BAB I
PENDAHULUAN

1.1. Latar Belakang

Twitter (yang sekarang bernama "X" setelah akuisisi oleh Elon Musk, tetapi untuk memudahkan, nama *Twitter* akan digunakan dalam penelitian ini) merupakan salah satu jaringan media sosial utama yang digunakan di dunia (Arazzi, M., dkk., 2023). *Twitter* terestimasi memiliki 368 juta pengguna aktif pada awal Desember 2022 di seluruh dunia (Stacy Jo Dixon, 2023). Hal tersebut membuat *Twitter* memiliki jangkauan global. Dengan jangkauan yang global tersebut, salah satu industri yang memanfaatkan jangkauan global sosial media seperti *Twitter* adalah industri *e-commerce.* Dengan memanfaatkan platform sosial media seperti *Twitter*, sebuah bisnis *e-commerce* dapat melakukan marketing dan meningkatkan interaksi dengan konsumen seperti dengan memposting promosi, mengadakan *giveaway*, memberikan *customer service,* dan lain-lainnya. Menurut Singh dan Cullinane (dalam Permatasari, A., dan Kuswadi, E., 2017) dengan menggunakan media sosial, konsumen dapat langsung memberikan tanggapan terhadap "pendapat, komentar, dan saran mengenai produk yang mereka tawarkan, sehingga konsumen dapat memperoleh produk yang mereka inginkan dan butuhkan".

Walaupun dengan keunggulan dari sosial media tersebut, terdapat pula sebuah isu yang sering muncul dan dapat mengefek upaya marketing bisnis *e-commerce* dan lain-lainnya pada platform media sosial. Isu tersebut dikarenakan aktivitas bot sosial. Dikalkulasikan bahwa mayoritas pengisi aktivitas di internet lebih banyak diisi oleh aktivitas bot daripada manusia (Li, Xigao, dkk., 2021). Di *Twitter*, terestimasi sekitar 48 juta penggunanya adalah akun yang terotomasi (Varol, dkk., 2017).

Sebuah bot adalah semacam software yang melakukan aktivitas terotomasi di internet. Bot dapat digunakan untuk melayani berbagai tujuan "Baik" ataupun "buruk", dimana bot tersebut dapat digunakan oleh sebuah pengguna melalui proses otomasi. Dengan pengotomasian, bot dapat digunakan secara jumlah besar dan frekuensi yang besar, sehingga dapat digunakan untuk mempengaruhi persepsi pengguna dalam suatu hal atau isu. Menurut laporan industri terbaru, bot bertanggung jawab atas 37,2% dari total lalu lintas terkait situs web, dan bot

”buruk” bertanggung jawab atas 64,7% dari keseluruhan lalu lintas bot (Li, Xigao, dkk., 2021).

Dari sisi bisnis *e-commerce* dalam media sosial seperti *Twitter,* sebuah bot yang “buruk” dapat mengganggu upaya marketing sebuah bisnis *e-commerce* melalui berbagai cara. Menurut (Schnebly & Sengupta, 2019) bot *twitter* memiliki kemampuan untuk mengubah topik yang sedang tren dengan menghasilkan banyak tweet yang membuat cerita palsu tampak memiliki lebih banyak penayang dengan mempromosikan atau men-tweet tentang topik tertentu bersama dengan hashtag terkait. Dengan kemampuan tersebut Bot dapat menimbulkan bencana bagi kehadiran media sosial sebuah perusahaan. Selain itu, menurut Kind dkk., (dalam Godulla, A. dkk., 2021) Bot sosial berpotensi berbahaya lainnya. Misalnya, dampaknya mengefek distorsi statistik ketika mengevaluasi data di jejaring sosial hingga perang dunia maya dan kejahatan priyayi. Maka dari itu kemampuan untuk mendeteksi bot sosial dan juga mengklasifikasikan bot sosial dan pengguna asli merupakan kemampuan yang sangat penting, dan dibutuhkan suatu solusi untuk mengklasifikasikan apakah sebuah post/*tweet* merupakan post/*tweet* yang *legitimate* atau sebuah bot.

Beberapa penelitian telah dilakukan untuk mendeteksi *Twitter* bot dengan beberapa metode dan algoritma. Seperti pada penelitian yang dilakukan oleh Schnebly dan Sengupta (2020) yang mengklasifikasikan bot *twitter* dengan membuat sebuah kumpulan fitur dari dataset dan menggunakan Algoritma *Machine Learning Random Forest* yang menghasilkan nilai akurasi sebesar (90.25%). Selain itu dilakukan pula penelitian oleh Kudugunta dan Ferrara (2018) terhadap pendeteksian bot menggunakan *Deep Neural Network* berbasis arsitektur *contextual long short-term memory* (LSTM) yang menghasilkan akurasi klasifikasi yang tinggi (AUC>96%).

Penelitian ini mengusulkan tentang bagaimana mengklasifikasikan bot pada *twitter* berdasarkan perilaku pengguna. Perilaku pengguna beracuan pada beberapa atribut sebuah *post* seperti sentimen positif/negatif, jumlah URL, jumlah *hashtag,* jumlah *mention,* jumlah *like*, dan lain-lainnya. Pemilihan atribut tersebut berdasarkan penelitian oleh Alarfaj, Fawaz Khaled, dkk. (2023), dimana telah diekstraksi beberapa konten fitur untuk melakukan deteksi bot *Twitter.* Metode

yang digunakan untuk melakukan klasifikasi adalah metode *semi-supervised support vector machine.* Penggunaan metode ini didasarkan penelitian (Mbona, I., & Eloff, J. H., 2023), dimana metode *semi-supervised support vector machine* (S3VM) mencapai hasil terbaik dari model *semi-supervised machine learning* lainnya. Pada penelitian ini, dalam pemanfaatan data tak berlabel akan digunakan proses *pseudo-labelling* untuk meningkatkan kinerja model klasifikasi.

