Prediksi *Retweet* Dengan Fitur Berbasis Pengguna dan Tingkat Sentimen Menggunakan Metode Klasifikasi Naive Bayes

Dionisio Febrianto[1], Jondri[2], Kemas Muslim Lhaksmana[3]

[1,2,3] Universitas Telkom, Bandung
[1]dionisiofebrianto@students.telkomuniversity.ac.id, [2]jondri@telkomuniversity.ac.id, [3]kemasmuslim@telkomuniversity.ac.id

Abstrak

Penggunaan jejaring sosial sebagai sarana membagian informasi telah banyak digunakan, salah satunya adalah jejaring sosial Twitter. Twitter memiliki fitur *retweet* untuk membagikan ulang sebuah *tweet* sehingga sebuah informasi dapat menyebar ke pengguna Twitter lainya. Penelitian ini bertujuan untuk membagun sistem prediksi *retweet* dengan fitur berbasis pengguna dan tingkat sentimen dari *tweet* dengan topik covid 19 berbahasa Indonesia. Hasil penelitian ini dapat digunakan untuk mendeteksi *influencer* yang *tweet*nya pada topik tertentu akan *retweet* oleh pengikutnya sehingga informasi yang dibagikan dapat menyebar luas. Untuk melakukan pemodelan digunakan Naïve Bayes sebagai metode klasifikasi. Hasil performa yang diperoleh dari sistem prediksi *retweet* ini yaitu akurasinya sebesar 0.78 dan *f1-score*nya sebesar 0.769.

Kata kunci : Twitter, Naïve Bayes, Fitur Berbasis Pengguna, Tingkat Sentimen

Abstract

The use of social networks as a means of sharing information has been widely used, one of which is the Twitter social network. Twitter has a retweet feature to re-share a tweet so that information can be spread to other Twitter users. This study aims to build a retweet prediction system with user-based features and the sentiment level of tweets with the topic of covid 19 in Indonesian. The results of this study can be used to detect influencers whose tweets on certain topics will be retweeted by their followers so that the information shared can be spread widely. To perform the modeling used Naïve Bayes as a classification method. The performance results obtained from this retweet prediction system are the accuracy of 0.78 and the f1-score of 0.769

Keywords: Twitter, Nave Bayes, User Based, Sentiment Level

1. Pendahuluan

Latar Belakang

Jejaring sosial semakin populer sebagai media untuk melakukan penyebaran informasi. Berdasarkan data dari datareportal.com pada januari 2021 Youtube menjadi sosial media paling banyak digunakan di Indonesia disusul Whatsapp, Instagram, Facebook dan Twitter pada urutan ke lima. Dengan menggunakan situs tersebut, orang – orang dapat membagikan informasi dengan berbagai topik sesuai dengan kesukaan dan minat mereka. Situs jejaring sosial seperti Facebook dan Twitter menunjukan potensi yang luar biasa untuk membuat konten menjadi populer secara instan[3]. Jejaring sosial menjadi tempat berbagai informasi populer seperti penyebaran berita terkini, penyebaran informasi selama keadaan darurat, kampanye, pemasaran dan lainya[3]. Memahami dan memprediksi penyebaran informasi di media sosial sangat penting[6]. Sebagai contoh terdapat 115 berita palsu pro Trump dibagikan sebanyak 30 juta kali di Facebook[12]. Berita palsu tersebut dibagikan dan berhasil menjangkau para pemilih Trump. Penyebaran informasi melalui jejaring sosial dapat dimanfaatkan untuk sistem rekomendasi, deteksi topik yang sedang populer, penyebaran kepercayaan, dan lainya[3]. Tujuan dari memahami penyebaran informasi adalah dapat memodelkan dan melakukan prediksi suatu peristiwa[2].

Twitter adalah salah satu situs jejaring sosial yang cukup banyak digunakan di indonesia.Pengguna_twitter dapat membuat akun dan membuat postingan yang biasa dikenal sebagai *tweet*. Pengguna twitter dapat membuat *tweet* dengan maksimal 240 karakter yang dapat terdiri dari *hashtag* dan *URL(Universal Resource Locator)* ataupun berisi gambar atau video. Selain itu pengguna Twitter dapat membagikan *tweet* milik orang lain yang disebut dengan *retweet*.

