BAB I
PENDAHULUAN

1.1 Latar Belakang

Dunia maya seperti layaknya media sosial merupakan sebuah revolusi besar yang mampu mengubah perilaku manusia, dimana relasi pertemanan serba dilakukan melalui medium digital menggunakan media baru (internet) yang dioperasikan melalui situs-situs jejaring sosial [1].

Twitter adalah salah satu situs *microblogging* paling populer yang dapat membagikan update status atau pesan yang tidak lebih dari 140 karakter kepada pengguna dalam satu jaringan. Pada pertengahan tahun 2010 Twitter memiliki lebih dari 106 juta pengguna diseluruh dunia dan terus meningkat setiap harinya sebanyak 300.000 pengguna dan Twitter setiap harinya mendapatkan lebih dari 3 juta request. Dari angka tersebut Indonesia menjadi negara yang menduduki peringkat 8 dalam mengakses situs Twitter. Twitter menerima *tweet* dari pengguna sebanyak 55 juta pesan setiap harinya [2]. Berdasarkan data tersebut, Twitter memiliki sumber data yang besar. Data dalam hal ini adalah tweet dari pengguna yang berjumlah sangat banyak. Hal ini merupakan sebuah sumber daya yang bisa kita manfaatkan untuk kepentingan tertentu misalkan untuk mengetahui berita COVID-19 di Indonesia berdasarkan *tweet* yang dikirim oleh pengguna yang berisi informasi berita COVID-19.

Pertumbuhan pengguna internet khususnya Twitter dari tahun ke tahun selalu meningkat cukup signifikan, hal tersebut sangat berdampak pada peristiwa penyebaran.

berita bohong atau hoax yang kian marak diperbincangkan oleh para netter di Indonesia. Pihak yang menyebarkan berita hoax ini memiliki tujuan, salah satunya adalah untuk menggiring opini masyarakat dan kemudian membentuk persepsi yang salah terhadap suatu informasi yang sebenarnya [3]. Banyaknya pengguna aktif media sosial di Indonesia ini sangat memudahkan pihak penyebar hoax dalam menjalankan aksinya.

Masyarakat tidak mampu membedakan berita *hoax* dan *non hoax* karena rendahnya narasi publik terhadap informasi di media sosial [4]. Pengaruh perkembangan teknologi bisa menjadi ancaman global termasuk Indonesia, khususnya terkait dengan penyebaran berita bohong (*hoax*).

Data mining adalah suatu proses ekstraksi atau penggalian data yang belum diketahui sebelumnya, data tersebut digali informasinya berdasarkan database besar untuk diambil keputusan yang penting [5]. Data mining mempunyai fungsi yang penting untuk membantu mendapatkan informasi yang berguna serta meningkatkan pengetahuan bagi pengguna, salah satu fungsinya yaitu klasifikasi. Klasifikasi adalah kumpulan model yang dapat melakukan ilustrasi atau menggambarkan dan membedakan kelas data atau konsep, dengan tujuan mampu menggunakan model untuk melakukan prediksi kelas dari objek yang label kelasnya tidak diketahui. Model tersebut didasarkan pada pola analisis kumpulan data latih [6].

Sebelum postingan dilakukan klasifikasi maka diperlukan tahap *preprocessing*. *Preprocessing* mempunyai tahapan sebagai berikut, yaitu *case folding*,

tokenizing, stopwords removal dan *stemming*. Proses *stemming* menjadi tahapan paling penting di dalam tahap *preprocessing* dikarenakan pada *stemming* terjadi proses penghilangan kata imbuhan. Sehingga kata menjadi kata dasar [7]. Salah satu metode untuk klasifikasi yang dapat digunakan adalah *Naïve Bayes Classifier*. *Naïve Bayes Classifier* telah banyak digunakan dalam penelitian text mining, beberapa kelebihan *Naïve Bayes Classifier* diantaranya adalah algoritmanya yang sederhana tetapi memiliki tingkat akurasi yang tinggi [8].

Penelitian yang berkaitan dengan NBC telah dilakukan diantaranya adalah Klasifikasi Berita Indonesia yang dilakukan oleh Arifiyanti, menggunakan NBC dengan *confix-stripping stemmer* mendapatkan hasil ketepatan klasifikasi sebesar 86.74% [4].

Dari referensi yang ada, dapat ditarik kesimpulan bahwa metode *Naïve Bayes Classifier* merupakan salah satu metode yang memiliki akurasi tinggi yang dapat digunakan peneliti dalam melakukan analisis. Selain itu, adapun kelebihan lain yang dimiliki oleh Algoritme *Naïve Bayes Classifier* yaitu kecepatan yang tinggi, mudah dipahami, pengkodeannya sederhana, lebih cepat dalam penghitungan, menangani kuantitatif dan data diskrit, dan lain-lain [9].

