Identifikasi *Fake Account* Twitter Menggunakan *Support Vector Machine*

Akbar Habib Buana Wibawa Putra[1], Anisa Herdiani2[2], Ibnu Asror[3]
[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]habibakb@students.telkomuniversity.ac.id, [2] anisaherdiani@telkomuniversity.ac.id,
[3]iasror@telkomuniversity.ac.id

Abstrak

Penggunaan media sosial sebagai sarana untuk melakukan kampanye sudah menjadi hal yang biasa sejak kemunculan media sosial seperti *twitter, facebook,* dan *instagram*. Masyarakat yang menggunakan media sosial harus mendaftar dan memiliki akun media sosial tersebut, namun masih banyak masyarakat yang menggunakan akun palsu. Akun palsu ini dapat disalah gunakan seperti penggunaan akun palsu untuk membantu kampanye yang dapat menimbulkan bias politik terhadap pengguna media sosial lainnya, ataupun penggunaan akun palsu untuk meningkatkan popularitas dari seorang pemilik akun media sosial. Oleh karena itu, diperlukan suatu langkah penilitian pada tugas akhir ini untuk mendeteksi akun-akun palsu tersebut. Penelitian ini menggunakan *Support Vector Machine* (SVM) untuk mengidentifikasi akun palsu. Pengidentifikasian diawali dengan ekstraksi data kemudian labeling data hingga proses klasifikasi menggunakan SVM. Hasil dari identifikasi fake account ini menghasilkan nilai akurasi sebesar 93.42% dan f1-score sebesar 92.92%

Kata kunci : Twitter, SVM, Follower Palsu, Akun Palsu

Abstract

The use of social media as a means to carry out campaigns has become commonplace since the emergence of social media such as Twitter, Facebook and Instagram. People who use social media must register and have a social media account, but there are still many people who use fake accounts. These fake accounts can be misused such as the use of fake accounts to help campaigns that can cause political bias against other social media users, or using fake accounts to increase popularity of someone social media account. Therefore, we need a research step in this thesis to detect these fake accounts. The research use Support Vector Machine (SVM) to identify fake accounts. Identification begins with data extraction then labeling the data to the classification process using SVM. The results of this fake account identification produce an accuracy value of 93.42% and f1-score of 92.92%

Keyword : Twitter, SVM, Fake Followers, Fake Accounts

1. Pendahuluan

Latar Belakang

Kementerian Komunikasi dan Informatika yang biasa disebut KemenKominfo menyatakan bahwa 95% dari seluruh pengguna internet di Indonesia menggunakan internet untuk mengakses media sosial. Twitter merupakan salah satu media sosial yang sering diakses oleh pengguna internet di Indonesia, dengan menempati peringkat kelima di dunia yang bersaing dengan USA, Brazil, Jepang dan Inggris.

Twitter sebagai media sosial yang memiliki dampak besar dalam beberapa bidang industri, bisnis serta politik, dimana banyak politikus yang mulai merambah ke jejaring sosial untuk menjangkau pemilih dan pendukung yang lebih banyak lagi dibandingkan dengan cara konvensional. Dengan adanya Twitter komunikasi antar politikus, pemilih dan pendukung akan lebih mudah namun masih banyak pengguna yang menggunakan akun palsu untuk memberikan pengaruh *negative* yang dapat mempengaruhi pandangan politik terhadap pengguna twitter lainnya [1]. Selain itu, akun palsu juga digunakan untuk menambah *followers*, *like* dan *retweet* dari akun seseorang sehingga dapat meningkatkan popularitas dan kredibilitas akun tersebut. Penggunaan akun palsu seperti ini biasanya didapatkan dari penyedia jasa media sosial yang menyediakan jasa untuk membeli *followers*, membeli *like*, ataupun *retweet* [2].