Retweet merupakan fitur yang disediakan oleh Twitter sebagai mekanisme penyebaran informasi[1]. Penelitian sebelumnya terkait penyebaran informasi di Twitter salah satunya adalah penelitian yang dilakukan oleh [3]. Peneliti [3] menggunakan fitur *latent topic information, user preferences, network connections, user activity & response, interaction, similarity between users, content-user similarity* dan menggunakan model *Bayesian network*. Sedangkan peneliti [6] menggunakan fitur berbasis pengguna, konten, dan waktu dengan menggunakan berbagai jenis metode klasifikasi.

Penelitian tugas akhir ini berfokus pada prediksi penyebaran informasi di jejaring sosial twitter dengan memprediksi apakah sebuah *tweet*(postingan) akan disebarkan(*retweet*) atau tidak dengan topik covid 19 . Pada penelitian ini, penulis akan menggunakan metode klasifikasi naïve bayes sebagai metode kalsifikasi. Metode klasifikasi naïve bayes dipilih karena metode klasifikasi naïve bayes memiliki tingkat akurasi yang lebih baik dibanding model klasifikasi lainya[5]. Fitur yang akan digunakan adalah fitur berbasis pengguna dan tingkat sentiment. Tingkat sentiment digunakan sebagai fitur karena *tweet* yang sangat positif atau negatif cenderung di *retweet* oleh orang lain[6].

Topik dan Batasannya

Berdasarkan latar belakang yang ada, maka dapat ditentukan topik masalah yaitu prediksi *retweet* dengan menggunakan fitur berbasis pengguna dan tingkat sentiment menggunakan metode klasifikasi naïve bayes. Adapun batasan masalah yang diangkat yaitu sistem dapat melakukan prediksi apakah sebuah *tweet* akan di *retweet* oleh pengguna Twitter lainya atau tidak.

Tujuan

Tujuan dari penelitian ini yaitu untuk membangun sistem prediksi *retweet* dengan fitur berbasis pengguna dan sentiment level menggunakan metode klasifikasi naïve bayes serta mengetahui performansi yang didapat dari sistem yang telah dibangun.

Organisasi Tulisan

Pada bab kedua akan dibahas mengenai penelitian sebelumnya yang serupa serta hasil penelitian tersebut serta membahas studi terkait penelitian tugas akhir ini. Pada bab tiga akan dibahas mengenai sistem yang akan dibangun, dan pada bagian empat akan dibahas mengenai hasil mengenai hasil evaluasi keluaran sistem yang telah dibangun kemudian analisis dari hasil pengujian yang telah dilakukan. Bagian terakhir berisi kesimpulan dari penelitian tugas akhir yang telah dilakukan.

2. Studi Terkait

2.1 Penelitian Serupa

Sejumlah penelitian terdahulu telah mempelajari proses dari penyebaran informasi melalui media social. Terdapat berbagai fitur yang digunakan dan algoritma klasifikasi yang digunakan. Pada penelitian yang dilakukan oleh [3] menggunakan beberapa fitur sebagai berikut: *Latent Topic Information, User Preferences, Network Connections, User Activity & response , interaction, similarity between users, content-user similarity*. Peneliti [3] memilih topik *Shopping,politics,social media,festive season* dalam melakukan pengujiannya dengan detail akurasi yang dihasilkan dinyatakan pada tabel 1

Tabel 1: Hasil akurasi peneliti [3]

	Precision	Recall	F1-Score
Shopping	86.36%	83.44%	84.87%
Politics	78.34%	91.54%	84.43%
Social Media	84.79%	82.00%	83.37%
Festive Season	78.08%	90.81%	83.96%

Peneliti [6] melakukan penelitian untuk melakukan prediksi *retweet* dari sebuah *tweet* dengan menggunakan tiga jenis fitur yaitu fitur berbasis pengguna, fitur berbasis konten, dan fitur berbasis waktu. Metode klasifikasi yang digunakan peneliti tersebut adalah *Naïve Bayes*(NB), *Support Vector Machine*(SVM), dan *Random Forest*(RF). Untuk melakukan pembagian data menggunakan 10-*fold Cross Validation*. Dari penelitian [6] dihasilkan bahwa *Random Forest* menghasilkan skor *F-measure* 5% lebih baik dibanding model lain dari total 16 juta *tweet*.