Berdasarkan uraian di atas, maka penulis akan melakukan penelitian sebagai skripsi dengan judul **"Implementasi *Naïve Bayes Classifier* untuk Mendeteksi Postingan Hoax terhadap COVID-19 di Twitter"**.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dapat dirumuskan rumusan masalah sebagai berikut:

1. Bagaimana penerapan algoritame *Naïve Bayes Classifier* dalam mendeteksi berita *hoax* di Twitter terhadap COVID-19?
2. Bagaimana tingkat akurasi dari algoritame *Naïve Bayes Classifier*?
3. Variabel apa saja yang berpengaruh dalam mendeteksi berita hoax di Twitter terhadap COVID-19?

1.3 Batasan Masalah

Agar penelitian ini dapat terfokus maka diperlukan adanya batasan-batasan untuk membatasi ruang lingkup dari penelitian ini. Adapun batasan-batasan dari penelitian ini adalah sebagai berikut :

1. Menggunakan metode klasifikasi yaitu metode *Naïve Bayes Classifier* untuk mengklasifikasikan postingan yang bersifat *hoax* dan *non hoax*.
2. Objek yang diteliti merupakan postingan mengenai covid-19 di Indonesia.
3. Komentar yang dipilih merupakan komentar berbahasa Indonesia.
4. Pengolahan dan pelabelan data dilakukan secara manual berdasarkan studi literatur dari beberapa sumber.
5. Sistem yang dibangun berbasis *website*.

1.4 Maksud dan Tujuan Penelitian

Maksud dan tujuan dari penelitian ini adalah sebagai berikut :

1. Membangun sistem yang dapat mengklasifikasikan postingan terhadap COVID-19 yang merujuk ke arah *hoax* atau pun *non hoax* menggunakan *Naïve Bayes Classifier*.
2. Memberikan evaluasi dari hasil klasifikasi dari sistem yang diusulkan.

1.5 Metodologi Penelitian

Untuk mendukung penelitian ini, penulis menggunakan beberapa metodologi penelitian di antaranya adalah :

1.5.1 Metode Pengumpulan Data

Metode pengumpulan data merupakan metode yang diperlukan untuk memperoleh informasi yang dibutuhkan untuk menunjang penelitian. Adapun metode pengumpulan data yang akan digunakan dalam penelitian ini adalah metode studi pustaka dan *crawling* data.

1.5.1.1 Studi Pustaka

Studi pustaka dilakukan sebagai proses menghimpun informasi yang relevan terhadap topik atau permasalahan dari buku-buku, karya ilmiah, jurnal-jurnal, skripsi maupun situs internet yang dapat membantu penelitian. Dalam hal ini penulis mencari informasi mengenai *Naïve Bayes Classifier*, Twitter, *K-Folf Cross Validation*, teori *hoax*, covid-19, dan informasi lain yang relevan.

1.5.1.2 Crawling Data

Pengumpulan data dari penelitian ini yaitu data yang diunduh dari server twitter berupa user dan tweet beserta atribut-atributnya sesuai kata kunci yang akan

dicari. Proses ini dilakukan dengan menggunakan API Twitter dan bahasa pemrograman Python.

1.5.2 Metode Analisis

Metode analisis dalam penelitian ini menggunakan metode *System Development Life Cycle* (SDLC) dengan tahapan-tahapan sebagai berikut:

1. Perencanaan Sistem (*System Planning*)
2. Analisis Sistem (*System Analysis*)
3. Perancangan Sistem (*System Design*)
4. Implementasi Sistem (*System Implementation*)
5. Pemeliharaan Sistem (*System Maintance*).

1.5.3 Metode Perancangan Sistem

Pada penelitian ini dalam membangun model klasifikasi postingan *hoax* COVID-19, penulis akan menggunakan *flowchart* dan DFD (*Data Flow Diagram*). *Flowchart* bertujuan untuk mengidentifikasi permasalahan-permasalahan yang ada dan kebutuhan-kebutuhan yang diperlukan agar dapat dievaluasi dan diusulkan menjadi ke dalam model yang lebih baik. Sedangkan DFD akan menerangkan aliran data dari program penelitian ini.

1.5.4 Metode Implementasi

1. Pengambilan data pada Twitter dengan menggunakan Python, berikuttahapan-tahapannya:
    1. Langkah pertama yaitu buka jupiter notebook dan panggil librarytweepy

dan csv. Tweepy disini akan digunakan untuk prosesauthentication API twitter dan untuk melakukan crawling twitter. Sementara csv untuk membuat csv file.

