ABSTRAK

Pemilihan calon presiden dilaksanakan setiap 5 tahun dengan berbagai kandidat yang mencalonkan diri, terutama dalam media sosial Twitter lebih sering ternjadi argumen seputar hal-hal politik yang tentunya banyak pengguna Twitter turut ikut berdiskusi tentang pemilihan calon presiden ini. Pengguna Twitter akan melakukan *tweet* untuk menyampaikan argumentasi dan diskusi terkait dengan pemilihan calon presiden ini. Oleh karena itu, penelitian ini berfokus pada *sentiment analysis* untuk melakukan penyimpulan respon pengguna terhadap pemilihan calon presiden serta melakukan validasi dengan mencari korelasi antara hasil survei elektabilitas dan data *sentiment* Twitter dengan menggunakan Korelasi Pearson.

Dalam pembangun mesin *sentiment*, metode *10-Fold Cross Validation* digunakan untuk mencari model mesin terbaik dari suatu dataset dengan pembagian data training dan data test sebesar 90:10. Lalu data alfabet akan diubah menjadi bentuk numerik dengan menggunakan metode pembobotan TF-IDF. Selanjutnya, untuk melakukan validasi dari model terbaik menggunakan *Confusion Matrix* untuk mendapat *f1-score*. Algoritma yang digunakan untuk membuat model adalah algoritma *Support vector machine* dengan kernel Gaussian RBF (Radial Basis Function). Hasil analisa dibandingkan dengan hasil survey elektabilitas portal berita yang memuat 3 calon tersebut dengan menggunakan Korelasi Pearson.

Berdasarkan hasil pencarian fold terbaik, ditemukan fold terbaik untuk masing-masing calon presiden yaitu fold ke-8 dengan *f1-score* 0,66 untuk calon Anies Baswedan dengan total 2.554 data training dan 283 data testing, fold ke-5 dengan *f1-score* 0,72 untuk calon Ganjar Pranowo dengan total 3.330 data training dan 370 data testing, dan fold ke-4 dengan *f1-score* 0,78 untuk calon Prabowo Subianto dengan total 3487 data training dan 387 data testing. Selanjutnya pada Korelasi Pearson, ditemukan koefisisen untuk masing-masing calon presiden yaitu Anies Baswedan dengan koefisien *sentiment* positif sebesar 0,994 dan koefisien *sentiment* negatif sebesar -0,994. Selanjutnya untuk calon presiden Ganjar Pranowo dengan koefisien *sentiment* positif sebesar -0,836 dan koefisien *sentiment* negatif sebesar 0,836. Lalu untuk calon presiden Prabowo Subianto dengan koefisiesn *sentiment* positif sebesar 0,789 dan koefisien *sentiment* negatif sebesar -0,789.

Penelitian ini menghasilkan fold terbaik untuk tiap data pada masing-masing calon presiden dengan ukuran *f1-score* untuk mencari model terbaik dari tiap fold. Pada Korelasi Peaason, pada calon Anies Baswedan dan Prabowo subianto, semakin tinggi *sentiment* positif, maka semakin tinggi juga data survei elektabilitas, sedangkan untuk calon Ganjar Pranowo, semakin rendah *sentiment* positif, maka semakin tinggi data survei elektabilitas. Untuk penelitian selanjutnya, dapat dilakukan penelitian yang membahas *hyper tuning* parameter dan menggunakan kernel lain pada algoritma *Support vector machine*.

Kata Kunci
NLP, *Pearson Correlation*, *Sentiment analysis*, SVM, TF-IDF

ABSTRACT

Elections for presidential candidates are held every 5 years with various candidates, especially on Twitter*, arguments about political matters often occur that many* Twitter *users participate in discussions about the election for presidential candidate.* Twitter *users will tweet to convey arguments and discussions related to the election. Therefore, this study focuses on sentiment analysis to infer user responses to the presidential election and validate it by looking for a correlation between electability survey results and* Twitter *sentiment data using Pearson Correlation.*

In sentiment analysis model, the 10-Fold Cross Validation method is used to find the best model from a dataset with a division of training data and test data with 90:10 split. Then the alphabetic data will be converted into numeric data using the TF-IDF weighting method. To validate the best model, Confusion Matrix is used to get the best f1-score. The model is using Support vector machine algorithm with the Gaussian RBF (Radial Basis Function) kernel. The results of the analysis are compared with the results of the news portal electability survey which contains the 3 candidates using Pearson Correlation.

Based on the search results for the best fold, the best fold was found for each presidential candidate, namely the 8th fold with an f1-score of 0.66 for candidate Anies Baswedan with a total of 2,554 training data and 283 testing data, the 5th fold with an f1-score of 0 .72 for the Ganjar Pranowo candidate with a total of 3,330 training data and 370 testing data, and the 4th fold with an f1-score of 0.78 for the Prabowo Subianto candidate with a total of 3,487 training data and 387 testing data. Furthermore, in the Pearson Correlation, a coefficient was found for each presidential candidate, namely Anies Baswedan with a positive sentiment coefficient of 0,994 and a negative sentiment coefficient of -0,994. Furthermore, for the presidential candidate Ganjar Pranowo with a positive sentiment coefficient of -0,836 and a negative sentiment coefficient of 0,836. Then for presidential candidate Prabowo Subianto with a positive sentiment coefficient of 0,789 and a negative sentiment coefficient of -0,789.

This study produces the best fold for each data on each presidential candidate with the f1-score to find the best model for each fold. In the Peason Correlation result, for candidate Anies Baswedan and Prabowo Subianto, the higher positive sentiment, the higher electability survey data, while in candidate Ganjar Pranowo, the lower positive sentiment, the higher electability survey data. For further research, can be discussed about hyper tuning parameters and using other kernels on Support vector machine algorithm.

Keywords
NLP, Pearson Correlation, Sentiment analysis, SVM, TF-ID