2.2 Twitter

Twitter adalah jejaring sosial yang didirikan pada tahun 2006. Twitter adalah sistem untuk berbagi sebuah informasi, dimana seorang pengguna dapat mengikuti pengguna lain untuk mendapat informasi yang pengguna tersebut bagikan[4]. Informasi tersebut terdiri dari pesan singkat yang disebut *tweet*[4]. Di twitter pengguna dapat membuat *tweet* tidak lebih dari 240 karakter, *tweet* tersebut dapat berisi *URL(Universal Resource Locator)*, *hashtag*(kata kunci yang diikuti dengan simbol "#"), *mentions*(menandai nama pengguna lainya dengan simbol "@"), dan *emoticons*[1].

Sebuah *tweet* yang dibuat oleh salah satu pengguna twitter dapat dibagikan oleh pengguna twitter lainnya, hal ini disebut sebagai *retweet*. *Retweeter*(orang yang melakukan *retweet*) diperbolehkan menambahkan komentarnya terhadap *tweet* tersebut dan teks dari *retweeter* akan ditempatkan di awal *retweet* tersebut[1]. Pengguna twitter menggunakan *hastag* (#) untuk mengidentifikasi topik tertentu[4]. *Hastag* atau *keywords* yang paling banyak diperbincangkan akan menjadi *trending topic*[4].

2.3 Naïve Bayes

Naive Bayes adalah salah metode klasifikasi yang menggunakan probabilitas dan statistik yang dikemukakan oleh Thomas Bayes. Probabilitas tersebut digunakan untuk memprediksi kelas dari data masukan[3]. Naïve Bayes memprediksi peluang yang akan datang berdasarkan pengalaman sebelumnya. Naïve Bayes memiliki asumsi bahwa setiap atributnya independen satu sama lain[3]. Jika y adalah suatu kelas dan $x_1, x_2, x_3, \dots, x_n$ adalah nilai yang akan di observasi, Maka nilai probabilitas masing-masing kelasnya adalah:

$$P(y|x_1, x_2, x_3, \dots, x_n) = \frac{\mathrm{P}(x_1, x_2, x_3, \dots, x_n|y)\mathrm{P}(y)}{\mathrm{P}(x_1, x_2, x_3, \dots, x_n)} \quad (1)$$

Dengan asumsi bahwa $x_1, x_2, x_3, \dots, x_n$ tidak saling tergantung(independen), maka probabilitas masing-masing kelasnya adalah:

$$P(y|x_1, x_2, x_3, \dots, x_n) = \frac{\mathrm{P}(y)\prod_{i=1}^{n} P(x_i|y)}{\mathrm{P}(x_1, x_2, x_3, \dots, x_n)} \quad (2)$$

Karena $\mathrm{P}(x_1, x_2, x_3, \dots, x_n)$ memiliki nilai yang sama untuk $P(y_i|x_1, x_2, x_3, \dots, x_n)$ dan $P(y_j|x_1, x_2, x_3, \dots, x_n)$, maka persamaan (2) dapat ditulis sebagai berikut:

$$P(y|x_1, x_2, x_3, \dots, x_n) = \mathrm{P}(y)\prod_{i=1}^{n} P(x_i|y) \quad (3)$$

Naïve Bayes memaksimalkan probabilitas dari masing masing kelas[13]. Kelas maksimum akan menjadi kelas keputusan

$$\hat{y} = \arg max\, \mathrm{P}(y)\prod_{i=1}^{n} P(x_i|y) \quad (4)$$

Untuk melakukan klasifikasi dengan data kontinyu digunakan persamaan *Densitas Gauss*

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma^2}} \quad (5)$$

Keterangan:

x: atribut

σ: simpangan baku

π: nilai phi

μ: rata rata

y: kelas yang dicari

2.4 Fitur Berbasis Pengguna

Seseorang yang sering berinteraksi dengan orang lain akan mendapat respon yang sesuai[6]. Jadi interaksi antar pengguna yang mengirim tweet akan diperhitungkan [6]. Berikut adalah fitur berbasis pengguna menurut [6]:

