Prediksi Penyebaran Informasi di Twitter dengan Metode Pembelajaran Mesin dengan Fitur Linimasa

http://dx.doi.org/10.28932/jutisi.v7i1.3324
Riwayat Artikel
Received: 23 Januari 2021 | *Final Revision*: 5 Maret 2021 | *Accepted*: 12 Maret 2021

Lucky Surya Haryadi[#1], Bernard Renaldy Suteja[✉#2]

[#]*Program Studi Magister Ilmu Komputer Fakultas Teknologi Informasi,*

Universitas Kristen Maranatha

Jl. Prof.drg. Suria Sumantri No.65, Bandung

[1]suryaharyadi@gmail.com

[2]bernard.rs@it.maranatha.edu

Abstract — **Social Media Network has been an important information source, and the information propagation within the network gave an impact on politics, marketing, and entertainment industry. Our aim is to predict a tweet whether the information will be propagated further. The previous research has focused on analyzing this task with a wide range of learning methods and features, such as content and account features. Timeline features are proposed as features that can further predict information propagation and as we compared the performance with content and account features. The dataset consists of 43.229 tweets, we predict the information propagation with logistic regression, support vector machines, and random forest learning method with these features. Our result indicates that the timeline feature can be a good candidate for predicting information propagation and the random forests learning method consistently performs better. From the training result, we further calculate feature importance. Recently tweets, engagement with another user and previous liked tweets on the timeline features contributed to more popular tweets.**

Keywords — **Information Diffusion, random forest, Predicting information propagation, Tweet propagation**

I. PENDAHULUAN

Sosial media telah digunakan sebagai salah satu sumber informasi yang digunakan secara umum[1]. Twitter, sebuah mikroblog, adalah sebagai salah satu sosial media yang dimanfaatkan oleh penyiar berita dari berbagai bidang seperti politik, pemasaran, dunia hiburan, olahraga, deteksi bencana alam hingga penyebaran berita bohong. Pertumbuhan informasi yang tercatat dengan berapa kali sebuah *tweet* dibagikan ulang disebut dengan *retweet*, dan jumlah kemunculannya dapat digunakan sebagai pengukur popularitas.

Penyebaran informasi di Twitter telah menjadi perhatian baik dalam bidang penelitian maupun bidang komersial. Twitter berisi cuitan singkat dan dibatasi 140 kata namun memiliki motif informasi seperti menggunakan @ untuk menyebut pengguna lain, tagar, dan media berisi foto atau video dimanfaatkan sebagai prediktor pertumbuhan *retweet*[2][3]. Suh et al menemukan hubungan antara popularitas *tweet* dengan popularitas pengguna yang diukur dengan jumlah (*follower*) pengikut[4]. Berkebalikan dengan hal yang berkembang dalam dunia penelitian, bidang komersial lebih berfokus pada aktivitas pengguna.

Para pembuat konten mencoba merumuskan hal yang dapat dilakukan pengguna yang tidak hanya didapatkan dari satu *tweet* namun melalui aktivitasnya yang tercermin pada linimasa[5][6]. Fitur yang akan diuji adalah fitur linimasa yang diambil berdasarkan 200 *tweet* sebelumnya sebagai prediktor jumlah *retweet*.

Berbagai macam metode pembelajaran dalam prediksi klasifikasi *retweet* telah diuji. Menggunakan pembelajaran mesin menyiratkan bahwa (1) setiap *tweet* diwakili oleh serangkaian fitur (2) satu set pelatihan digunakan untuk belajar dan pengujian. Xu et al [7] melakukan prediksi dalam perspektif twitter individu dan mencoba melakukan prediksi dengan metode *decision tree*, regresi logistik dan s*upport vector machine* (SVM)[7]. Metode pembelajaran *random forest* telah diujikan dalam penelitian sebelumnya dan menunjukan hasil yang lebih baik dibanding metode yang sebelumnya diuji[8][9][10].

Tujuan penelitian ini adalah melihat (i) apakah dengan mengamati aktivitas user pada linimasa dapat menjadi fitur

yang dapat memprediksi penyebaran informasi dan (ii) mencari algoritma pembelajaran mesin yang terbaik dengan memanfaatkan fitur pengguna, fitur konten dan fitur lini masa dan (iii) mencari prediktor fitur terbaik dengan mencari *feature importance* dari hasil pelatihan pembelajaran mesin.