Topik dan Batasannya

Seiring dengan adanya masalah tersebut diperlukan sebuah sistem untuk mendeteksi akun-akun palsu. Sistem ini diperlukan karena banyak akun palsu yang digunakan untuk menambah popularitas seseorang dengan menggunakan *follower* palsu atau *fake followers*. Perancangan sistem ini akan dibuat menggunakan beberapa tahapan diantarnya ekstraksi data kemudian labelling dan kemudian dilakukan proses klasifikasi menggunakan SVM. Pemilihan metode SVM ini dikarenakan pada jurnal penilitian "*Fake Account Detection in Twitter Based on Minimum Weighted Feature Set*" menyatakan bahwa metode SVM memiliki tingkat akurasi yang tinggi, yaitu 86.62% dalam melakukan pendeteksian terhadap akun palsu [3]. Sistem ini juga diharapkan dapat memberikan informasi klasifikasi terkait ciri-ciri akun palsu, seperti akun yang tidak *verified*, melihat jumlah *followers* dan *following* akun tersebut, serta melihat jumlah tweet yang dilakukan selama 2 minggu terakhir, sehingga

memperoleh hasil performansi yang tinggi dalam klasifikasi pendeteksian akun palsu untuk mengurangi penggunaan akun palsu pada *twitter*. Akun palsu yang akan diidentifikasi adalah akun yang digunakan sebagai penambah popularitas dari suatu akun *twitter* dengan menambah *followers* pada akun tersebut yang didapat dari akun penyedia jasa penambah *followers*.

Tujuan

Pada penelitian ini dilakukan proses pendeteksian akun palsu pada media sosial *twitter* menggunakan algoritma SVM. Kemudian sistem mengukur nilai akurasi pada perhitungan SVM dalam mendeteksi akun palsu dengan cara melakukan penyaringan atribut kembali yang kemudian dipilih sebagai atribut yang akan digunakan untuk klasifikasi serta menerapkan proses evaluasi dengan perhitungan *precision*, *recall* dan *f1 – score*.

2. Studi Terkait

2.1. Akun Palsu

Identitas merupakan hal-hal yang melekat pada manusia, namun terpisah dari tubuh manusia salah satu contoh identitas manusia seperti nama seseorang. Selain itu, contoh lainnya seperti KTP atau paspor yang mengandung nama, tempat dan tanggal lahir, kewarganegaraan, sidik jari digital, dan foto dari orang tersebut serta identitas harus unik dimana setiap objek pengenal hanya mangacu pada satu orang saja [4]. Keaslian identitas seseorang diperiksa dan diverifikasi oleh lembaga pemerintah seperti kelurahan dan kecamatan. Salah satu contohnya adalah KTP, kelurahan memastikan bahwa nama, sidik jadi, foto, tempat dan tanggal lahir, dan sebagainya benar benar dimiliki oleh orang yang sama. Pada situs media sosial seorang pengguna biasanya diidentifikasi oleh suatu profil yang dapat mengandung nama, foto, alamat dan tanggal lahir. Namun, yang mengatur dan membuat profil tersebut bisa saja bukan orang yang memiliki identitas yang dicantumkan pada profilnya. Hal ini disebut identitas palsu dimana seseorang menggunakan identitas orang lain. Seseorang juga bisa membuat profil yang menggunakan nama yang dibuat-buat dan informasi lain yang tidak berhubungan dengan siapapun di negara manapun, ini disebut identitas yang dipalsukan [4].

Situs media sosial seperti Twitter dan Facebook memiliki beberapa peraturan terkait akun atau profil palsu. Seperti pada Twitter akun yang dianggap palsu memiliki ciri seperti:

1. Menggunakan foto stok atau foto *avatar* curian
2. Menggunakan *bio* profil curian atau hasil salinan
3. Menggunakan informasi profil yang sengaja menyesatkan, termasuk lokasi profil

2.2. SVM (Support Vector Machine)

Support Vector Machine (SVM) merupakan salah satu metode yang digunakan untuk klasifikasi dan regresi .Pada *Support Vector Machine* akan mengklasifikasikan informasi dengan cara menemukan *hyperplane* yang memisahkan semua informasi dari satu jenis dan jenis klasifikasi lainnya. Hyperplane terbaik untuk metode SVM adalah yang memiliki jarak terbesar antar kelas. Sebuah SVM mengklasifikasikan data dengan menemukan hyperplane yang memisahkan semua aspek pengetahuan dari satu kategori dari kelas-kelas lain [1]. SVM menggunakan model linear sebagai *decision boundary* dengan bentuk umum sebagai berikut:

$$y(x) = w^T\theta(x) + b \quad (1)$$

Dimana **x** adalah vektor input, **w** adalah parameter bobot (*weight*), **θ(x)** adalah fungsi basis, dan *b* adalah suatu bias. Bentuk model linear yang paling sederhana untuk *decision boundary* adalah:

$$\mathrm{y(x)} = \mathrm{w^T x} + \mathrm{w_0} \quad (2)$$

Dimana **x** adalah vektor input, **w** adalah vektor bobot dan **w0** adalah bias. Sehingga *decision boundary* yang didapat adalah **y(x)** = 0 yaitu suatu *hyperplane* berdimensi (D-1). *Decision bounday* ini akan mengklasifikasikan suatu vektor input **x** ke kelas sampel positif (+1) jika **y(x) ≥ 0**, dan kelas sampel negatif (-1) jika **y(x) < 0**. Untuk menentukan *decision boundary* (DB), yaitu suatu model linear atau *hyperplane* **y(x)** dengan parameter **w** dan **b**, SVM menggunakan konsep margin yang didefiniskan sebagai jarak terdekat antara DB dengan sembarang data training [5].

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara DB dan titik terdekatnya, yaitu $\frac{1}{\|w\|}$. Hal ini dapat dirumuskan sebagai *Quadratic Programing problem*, yaitu mencari titik minimal persamaan (3), dan harus memenuhi batasan pada persamaan (4)

$$\min_w \tau(w) = \frac{1}{2}\|w\|^2 \quad (3)$$

$$y_i(x_i.w + b) - 1 \geq 0, \forall i \quad (4)$$

Permasalahan ini dapat diselesaikan dengan berbagai teknik komputasi, salah satunya adalah *Lagrange Multiplier*.

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i \left(y_i\left((x_i.w+b)-1\right)\right) \quad (i = 1,2,\dots,l) \tag{5}$$

Pada persamaan (5) α_i adalah pengali Lagrange (*Lagrange Multiplier*), yang bernilai nol atau positif ($\alpha_i \geq 0$).Nilai optimal dari persamaan (5) dapat dihitung dengan meminimalkan L terhadap w dan b, dan memaksimalkan L terhadap α_i. Dengan memperhatikan sifat bahwa pada titik optimal gradient L=0, persamaan (5) dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung α_i saja, seperti pada persamaan (6).

Maksimalkan:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j x_i.x_j \tag{6}$$

Bergantung pada:

$$\alpha_i \geq 0 \ \ (i = 1,2,\dots,l) \ \ \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{7}$$

Dari hasil perhitungan ini diperoleh α_i yang kebanyakan bernilai positif. Data yang berkorelasi dengan α_i yang positif inilah yang disebut sebagai support vector [6].

2.3. Evaluation

Precision dan *recall* merupakan salah satu teknik yang dapat digunakan dalam menghitung performansi dari sistem. Pada perhitungan ini membutuhkan beberapa komponen diantaranya:

1. *True Positive* (TP)
   Dokumen relevan yang diambil/teridentifikasi dengan benar sebagai dokumen relevan.
2. *True Negative* (TN)
   Dokumen tidak relevan yang tidak diambil/tidak teridentifikasi sebagai dokumen relevan.
3. *False Positive* (FP)
   Dokumen tidak relevan namun diambil/teridentifikasi sebagai dokumen relevan.
4. *False Negative* (FN)
   Dokumen relevan namun tidak diambil/tidak teridentifikasi sebagai dokumen relevan.

Tabel 1 Table Confusion Matrix [7]

	Retrieved	*Not retrieved*
Relevant	*True Positive (TP)*	*False Negative (FN)*
Not relevant	*False Positive (FP)*	*True Negative (TN)*

Performansi yang akan diuji diantaranya:

1. Precision

Precission yaitu sebuah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Rumus pernyataan *precision* sebagai berikut.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} * 100\% \tag{8}$$

2. Recall

Recall merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Rumus pernyataan *recall* sebagai berikut.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} * 100\% \tag{9}$$

3. F1-score

F1-score merupakan hasil rata-rata antara precision dan recall. F1-score ini muncul untuk menyetarakan nilai precision dan recall yang sering terpaut jauh. F1-score juga diartikan sebagai penyetaraan nilai precision dan recall [7].

