1. Pendahuluan

Dengan banyaknya informasi tersebar di Twitter melalui *tweet* pengguna, suatu *tweet* dapat digunakan untuk memberi suatu prediksi tentang apa yang populer dan mengapa suatu informasi dapat menjadi populer dengan cepat. Memodelkan difusi informasi di era digital ini sangat penting untuk mengetahui cara penyebaran informasi dan cara mengendalikannya. Dengan mempelajari difusi informasi kita dapat mengetahui bagaimana seseorang dapat terpengaruh oleh suatu pendapat dan bagaimana suatu informasi dapat menjadi populer secara cepat.

Penelitian ini akan menganalisa apakah suatu tweet akan di retweet (atau bisa disebut dengan tersebar), apakah level penyebaran suatu tweet dapat dimodelkan dan apakah kita dapat meramalkan penyebaran tweet yang baru. Kali ini saya menggunakan model Naïve Bayes untuk memodelkan penyebaran tweet dan memprediksi penyebaran tweet yang baru.

Dari penelitian lain didapatkan bahwa suatu tweet didapatkan hasil bahwa suatu tweet dapat diprediksi kepopulerannya dengan menggunakan model klasifikasi pembelajaran mesin[1]. Di dalam paper tersebut ada tiga kelompok fitur yang terdiri dari *User-based feature, Content-based feature, dan time-based feature. Content-based feature* terdiri dari fitur yang berhubungan langsung dengan konten dari suatu tweet. User-based feature berisi fitur yang ada di dalam suatu user. *Time-based feature* terdiri dari fitur yang terdiri tentang waktu dari suatu post seperti contoh kapan suatu tweet disebar, apakah di tweet di saat siang hari atau malam hari, dan sebagainya. Tugas akhir ini mengambil *User-based feature* sebagai fitur yang digunakan untuk di proses di model *naive-bayes.*

User-based features adalah fitur yang mengandung informasi interaksi antar pengguna. Fiturnya terdiri dari *total of tweets* yang berisi total *tweet* yang telah dibuat oleh pengguna, *number of followers* yang berisi jumlah *followers, following* yang berisi akun mana yang pengguna follow, *age of user* yang berisi umur pengguna, *number of likes* berisi jumlah tweet yang ditandai sebagai suka, dan interaksi antar *tweet* dan *retweet*. Fitur ini dipilih dengan harapan dapat mendapatkan kriteria tweet yang akan menjadi viral berdasarkan umur pengguna, *jumlah followers, jumlah following,* dan interaksi dengan pengguna.

Naïve Bayes adalah model pembelajaran mesin *supervised* yang akan memodelkan suatu data untuk diklasifikasi. *Naïve Bayes* akan mengaplikasikan teori bayes dengan asumsi "naif" dari kemandirian bersyarat dari setiap pasangan fitur dengan nilai yang sudah ada. Di lapangan, *Naïve Bayes* diaplikasikan di berbagai situasi nyata seperti pendeteksi spam, klasifikasi dokumen, dll. *Naïve Bayes* juga membutuhkan data latih yang sedikit untuk untuk memberi hasil prediksi model. *Naïve Bayes* bisa menjadi model yang cepat jika dibandingkan dengan model yang lebih rumit. Pemisahan dari kelas distribusi fitur bersyarat adalah setiap distribusi dapat diprediksi satu-satu.[10]

1.1 Latar Belakang

Penelitian ini mempunyai latar belakang untuk mengetahui apakah suatu tweet dapat diprediksi akan menjadi populer dengan *user-based feature*. Dengan ini juga dapat diketahui apakah umur akun pengguna, banyaknya teman pengguna, dan banyaknya likes akan berpengaruh besar terhadap kemungkinan tweet yang akan di post pengguna akan menjadi populer dan tersebar. Penelitian ini menggunakan algoritma *Naive Bayes* karena algoritma ini cukup sederhana dalam pengaplikasiannya dan dengan algoritma ini akan membuktikan apakah dengan algoritma yang tepat akan menghasilkan hasil prediksi yang akurat.

1.2 Topik dan Batasannya

Sesuai dengan judul diatas, peneliti akan melakukan prediksi tweet yang viral dengan model *Naive Bayes.* Data mulai dikumpulkan selama seminggu sebanyak 3 kali dengan topik BLM (*Black Lives Matter*). Total lama pengumpulan data selama tiga minggu. Data diambil menggunakan website bernama Netlytics yang dapat mengumpulkan data tweet maksimal 2500 tweet untuk versi gratis. Karena keterbatasan sumber daya komputer maka total 7500 data yang terkumpul selama tiga minggu lebih dari cukup untuk dimasukan ke dalam model.

Jika dibandingkan dengan penelitian lain dataset yang mereka punya jauh lebih banyak sebanyak lebih dari enam juta tweet dibandingkan dengan penelitian ini sebanyak 7500 tweet dan mereka juga menyewa komputer dengan kemampuan komputasi dan sumber daya RAM yang lebih tinggi dibanding komputer pada umumnya[1]. Penelitian tersebut menggunakan platform Osirim-IRIT dengan RAM 64gb sementara penelitian ini diproses menggunakan Prosesor Intel Core i5 3230m dan RAM 8Gb[1].

1.3 Tujuan

Tujuan utama dari penelitian ini adalah untuk memprediksi apakah suatu tweet akan menjadi populer dengan user based feature. Dataset tersebut akan diuji menggunakan model Naive Bayes untuk mengetahui apakah model ini cocok untuk diaplikasikan ke dataset ini. hasil dari model tersebut akan dianalisa Setelah itu dilakukan proses evaluasi untuk menganalisa nilai F1-skornya untuk mencari tahu apakah model ini cocok untuk memprediksi *tweet* yang populer.

1.4 Organisasi Tulisan

Setelah ini akan dilanjutkan dengan empat bab selanjutnya yang terdiri dari: Studi terkait, Sistem yang dibangun, Evaluasi, dan Kesimpulan.