2. Langkah selanjutnya yaitu masukkan consumer_key, consumer_secret, Access_token dan access_token_secret yang didapatkan pada saat melakukan pembuatan API twitter.
3. Melakukan prosess authentication API twitter.
4. Langkah selanjutnya yaitu memberi nama tempat penyimpanan data crawling, kemudian membuatnya menggunakan fungsi csv.writer.
5. Langkah terakhir yaitu melakukan crawling twitter dengan mencari tweet dengan Hashtag #covid-19 dengan jumlah yang dibutuhkan. Kemudian bisa menentukan Bahasa yang digunakan, menggunakan "id"(Indonesia). Selanjutnya menentukan mulai kapan ingin mengambil data pada twitter, kemudian pada baris terakhir yaitu memasukkan tweet kedalam csv.
6. Setelah itu bisa melihat hasilnya dengan membuka file csv.

2. *Preprocessing* Data

Preprocessing merupakan tahap awal dari *text mining* untuk mengubahdata sesuai dengan format yang dibutuhkan, berikut tahapan-tahapannya:

1. *Case Folding*, tahapan awal pada *preprocessing* yang bertujuan untuk membersihkan tweet meliputi: menghilangkan URL, menghilangkan *emoticon*, menghilangkan simbol, menghilangkan *@username*,

menghilangkan *hashtag*, proses penyeragaman bentuk huruf menjadi huruf kecil.

2. *Tokenization*, Langkah *tokenization* merupakan langkah membagi kalimat *tweet* menjadi kata-kata yang menyusun kalimat tersebut. Tahap ini digunakan untuk memudahkan analisis kata-kata tersebut lebih lanjut.
3. *Stopwords Removal* adalah tahap menghilangkan kata-kata yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Contoh *stopword* dalam bahasa Indonesia adalah "yang", "dan", "dari", "di", dll. Kata-kata *stopword* dihilangkan karena memiliki informasi rendah dari sebuah kalimat, sebagai gantinya dapatberfokus pada kata-kata penting yang lainnya. Pada kasus ini peneliti menggunakan *library* dari *Python Sastrawi* yang dimana *library* tersebut juga dapat digunakan untuk menghilangkan kata tambahan pada *preprocessing*.
4. *Stemming* merupakan proses untuk merubah token-token yang telah melewati proses sebelumnya ke dalam bentuk kata dasar. Dalam prosesnya, peneliti menggunakan *library Sasatrawi*

3. Klasifikasi *Naïve Bayes Classifier*

Klasifikasi yang dilakukan menggunakan metode *Naïve Bayes Classifier*. Metode *Naïve Bayes Classifier* memiliki 2 proses yaitu proses *training* (pelatihan) dan *classification* (klasifikasi). Pada tahap *training* akan dilakukan pencarian probabilitas data *training* pada masing-masing kelas yang telah

melewati proses sebelumnya yaitu pelabelan dan *preprocessing* data. Untuk hasil klasifikasinya akan ditentukan dari nilai probabilitas yang paling besar apakah tergolong pada kelas *hoax* atau *real*.

4. Pengujian

   Pengujian sistem dilakukan dengan menggunakan *Black box Testing* dan pengujian model menggunakan *K-Fold Cross Validation* dan *Confusion Matrix*.

1.5.5 Metode Pengujian

Pengujian yang dilakukan pada penelitian ini menggunakan *K-Fold Cross Validation* dan *Confusion Matrix* untuk mengetahui *accuracy error, precision dan recall* dari seluruh *fold*. Serta menggunakan pengujian *Black Box Testing* untuk menguji apakah program memenuhikebutuhan (*requirements*) yang disebutkan dalam spesifikasi atau tidak.

1.6 Sistematika Penulisan

Penelitian ini dituliskan dalam suatu sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini berisi mengenai latar belakang penelitian, rumusan penelitian, batasan penelitian, maksud dan tujuan penelitian, metode penelitian yangdigunakan serta sistematika penulisan penelitian.

BAB II LANDASAN TEORI

Pada bab ini berisi kajian pustaka yang berisi penelitian-penelitian lain yang relevan serta teori-teori yang dijadikan dasar dalam penelitian ini.

BAB III ANALISIS DAN PERANCANGAN

Pada bab ini memberikan penjelasan mengenai analisis kebutuhan dan perancangan sistem yang akan diusulkan dalam penelitian ini.

BAB IV IMPLEMENTASI DAN PEMBAHASAN

Pada bab ini berisi mengenai implementasi rancangan sistem yang akan dibuatserta akan dijelaskan pula mengenai pengujian dan evaluasi dari sistem yang akan dibuat.

BAB V PENUTUP

Pada bab ini berisi kesimpulan dan saran dari penulis mengenai penelitian yangtelah dilakukan.

DAFTAR PUSTAKA

Berisi sumber-sumber yang digunakan penulis dalam melakukan penelitian.