$$\text{F1} - \text{Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} * 100\% \tag{10}$$

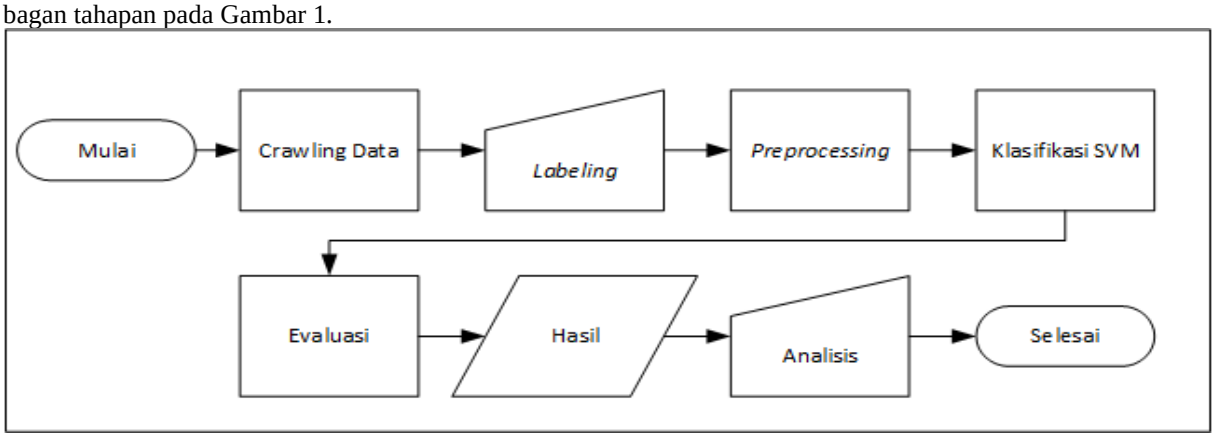
4. Accuracy

Accuracy adalah kedekatan antara nilai prediksi dengan nilai aktual, akurasi memberikan bobot yang sama untuk kesalahan label dari kedua jenis.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (11)$$

3. Sistem yang Dibangun

Metodologi penelitian yang digunakan untuk membangun sistem dalam penelitian ini digambarkan dengan bagan tahapan pada Gambar 1.



Gambar 1 Alur penelitian

Berikut merupakan penjelasan tahapan-tahapan dari alur pada penelitian ini.

1. Crawling

   Pengumpulan data yang akan dilakukan menggunakan *library Tweepy* pada *Python*. Data yang diambil diantaranya adalah data akun dan data tweet selama 3 bulan terakhir yang berjumlah 790 akun dengan berdasarkan beberapa atribut sebagai berikut:

   a. Jumlah followers
   b. Jumlah following
   c. Jumlah likes 500 tweet terakhir
   d. Jumlah hastag yang digunakan pada 500 tweet terakhir
   e. Jumlah total tweet
   f. Jumlah mention
   g. Jumlah tweet dalam 2 minggu terakhir maksimal 0-500 tweet
   h. Status *verified*
   i. Pemakaian *device* Android, web dan iPhone
   j. Status geolokasi
   k. Status akun (*protected* atau *public*)

   Atribut yang dipilih seluruhnya berdasarkan kepada penelitian sebelumnya pada [3] yang menggunakan atribut-atribut tersebut dalam megidentifikasi akun palsu. Beberapa atribut dipilih agar data yang didapat bisa relevan dengan data yang ada saat ini. Contohnya jumlah followers dan jumlah following. Pada penelitian sebelumnya [3] kedua atribut ini diberi batasan jumlah minimal untuk kemudian dijadikan nilai benar salah atau *boolean*. Pada penelitian ini jumlah followers dan following tidak dibatasi karena tidak terdapat batas minimal yang jelas untuk jumlah followers ataupun following agar dapat ditentukan akun tersebut asli ataupun palsu karena ada beberapa akun asli yang memiliki followers sedikit namun memiliki following banyak yang biasanya merupakan ciri-ciri dari akun palsu.

