Optimalisasi *Hyperparameter* pada *Logistic Regression* Menggunakan *Grid Search* untuk Mendeteksi Rumor pada Twitter

Vrieza Rizqya Fajrul Rahman[1], Erwin Budi Setiawan[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]vriezarhmn@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

Abstrak

Twitter merupakan media sosial populer yang digunakan oleh masyarakat di seluruh dunia dan menjadi media sosial pertama dan tercepat dalam hal penyebaran berita. Pada tahun 2019, jumlah pengguna Twitter mencapai 134 juta pengguna aktif setiap harinya. Berita yang tersebar dengan cepat, tanpa ada supervisi, meningkatkan jumlah penyebaran berita atau isu yang belum tentu benar adanya, yang disebut dengan rumor. Penyebaran rumor dengan cepat dapat menyebabkan opini publik menjadi salah arah. Oleh karena itu, dibangun sistem untuk melakukan klasifikasi terhadap tweet ke dalam dua kelas, yaitu rumor dan non rumor. Metode yang digunakan pada penelitian ini untuk melakukan klasifikasi yaitu *Logistic Regression*, sebuah metode klasifikasi yang cukup populer untuk permasalahan biner. Sebelum dilakukan klasifikasi, tweet yang diambil harus melewati *pre-processing*. *Feature Extraction* yang digunakan dalam metode ini adalah *TF-IDF (Term Frequency – Inverse Document Frequency)*. Untuk meningkatkan performa *Logistic Regression*, digunakan juga *Grid Search* untuk mencari *Hyperparameter* terbaik. Dari hasil pengujian yang dilakukan, data uji sebesar 10% dengan kombinasi Fitur *TFIDF Unigram, Bigram, Trigram* dan Fitur Twitter setelah optimalisasi Hyperparameter menghasilkan nilai akurasi tes sebesar 72,03% dan akurasi train sebesar 77,60%.

Kata kunci : Twitter, Rumor, TF-IDF, Logistic Regression, Hyperparameter

Abstract

Twitter is a popular social media used by people around the world and is the first and fastest social media in terms of spreading news. In 2019, the number of Twitter users reached 134 million daily active users. News that spreads quickly, without any supervision, increases the number of news spreads or issues that are not necessarily true, which are called rumors. The rapid spread of rumors can lead to misguided public opinion. Therefore, a system was built to classify tweets into two classes, namely rumors and non-rumors. The method used in this study to classify is Logistic Regression, a popular classification method for binary problems. Before classification, the tweet that is taken must go through pre-processing. The Feature Extraction used in this method is TF-IDF (Term Frequency - Inverse Document Frequency). To improve Logistic Regression performance, Grid Search is also used to find the best Hyperparameter. From the results of the tests carried out, the test data is 10% with the combination of the TFIDF Unigram feature and the Twitter feature after optimizing the Hyperparameter resulting in a test accuracy value of 72,03% and a train accuracy of 77,60%.

Keywords: Twitter, Rumour, TF-IDF, Logistic Regression, Hyperparameter