehingga menemukan setiap kata dasar dari kata yang terdapat dalam sebuah opini. Proses ini merupakan kelanjutan dari proses *stopword removal*.

5. Feature List

Feture List merupakan tahapan terakhir dari proses *Preprocessing*. di tahap ini kita akan mendapatkan hasil akhir berupa *term positive* dan *term negative* yang dapat di implementasikan pada algoritma yang telah di pilih.

1.6.4 Implementasi

Pada tahap implementasi di lakukan penghitungan nilai akurasi dari hasil *preprocessing* yang telah dilalalui menggunakan algoritma *Naïve Bayes Classifier* dan *Support Vector Machine* pada data klasifikasi *Sentiment Analysis*.

1.6.5 Testing

Tahap ini merupakan pengujian terhadap implementasi yang telah di lakukan dengan membandingkan setiap data berdasarkan jumlah opini *negative* dan opini *positive* yang di variasikan kuantitas perbandingannya.

1.7 Sistematika Penulisan

Untuk membantu memahami setiap isi dalam karya tulis ilmiah ini, maka penulis mencoba membuat sistematika pennyusunan sebagai berikut :

BAB I PENDAHULUAN

Dalam bab ini penulis mencoba menjelaskan latar belakang masalah yang di temukan, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian, dan sistematika penelitian.

BAB II LANDASAN TEORI

Bab kedua ini akan menguraikan beberapa teori yang menjadi dasar dalam penelitian yang di lakukan. Landasan teori yang ada akan dapat menjelaskan definisi atau model yang berkaitan dengan penelitian ini secara detail. Pada bab ini juga akan di jelaskan tentang gambaran umum mengenai penelitian yang meliputi objek penelitian, data yang digunakan, *tools* atau *software* yang dipakai untuk melakukan penelitian.

BAB III ANALISIS

Di dalam bab ini akan diuraikan mengenai analisis terhadap data yang telah di peroleh melalui cara pembangunan metode sentiment analysis terhadap data itu sendiri sehinnga peneliti bisa mendapatkan nilai akurasi dari *term positive* dan *term negative* untuk penelitian ini.

BAB IV IMPLEMENTASI DAN PEMBAHASAN

Pada bab ini akan menjelaskan tentang implementasi analisis yang telah di lakukan kedalam algoritma yang dipilih. Selain itu setiap hasil dari algoritma yang ada akan dibandingkan dan dicari kelebihan serta kekurangan setiap proses algoritma yang ada.

BAB V PENUTUP

Bab terakhir adalah bab yang berisi tentang kesimpulan dari skripsis yang telah dibuat dan juga berisi saran yang di harapkan dapat bermanfaat bagi pengembangan penelitian dimasa yang akan datang.