   Atribut jumlah likes, hashtag, total tweet, mention dan jumlah tweet 2 minggu terakhir dipilih dengan tujuan untuk menentukan keaktifan dari suatu akun. Atribut status *verified* dipilih karena untuk mendapatkan status tersebut pemilik akun perlu untuk secara personal datang atau meminta ke Twitter untuk dijadikan akun *verified* yang berarti akun *verified* dapat dipastikan sebagai akun yang *real*. Namun tidak semua akun yang belum *verified* merupakan akun *fake*. Atribut pemakaian *device* juga didapat dari penelitian sebelumnya [3], namun jika pada penelitian sebelumnya dibatasi hanya *device* iPhone pada penelitian ini pemakaian *device* Android dan web juga dipilih karena rasio pemakaian *device* Android dan web pada pemilik akun *real* lebih tinggi jika dibandingkan dengan pemakaian *device* iPhone. Contoh dataset dilampirkan pada lampiran 1.

2. Labeling

   Proses *labeling* ini dilakukan dengan melakukan pelabelan pada dataset, dataset yang diberikan label terdiri dari dataset gabungan antara data fake dan data asli yang akan menjadi satu dataset. Proses *labeling* diambil akun asli sebanyak 339 yang terdiri dari beberapa akun pengguna yang dikenal dan sisanya 451

akun palsu yang berasal dari penyedia jasa akun palsu. Berdasarkan ekstraksi data yang menjadi parameter kemudian akun yang palsu didapatkan keputusan dengan memberi label "fake" dan akun asli diberi label "real". Proses *labeling* dilakukan secara manual, proses pelabelan 339 akun asli dengan cara melihat kesesuaian kriteria akun asli, seperti *username* yang dikenal secara pribadi dan akun yang *verified* otomatis termasuk akun asli. Sedangkan, pelabelan akun fake dilabelkan berdasarkan akun yang didapatkan dari penyedia jasa akun palsu.

3. Preprocessing

   Setelah dataset diberi label pada proses labeling, proses selanjutnya ialah preporcessing data dimana data-data yang bersifat non-numerik dirubah menjadi data numerik. Untuk data *boolean* (benar/salah), data yang bernilai benar (TRUE) diubah menjadi 1 dan data yang bernilai salah (FALSE) diubah menjadi 0, seperti yang terlampir pada lampiran 2.

   Setelah mengubah data non-numerik menjadi data numerik langkah selanjutnya adalah melakukan *scaling* pada masing-masing atribut dengan menggunakan *MinMaxScaling*, yang bertujuan untuk memastikan suatu fitur tidak lebih mempengaruhi hasil prediksi daripada fitur lainnya.

4. Klasifikasi

   Klasifikasi merupakan proses pendeteksian akun palsu yang akan diidentifikasi dengan menggunakan metode SVM. Sebelum masuk ketahapan klasifikasi dataset terlebih dahulu dilakukan *scalling* untuk memastikan 1 fitur tidak mempengaruhi hasil prediksi dari satu fitur yang lain. Pada proses klasifikasi dengan metode SVM, digunakan *library* sklearn.svm pada bahasa pemrograman python. Terdapat beberapa parameter pada fungsi SVC pada *library* sklearn.svm, diantaranya adalah:

   a. Parameter C, parameter ini berfungsi untuk menentukan seberapa banyak kesalahan klasifikasi yang ingin dihindari dari tiap contoh *training*. Nilai C yang besar akan menghasilkan *hyperplane* dengan margin yang lebih kecil jika *hyperplane* tersebut bisa mendapatkan semua poin *training* yang diklasifikasikan dengan benar. Sebaliknya nilai C yang kecil akan menghasilkan *hyperplane* dengan margin yang besar meskipun *hyperplane* tersebut mendapatkan beberapa kesalahan klasifikasi. Parameter C yang digunakan pada penelitian ini adalah parameter C *default* atau C=1.
   b. Parameter kernel, parameter ini berfungsi untuk menentukan kernel apa yang akan digunakan oleh algoritma SVC. Nilai yang dapat dipilih antara lain adalah *linear*, *poly*, *rbf*, *sigmoid*, *precomputed*, atau sebuah fungsi. Parameter kernel yang digunakan pada penelitian ini adalah parameter kernel *rbf*, dimana kernel ini dapat memetakan inputan dengan nilai dimensi tak hingga.
   c. Parameter degree, parameter yang digunakan oleh kernel *poly* atau polinomial untuk menentukan derajat atau pangkat dari fungsi kernel polinomial. Secara *default* bernilai 3 untuk fungsi polinomial pangkat 3.
   d. Parameter gamma, parameter ini merupakan koefisien untuk fungsi kernel *rbf*, *poly*, dan *sigmoid*, yang berfungsi untuk mengubah hasil fungsi kernel dengan tujuan untuk mendapatkan *hyperplane* yang memiliki margin paling besar. Parameter gamma yang digunakan pada penelitian ini adalah *auto* yang merupakan nilai *default* dari parameter gamma yang menggunakan rumus 1 / jumlah_fitur.
   e. Paramter coef0, merupakan parameter koefisien yang hanya digunakan dalam kernel *poly*, dan *sigmoid*.
   f. Parameter shringking, adalah parameter bernilai benar atau salah yang akan menentukan apakah algoritma akan menggunakan penyusutan heuristik atau tidak.
   g. Parameter class_weight, dimana parameter ini akan menentukan bobot dari masing-masing kelas. Nilai yang digunakan adalah nilai *balanced* yang berfungsi secara otomatis menyeimbangkan bobot dari masing-masing kelas berbanding terbalik dengan frekuensi kelas pada data input.
   h. Parameter max_iter, yang berfungsi untuk menentukan batas jumlah iterasi maksimal. Penelitian ini menggunakan nilai *default* (-1) dari parameter max_iter yang berarti tidak akan ada batas iterasi dari algoritma.

4. Pengujian dan analisis

1. Pengujian

   Berikut merupakan tahapan skenario pengujian pada penelitian ini diantaranya.

   a. Pada pengujian ini menggunakan *K-fold cross validation* dengan jumlah *fold* atau K=10 yang berarti 90% dataset digunakan untuk proses *training* dan 10% sisanya akan digunakan sebagai *testing* pada

setiap *fold* atau perulangan dari *cross validation*. Penggunaan K=10 karena semakin besar nilai K – *fold* maka semakin besar data *training* yang digunakan dan semakin kecil data *testing* yang digunakan sehingga dapat meningkatkan nilai akurasi [6].

b. Kemudian diseuaikan dengan model SVM dan dari model tersebut akan menghasilkan nilai prediksi.
c. Menghitung hasil nilai prediksi untuk mendapatkan nilai *precision, recall, f1-score* dan *accuracy*.
d. Kemudian dari 11 fitur yang ada pada dataset dikombinasikan untuk menguji dampak dari suatu fitur terhadap hasil prediksi. Fitur-fitur yang sudah dikelompokkan kemudian diujikan kedalam K-fold cross validaton dengan K = 10.

2. Analisis.

Proses pengujian sistem telah menghasilkan nilai *precision, recall, f1-score* dan *accuracy* [8] guna mengetahui seberapa baik metode dan sistem yang dibangun untuk melakukan proses identifikasi akun palsu . Berikut merupakan tabel 2 yang memuat informasi hasil dari pengujian sistem menggunakan *K-fold cross validation* yang dilakukan pada penelitian ini.

Table 2 Tabel hasil pengujian

Hasil pengujian	Seluruh Atribut	Status Verified	Jumlah Tweet 2 Minggu Terakhir	Jumlah Tweet, Status Verified, Device Android, Device Web, Geolokasi	Jumlah Tweet, Device Android, Device Web, Geolokasi, Status Protected
Precision	88.12 %	64.61 %	81.87 %	90.89 %	90.89 %
Recall	94.59 %	100 %	89.80 %	95.32 %	95.32 %
F1-Score	90.88 %	76.66 %	85.08 %	92.92 %	92.92 %
Accuracy	91.14 %	70.00 %	84.94 %	93.42 %	93.42 %

Table 2 menunjukan hasil performansi setelah dilakukan pengujian *K-fold cross validation*, beberapa hal yang sangat mempengaruhi hasil performansi diantaranya.