II. Penelitian Terkait

Xu, et al. [7] menganalisis perilaku retweet pengguna pada tingkat individu dan berpendapat bahwa fitur terpenting bagi orang umum adalah sosial. Dengan melakukan perbandingan antar fitur, diidentifikasi faktor yang terkait dengan perilaku *retweet* pengguna adalah fitur akun. Kita juga menambahkan beberapa fitur baru yaitu fitur linimasa dengan pembanding fitur akun, dan fitur konten yang telah dimanfaatkan.

Cheng[9] mengembangkan kerangka kerja agar dapat memprediksi aliran informasi pada media sosial. Aliran informasi dijadikan sebagai masalah klasifikasi biner pembagian ulang *k* pertama dari aliran informasi diamati dan diprediksi apakah ukuran akhirnya dari aliran informasi mencapai media ukuran semua aliran informasi dengan setidaknya dua kali. Ini memungkinkan aliran informasi bervariasi dengan dapat diprediksi dengan lebih sederhana.

Dengan mengamati 5 pembagian ulang pertama dari penyebaran ulang dan diprediksi apakah akan mencapai ukuran penyebaran ulang nilai median[10]. Secara keseluruhan, sementara setiap set fitur secara individual lebih baik daripada memprediksi secara acak, itu adalah set fitur temporal yang mengungguli semua individu lainnya.

Model penyebaran secara kompleks dengan struktur jaringan *social reinforcement* dan *homophily*[10]. Weng menguji model penyebaran kontaminasi *meme* menyebar seperti kontaminasi sederhana dan pola penyebaran dapat diprediksi melalui pola penyebaran awal pada komunitas. Semakin awal *meme* menyebar pada komunitas maka *meme* akan semakin viral.

Dengan menerapkan algoritma pembelajaran mesin, *random forest*, untuk memprediksi *meme* mana yang dapat menjadi viral di masa mendatang. Fitur berbasis komunitas yang digunakan dalam pengklasifikasi secara signifikan meningkatkan hasil prediksi.

Yang et al.[11] juga mempelajari proses *retweet* di jejaring sosial. Dari pengamatan mereka pada data twitter, ditemukan bahwa hampir 25,5% dari *tweet* yang diposting oleh pengguna sebenarnya di-*retweet* dari status teman mereka. Dari situ, mereka mengusulkan kerangka kerja semi-supervisi pada model graph faktor untuk memprediksi perilaku *retweet* pengguna Twitter. Fitur preferensi riwayat pengguna, konten pesan, dan informasi jejak dianggap tetapi tidak dijelaskan secara eksplisit. Hal ini menunjukan bahwa data histori *tweet* seseorang dapat menjadi acuan prediksi *retweet* berikutnya.

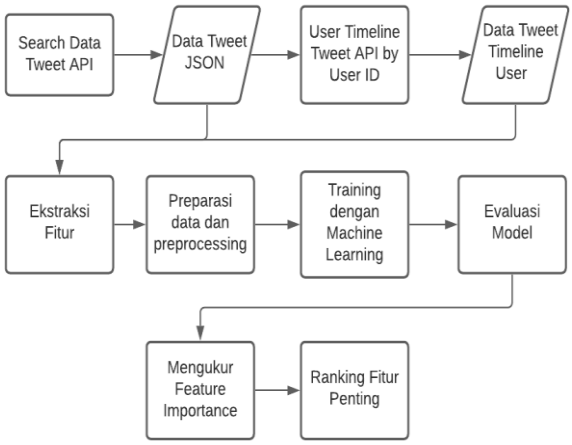
Bunyamin[8] dalam deteksi popularitas menemukan metode pelatihan yang memiliki performa terbaik adalah *random forest* dengan memanfaatkan fitur pengguna dan fitur konten. Dalam pencarian *feature importance* digunakan metode *mean decrease importance* (MDI) pada *random forest* dan *recursive feature elimination* (RFE) pada pelatihan regresi logistik, lalu nilai kepentingannya diurutkan.