Pada pembangunan model klasifikasi di penelitian ini, dimanfaatkan metode pembelajaran mesin *semi-supervised learning.* Menurut Chappele dkk. dan Zhu (dalam Van Engelen, J. E., & Hoos, H. H., 2020.), *semi-supervised learning* adalah cabang pembelajaran mesin yang bertujuan untuk menggabungkan kedua metode, yaitu *supervised learning* dan *unsupervised learning.* Metode klasifikasi *semi-supervised* sangat relevan pada skenario dimana data berlabel langka. Maka dari itu, metode *semi-supervised* tersebut bertujuan untuk menggunakan dataset yang tidak berlabel dan juga dataset yang berlabel.

Pada penelitian ini, dataset tak berlabel yang digunakan merupakan dataset yang dikumpulkan dari *Twitter*, dimana data-data dalam dataset tersebut diambil dengan menggunakan teknik *Twitter Data Crawling. Twitter Data Crawling* adalah sebuah aksi dimana data *Tweet* user dikumpulkan dan diambil menggunakan *API* yang disediakan *Twitter* (Sohail dkk., 2021). Sedangkan untuk dataset berlabel yang akan digunakan berasal dari dataset *cresci-2017* (Cresci, S., dkk., 2017*)* dari *botometer* dimana dataset tersebut berisi data-data akun dan *tweet* pengguna asli dan bot sosial. Kedua dataset tersebut dapat digunakan untuk membantu melatih model klasifikasi dan digunakan untuk proses *pseudo-labeling*.

Berdasarkan penjelasan diatas, pada penelitian ini akan membahas pembuatan model klasifikasi *post social bot twitter/x* dengan *semi-supervised support vector machine.*

1.2. Rumusan Masalah

Berdasarkan latar belakang diatas rumusan masalah adalah sebagai berikut:

1) Bagaimana kinerja dari model klasifikasi dalam mengklasifikasikan bot sosial *Twitter* menggunakan *Semi-Supervised Learning* dan *Support Vector Machine*?

2) Bagaimana perbedaan kinerja dari model SVM yang dilatih dengan data berlabel dan kinerja model SVM setelah dilatih dengan data *pseudo-labeled*?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dari penelitian adalah sebagai berikut:

1) Mengetahui kinerja model klasifikasi dalam mengklasifikasikan bot sosial *Twitter* menggunakan *Semi-Supervised Learning* dan *Support Vector Machine.*
2) Mengevaluasi perbedaan kinerja dari model SVM yang dilatih dengan data berlabel dan kinerja model SVM setelah dilatih dengan data *pseudo-labeled*.

1.4. Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat menjadi acuan bagi peneliti-peneliti lainnya dalam melakukan penelitian dan juga bermanfaat untuk mengembangkan ilmu pembelajaran mesin, deteksi bot sosial, media sosial dan lain-lainnya.

1.5. Batasan Masalah

Berdasarkan yang telah dijelaskan diatas, fokus dan batasan masalah dari penelitian adalah sebagai berikut:

1) Alat yang digunakan untuk melakukan *Twitter Data Crawling* adalah *Tweet-Harvest*. Data yang dikumpulkan berupa *post (tweet)* berbahasa Inggris yang publik, dikumpulkan dari tanggal 4 Juli 2024, dan dengan kata kunci "#Amazon", "#eBay", dan "#AliExpress"
2) Atribut data *Twitter* yang dikumpulkan dengan *data crawling* adalah "conversation_id_str, created_at, favorite_count, full_text, id_str, image_url, in_reply_to_screen_name, lang, location, quote_count, reply_count, retweet_count, tweet_url, user_id_str, username"
3) Penggunaan pendekatan *Semi-Supervised Learning* menggunakan metode *Pseudo-labelling* untuk memanfaatkan dataset tak berlabel
4) Algoritma *Support Vector Machine* menggunakan parameter kernel *Radial Basis Function* (RBF) dan nilai *C*= 0.001, C= 0.01, C= 0.1, C= 1, C= 10, C= 100.

1.6. Sistematika Pelaporan Skripsi

Sistematika pelaporan skripsi ini terdiri dari 5 (lima) bab yaitu sebagai berikut:

BAB I PENDAHULUAN

Bab ini merupakan awal dari penelitian. Bab ini terdiri dari latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika pelaporan skripsi. Pada bab ini dijelaskan permasalahan utama dari penelitian yang dilakukan.

BAB II KAJIAN PUSTAKA

Bab ini berisi teori-teori yang relevan dan berkaitan dengan penelitian ini. Teori yang dijelaskan dijadikan sebuah landasan dalam penulisan penelitian ini.

BAB III METODOLOGI PENELITIAN

Bab ini berisi penjelasan mengenai metodologi penelitian yang akan digunakan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi penjabaran hasil dan pembahasan dari penelitian yang telah dilakukan.

BAB V SIMPULAN DAN SARAN

Bab ini berisi simpulan yang telah didapat dari penelitian ini, simpulan merupakan jawaban dari rumusan masalah yang ada pada Bab I. Selanjutnya dilampirkan saran dan berbagai hal yang dapat dilakukan untuk meningkatkan penelitian di kedepannya kepada penelitian selanjutnya, pembaca dan peneliti lainnya.