1. Hasil pengujian pada penelitian ini menghasilkan nilai akurasi tertinggi sebesar 93.42% hal tersebut merupakan hasil dari rata rata setiap fold untuk nilai akurasi. Hal yang mempengaruhi nilai tersebut dikarenakan hasil prediksi dan hasil labelling tidak terpaut jauh perbedannya yang memiliki dampak terhadap nilai pada confusion matrix baik.
2. Banyaknya parameter dan kombinasi parameter yang digunakan dalam menentukan akun palsu juga mempengaruhi kepada baik buruknya hasil akhir pada akurasi dari penelitian ini. Jika semua parameter digunakan akan memungkinkan hasil identifkasi akun palsu akan menjadi lebih baik dibandingkan dengan jika hanya sebagian parameter yang digunakan, dan parameter jumlah tweet 2 minggu terakhir memiliki akurasi tertinggi diantara atribut individual lainnya.
3. Kombinasi atribut dengan nilai akurasi terbaik adalah kombinasi atribut jumlah tweet 2 minggu terakhir, status *verified*, pemakaian *device* Android, pemakaian *device* web, dan status geolokasi juga kombinasi atribut atribut jumlah tweet 2 minggu terakhir, pemakaian *device* Android, pemakaian *device* web, status geolokasi, dan status *protected* dengan keduanya memiliki nilai akurasi 93.42%.

5. Kesimpulan

Berdasarkan hasil pengujian serta analisis sebelumnya, maka diperoleh kesimpulan sebagai berikut

1. Berdasarkan hasil penelitian ini fake account dapat diidentifikasi dengan menggunakan metode SVM (*Support vector machine*) dengan menghasilkan *accuracy* sebesar 93.42% jika menggunakan seluruh atribut jumlah tweet, memakai twitter untuk Android, pemakaian twitter untuk web, Geolokasi, status protected atau status verified. Hal ini menandakan bahwa metode SVM memiliki performa yang cukup baik dalam melakukan identifikasi jika dibandingkan dengan hasil penelitian yang berjudul "*Fake Account Detection in Twitter Based on Minimum Weighted Feature Set*" pada [3] dimana hasil akurasi pada penelitian tersebut ketika menggunakan metode SVM hanya sebesar 86.62%.
2. Berdasarkan hasil penelitian, parameter *verified status* akun tidak memiliki dampak yang besar terhadap hasil penelitian jika dibandingkan dengan parameter jumlah tweet 2 minggu terakhir. Ini dikarenakan status *verified* dari suatu akun hanya berlaku untuk akun yang sudah terverifikasi, sedangkan untuk akun-akun yang belum terverifikasi tidak memiliki status *verified* sehingga dengan parameter *verified* saja belum cukup untuk menentukan asli atau tidaknya suatu akun.

3. Parameter jumlah *followers* dan *following* berdampak kecil terhadap hasil penelitian karena terdapat akun asli yang memiliki jumlah *followers* dan *following* sedikit, namun terdapat akun palsu yang memiliki *followers* dan *following* banyak.

Daftar Pustaka

[1] P. S. Rao, G. Narsimha and J. Gyani, "Fake Profile Identification in Online Social Networks Using Machine Learning and NLP," *International Journal of Applied Engineering Research*, vol. 13, no. 6, pp. 4133-4136, 2018.

[2] J. Castellini, V. Poggioni and G. Sorbi, "Fake twitter followers detection by denoising autoencoder," in *Proceedings - 2017 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2017*, Perugia, 2017.

[3] A. E. Azam, A. M. Idrees, M. A. Mahmoud and H. Hefny, "Fake Account Detection in Twitter Based on Minimum Weighted Feature set," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 10, no. 1, pp. 13-18, 2016.

[4] A. Romanov, A. Semenov, O. Mazhelis and J. Veijalainen, "Detection of Fake Profiles in Social Media," *Proceedings of the 13th International Conference on Web Information Systems and Technologies (WEBIST 2017)*, pp. 363-369, 2017.

[5] R. M. Neal, "Pattern Recognition and Machine Learning," *Technometrics*, vol. 49, no. 3, pp. 366-366, 8 2007.

[6] A. S. Nugroho, A. B. Witarto and D. Handoko, "Application of Support Vector Machine in Bioinformatics," in *Proceeding of Indonesian Scientific Meeting in Central Japan*, Gifu, 2003.

[7] E. Gaussier and C. Goutte, "A Probabilistic Interpretation of Precision, Recall and F1-score with implication for evaluation," Springer, 2005.

[8] J. Davies and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, 2006.