Tabel 2: Atribut fitur berbasis pengguna

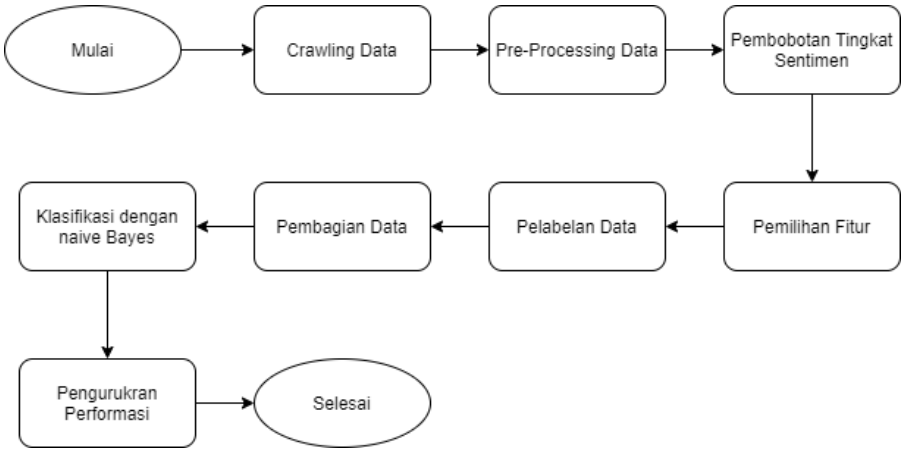
Features	Deskripsi	Tipe data
1. Total_ofOtweet	Total *tweet* sebelumnya yang telah di posting oleh pengguna.	*Nummeric*
2. No_of_followers	Jumlah orang mengikuti pengguna	*Nummeric*
3. No_of_followees	Jumlah orang yang dikuti oleh pengguna	*Nummeric*
4. Age_of_account	Jumlah hari sejak akun dibuat user	*Nummeric*
5. No_of_favourite	Jumlah *tweets* yang disukai oleh pengguna	*Nummeric*
6. No_groups_user_belongs	Jumlah grub pengguna	*Nummeric*
7. Aver_favou_per_day	Rata-rata like yang diterima perhari	*Nummeric*
8. Aver_tweets_per_day	Rata-rata *tweets* yang dibuat perhari	*Nummeric*
9. User_name_len	Panjang *username* pengguna	*Nummeric*

2.5 Sentiment Analysis

Sentimen analisis atau penambangan opini adalah studi yang mempelajari dan menganalisis pendapat orang, sentiment, evaluasi ,sikap, penilaian dan emosi terhadap *entities* seperti produk,layanan,organisasi,individu,masalah,peristiwa dan lainya [7]. Dalam kehidupan sehari-hari perusahaan yang menjual sebuah produk ingin mencari opini dari konsumen tentang produk yang mereka jual. Karena itu diperlukan sentimen analisis untuk mengambil sebuah keputusan. Sentimen analisis menghasilkan nilai positif atau negatif terhadap sebuah produk, layanan, organisasi, individu, masalah ,peristiwa dan lainya[7]. Tingkat sentimen dari *tweet* yang sedang ramai dibicarakan memiliki nilai sangat positif atau sangat negative dan *tweet* ini lebih cenderung di *retweet* [6].

3. Sistem yang Dibangun

3.1 Gambaran Sistem



Gambar 1: Gambaran sistem

3.2 Pengumpulan Data

Pengumpulan data dilakukan dengan proses *crawling* data Twitter dengan menggunakan twitter API(*Application program interface)* yang telah disediakan oleh Twitter. Tahapan pengumpulan data ini dilakukan dengan mencari *tweet* berbahasa Indonesia dengan kata kunci "covid 19"

3.3 *Pre-Processing* Data

Pre-processing data merupakan tahapan yang bertujuan untuk mengolah data untuk menyiapkan data yang akan menjadi *input* ke dalam sistem. Berikut adalah Langkah-langkah *pre processing* data yang dilakukan:

3.3.1 Data *cleaning*

Proses pembersihan data dilakukan untuk menghilangkan data yang memiliki nilai atribut yang kosong dan menghilangkan data ganda.

