1. Pendahuluan

Latar Belakang

Twitter merupakan sebuah sosial media yang sangat mudah dijadikan sarana untuk berbagi informasi. Pada Twitter, setiap pengguna dibolehkan untuk membuat profil Twitter, untuk menulis pesan, maupun untuk berbagi informasi dengan pengguna lain [1]. Melalui *tweet,* informasi bisa disebarkan ke publik secara real-time. Setiap *tweet* juga bisa disisipkan berupa foto, video, maupun tautan website. Selain itu, Twitter juga mempunyai fitur *Retweet* yang berfungsi untuk memposting ulang *tweet* pengguna lain ataupun *tweet* milik sendiri untuk dibagikan ke *followers* pengguna. Hal inilah yang membuat *Retweet* menjadi kunci mekanisme dari sebuah difusi informasi mengapa informasi bisa menyebar [2].

Dari banyaknya *tweet* yang dibagikan di Twitter, ada *tweet* yang mendapatkan *retweet* dan ada juga yang tidak mendapatkan *retweet.* Hal tersebut bisa menjadi potensi untuk dijadikan sebuah penelitian untuk keperluan penyelesaian masalah. Contohnya bisa di berbagai hal seperti pada pemasaran yang membutuhkan *engagement* yang besar, analisis bisnis, hingga membantu dalam memutuskan sebuah prediksi. Sehingga permasalahan pada bidang tersebut bisa terbantu [2].

Maka dari itu diperlukan sebuah pemodelan yang bisa digunakan untuk memprediksi apakah sebuah *tweet* mendapatkan *retweet* atau tidak, yang mana hasilnya berupa skor akurasi pada prediksi *retweet*. Prediksi *retweet* memiliki banyak fitur yang bisa digunakan untuk membangun model prediksi retweet tersebut. Fitur yang bisa digunakan antara lain fitur *User-based*, *Time-based,* dan *Content-based*. Dimana *User-based* berupa informasi terkait profil pengguna. Lalu, *Time-based* berupa fitur yang berkaitan dengan waktu sebuah *tweet* tersebut dipublikasikan. Hasil hipotesa yang disebutkan oleh Thi Bich Ngoc Hoang dan Josiane Mothe, pada "waktu senggang", sebuah *tweet* lebih mudah mendapatkan *retweet*. Sedangkan *Content-based* berupa informasi yang berkaitan dengan isi pada sebuah *tweet* [3].

Pada penelitian yang sudah dilakukan oleh Ishita Daga, dkk. melakukan pencarian *accuracy* dari prediksi *Retweets* dan *Likes* pada Twitter dengan *Text Embedding* TF-IDF dan Doc2Vec dengan menggunakan beberapa metode klasifikasi untuk pembandingnya. Akurasi pada prediksi *Retweets* dengan metode *Logistic Regression* menghasilkan akurasi 70.1%, dengan metode *SVM* menghasilkan akurasi 62.9%, dengan metode *Random Forest* menghasilkan akurasi 49.8%, dengan metode *Neural Network* menghasilkan akurasi 39.1%, dan menggunakan metode *Multinomial Naïve Bayes* 75.9%[4]. Sehingga dapat disimpulkan bahwa Text Embedding dan jenis metode klasifikasi sangat mempengaruhi akurasi. Sehingga itu diperlukan pemodelan dengan metode lain untuk mendapatkan akurasi lebih baik.

Pada penelitian ini, penulis berfokus pada pemodelan untuk pencarian akurasi prediksi *Retweets* dengan menggunakan fitur User-based dan Content-based dengan menggunakan Metode *Ensemble Stacking*. *Ensemble Stacking* itu sendiri merupakan gabungan dari beberapa metode klasifikasi Machine Learning. Pada *base-learner* ini menggunakan 3 buah metode klasifikasi, yaitu *Random Forest*, *Gradient Boost,* dan *Support Vector Machine*(SVM). Sedangkan pada *meta-learner* digunakan metode klasifikasi *Support Vector Machine*(SVM). Sehingga diharapkan dengan menggunakan penumpukan metode klasifikasi(*Stacking*) bisa mendapatkan akurasi yang lebih baik dibandingkan hanya menggunakan satu metode klasifikasi.

Topik dan Batasannya

Berdasarkan latar belakang yang sudah disebutkan, maka dapat ditentukan topik masalahnya adalah prediksi retweet menggunakan fitur User-based dan Content-based menggunakan metode Ensemble Stacking yang terdiri dari 3 *base-learner* dan 1 *meta-learner.* Sedangkan untuk Batasannya adalah data yang digunakan adalah data tweet yang diambil dengan kata kunci "kominfo " dalan rentang waktu April hingga Juli 2022.

Tujuan

Penelitian ini mempunyai tujuan untuk membangun sistem yang dapat memprediksi *retweet* dengan fitur *user-based* dan *content-based* menggunakan metode Ensemble Stacking.

Organisasi Tulisan

Pada tulisan selanjutnya akan dijelaskan tentang studi terkait tentang penelitian dan hasil penelitian tersebut. Lalu, pada bagian selanjutnya menjelaskan tentang urutan sistem yang akan dibangun untuk melakukan penelitian. Pada bagian evaluasi menjelaskan tentang hasil dari penelitian dan analisis hasil penelitian dengan melakukan beberapa skenario yang berbeda. Pada bagian terakhir berisi kesimpulan dari penelitian dan saran untuk penelitian selanjutnya.