3.3.2 Data *Normalization*

Data *Normalization* yang dilakukan menggunakan metode *Min-Max Normalization*. *Min-Max Normalization* bekerja dengan merubah setiap fitur yang ada menjadi nilai rentang antara 0 dan 1 dengan persamaan sebagai berikut:

$$X_{new} = \frac{X_{old} - X_{Max}}{X_{Max}} \quad (6)$$

3.4 Pemberian Bobot Tingkat Sentimen

Proses pembobotan tingkat sentimen dilakukan untuk memberikan nilai sentimen dari setiap *tweet* yang telah didapat pada proses pengumpulan data. Pemberian nilai tingkat sentiment dilakukan menggunakan *tools orange data mining*.

3.5 Pemilihan Fitur

Tahapan pemilihan fitur dilakukan untuk menghilangkan data yang dianggap tidak relevan untuk digunakan pada proses klasifikasi. Fitur yang akan digunakan pada penelitian ini yaitu: jumlah *tweet*, jumlah *like* yang diterima sebuah akun, jumlah pengikut, jumlah yang diikuti, rata rata *tweet* per hari, rata rata *like* perhari, status verifikasi akun, umur akun, tingkat sentiment.

3.6 Pelabelan Data

Pada tahapan ini melakukan pelabelan data dari dataset menjadi dua kelas yaitu kelas 0 dan 1. Kelas 0 terdiri dari data yang memiliki jumlah *retweet* sama dengan nol. Sedangkan kelas 1 terdiri dari data yang memiliki jumlah *retweet* lebih dari atau sama dengan 1.

3.7 Pembagian Data

Dataset yang digunakan akan dibagi menjadi *validation fold* dan *traning fold* dengan metode *K-fold cross validation* . Nilai K yang digunakan pada penelitian ini adalah 5.

3.8 Perhitungan Performansi

Perhitungan performansi dari model klasifikasi yang telah dibangun adalah menggunakan perhitungan akurasi, dan *f1-score*.

4. Evaluasi

Pada penelitian tugas akhir ini menggunakan dataset berjumlah 14218 data. Kelas 0(*tweet* yang tidak diretweet) terdapat 11537 data dan kelas 1(*tweet* yang diretweet) terdapat 3615 data. Karena pada dataset terjadi *imbalance class* maka pada pengujian dilakukan sampling dengan metode *random oversampling* dan *random undersampling*.

4.1 Hasil Pengujian

4.1.1 Hasil Pengujian Tanpa *Sampling*

Pada pengujian tanpa *sampling* dilakukan dengan membagi dataset menjadi data latih dan data uji. Pembagian dataset tersebut dilakukan dengan metode *k-fold cross validation* dengan nilai k = 5. Metode klasifikasi yang digunakan pada penelitian ini adalah naïve bayes yang akan melakukan pemodelan untuk memprediksi *retweet*. Hasil evaluasi dihitung menggunakan perhitungan akurasi dan *f1-score*. Berikut adalah hasil dari pengujian yang dilakukan:

Tabel 3: Hasil pengujian tanpa sampling

Nilai K	Akurasi	*F1-Score*
1	0.79	0.78
2	0.77	0.76
3	0.76	0.76
4	0.79	0.77
5	0.77	0.76
Rata-rata	0.78	0.769

4.1.2 Hasil Pengujian Dengan *Random Undersampling*

Pengujian dengan menggunakan metode *random undersampling* dilakukan dengan cara data pada kelas mayoritas akan dikurangi secara acak sehingga data latih pada kelas mayoritas(kelas 0) memiliki jumlah yang sama dengan kelas minoritas(kelas 1) yaitu sejumlah 3615 data. Berikut merupakan hasil dari pengujian yang dilakukan:

Tabel 4: Hasil pengujian dengan *undersampling*

Nilai K	Akurasi	*F1-Score*
1	0.65	0.63
2	0.64	0.62
3	0.65	0.62
4	0.63	0.60
5	0.65	0.63
Rata-rata	0.65	0.625

4.1.3 Hasil pengujian Dengan *Random Oversampling*

Pengujian dengan menggunakan metode *random oversampling* dilakukan dengan cara data pada kelas minoritas akan ditambah hingga banyak data pada kelas minoritas sama dengan jumlah data pada kelas mayoritas yaitu sebanyak 11537 data. Data yang ditambahkan pada kelas minoritas berasal dari duplikasi data pada kelas minoritas yang dipilih secara acak. Berikut merupakan hasil dari pengujian yang dilakukan:

Tabel 5: Hasil pengujian dengan *oversampling*

Nilai K	Akurasi	*F1-Score*
1	0.64	0.62
2	0.64	0.62
3	0.64	0.62
4	0.66	0.63
5	0.64	0.61
Rata-rata	0.64	0.625

4.2 Analisis Hasil Pengujian

Berdasarkan hasil pengujian untuk melakukan prediksi *retweet* dengan fitur berbasis pengguna dan tingkat sentiment menggunakan metode klasifikasi *naïve bayes* menghasilkan nilai akurasi dan *f1-score* terbaik dengan data yang tidak dilakukan sampling yaitu sebesar 0.78 untuk rata-rata akurasi dan 0.76 untuk rata-rata nilai *f1-score*. Sedangkan untuk pengujian menggunakan *random undersampling* didapat akurasi rata-rata sebesar 0.65 dan 0.625 untuk nilai rata-rata *f1-score*. Untuk pengujian dengan menggunakan metode *random oversampling* didapat rata rata akurasi sebesar 0.64 dan 0.625 untuk rata-rata nilai *f1-score*. Dari hasil pengujian tersebut didapat bahwa hasil pengujian dengan tanpa menerapkan metode *random undersampling* dan *random oversampling* memiliki nilai akurasi dan *f1-score* yang lebih baik.

5. Kesimpulan

Dari penelitian tugas akhir ini didapat kesimpulan prediksi *retweet* dengan menggunakan fitur berbasis pengguna dan tingkat sentimen menggunakan metode klasifikasi *naïve bayes* mendapatkan hasil performa yang cukup baik dengan pembagian dataset menggunakan 5-*fold cross validation*. Performa terbaik didapat pada pengujian dengan tidak melakukan sampling pada *dataset* yang mengalami *imbalance class* hal ini dapat ditunjukan pada tabel 3. Untuk mengatasi dataset yang mengalami *imbalance class* digunakan metode *random undersampling* dan *random oversampling* namun hasil performa dari kedua metode tersebut mengalami penurunan hal ini dapat ditunjukan pada tabel 4 dan 5. Performa terbaik yang didapat yaitu 0.78 untuk nilai rata-rata akurasi dan 0.76 untuk rata-rata nilai *f1-socre*.

REFERENSI

[1] Firdaus, S. N., Ding, C., & Sadeghian, A. (2018). Retweet: A popular information diffusion mechanism–A survey paper. *Online Social Networks and Media*, *6*, 26-40.

[2] Molaei, S., Zare, H., & Veisi, H. (2020). Deep learning approach on information diffusion in heterogeneous networks. *Knowledge-Based Systems*, *189*, 105153.

[3] Varshney, D., Kumar, S., & Gupta, V. (2017). Predicting information diffusion probabilities in social networks: A Bayesian networks based approach. *Knowledge-Based Systems*, *133*, 66-76.

[4] Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010, July). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)* (Vol. 6, No. 2010, p. 12).

[5] Xhemali, D., J HINDE, C., & G STONE, R. (2009). Naïve bayes vs. decision trees vs. neural networks in the classification of training web pages. *D. XHEMALI, CJ HINDE and Roger G. STONE," Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages", International Journal of Computer Science Issues, IJCSI, Volume 4, Issue 1, pp16-23, September 2009*, *4*(1)

[6] Hoang, T. B. N., & Mothe, J. (2018). Predicting information diffusion on Twitter–Analysis of predictive features. *Journal of computational science*, *28*, 257-264

[7] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

[8] Sembodo, J. E., Setiawan, E. B., & Baizal, Z. A. (2016, August). Data Crawling Otomatis pada Twitter. In *Indonesian Symposium on Computing (Indo-SC)* (pp. 11-16).

[9] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6.

[10] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437.

[11] Singh, N., Singh, A., & Sharma, R. (2020). Predicting Information Cascade on Twitter Using Random Walk. *Procedia Computer Science*, *173*, 201-209.

[12] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-36.

[13] Nugroho, A., & Subanar, S. (2015). Klasifikasi Naïve Bayes untuk Prediksi Kelahiran pada Data Ibu Hamil. *Berkala Ilmiah MIPA*, *23*(3), 241908.

[14] Kemp, S. 2021. Digital 2021: Indonesia. [Online] Available at: https://datareportal.com/reports/digital-2021-indonesia [Accessed 28 July 2021].