Keterbatasan model yang dialami adalah tidak dapat memasukan fitur struktural dan fitur temporal. Fitur struktural membutuhkan model jeram informasi yang mengetahui dari siapa seorang pengguna mendapatkan informasi dalam bentuk graf. Namun *Application Programming Interface* (API) twitter pada saat ini tidak dapat mengetahui dari mana pengguna mendapatkan informasi dan melakukan *retweet*.

Fitur temporal yang biasa digunakan adalah mengukur kecepatan antar *retweet*. Dengan dataset yang kami miliki, terlalu banyak data yang hanya memiliki 1 *retweet* (RT) sehingga kesulitan untuk melakukan analisa dalam kecepatan.

III. Model dan Fitur

A. Desain Eksperimen



Gambar 1. Skema besar eksperimen dari mengumpulkan data Twitter hingga mencari Fitur Penting

Eksperimen ini seperti pada Gambar 1, dimulai dengan pengumpulan data dari API Twitter yaitu search API dan API linimasa pengguna. Data mentah yang didapatkan dalam bentuk *JavaScript Object Notation* (JSON) akan diseleksi fitur yang terbagi atas fitur akun, fitur konten dan fitur linimasa. Dataset yang telah dikumpulkan akan dilakukan pelatihan dengan metode regresi logistik, *Support Vector Machine* (SVM) dan *random forest*, sedangkan pengujian dengan metode pembelajaran mesin dan model akan dievaluasi dan dibandingkan antara tiga fitur dan model pembelajaran mesin yang diujikan. Hasil pelatihan akan dikalkulasi *feature importance* dan akan diurutkan fitur yang paling mempengaruhi model.

B. Dataset

Algoritma dalam kemunculan *Tweet* di linimasa berubah seiring waktu, dalam pertimbangan ini maka dilakukan *search* API *tweet* yang terbaru dalam bahasa Indonesia. Sebanyak 6 tagar digunakan dengan topik yang berbeda agar sistem ini dapat dimanfaatkan dalam berbagai bidang. Pengambilan data konten dengan search API Twitter dan akun diambil 6 tagar dalam rentang dari 7 Oktober hingga 10 November seperti pada Tabel 1. Pencarian dengan API search dan hasil pencarian akan didapatkan file JSON seperti contoh pada Gambar 2.

{ 'id': 1361134118998663168,
'id_str': '1361134118998663168',
'text': 'RT @musniumar: #Jerman keok sama Spanyol 6-0 boss',
'in_reply_to_status_id_str': None,
'in_reply_to_user_id': None,
'in_reply_to_user_id_str': None,
'in_reply_to_screen_name': None, 'user':
{'id': 826110341268463617,
'screen_name': 'NandaDaulay7',
'url': None,
'entities': {'description': {'urls': []}},
'protected': False,
'followers_count': 959,
'friends_count': 680,
'listed_count': 0,
'created_at': 'Mon Jan 30 16:50:28 +0000 2017',
'favourites_count': 30462, }

Gambar 2. Hasil pencarian yang ditampilkan dalam bentuk JSON

Proses yang dilakukan adalah mengumpulkan data yang terdapat pada JSON seperti mengumpulkan fitur akun seperti jumlah list, *followers_count*, *following_account*, umur akun, dan *verified*. Dari teks yang didapat dilakukan ekstraksi dengan melihat beberapa tanda seperti # sebagai tagar, dan @ sebagai penyebutan pada akun lainya yang akan digunakan sebagai fitur akun. Terdapat 43.229 *tweet* dengan jumlah *tweet* unik sebanyak 4.773. Dipilih berbagai topik dengan tujuan untuk mencari motif fitur yang umum pada prediksi penyebaran informasi.

TABEL I
TAGAR DAN JUMLAH *TWEET* YANG TELAH DIKUMPULKAN MELALUI TWITTER API

No.	Tagar	Jumlah Tweet	Topik
1	#Jerman	5082	Olah Raga
2	#UUCiptaker	16310	Politik
3	#merapi	7925	Bencana Alam
4	#mositidakpercaya	5919	Politik
5	#pahlawan	4833	Umum
6	#pilpres	3160	Politik

Dalam *tweet* unik sebanyak 4.773 tersebut juga dilakukan pengambilan data melalui API *user_timeline* sebanyak 200 *posting* dalam aktivitasnya dan agregasi data dalam fitur linimasa pengguna. Fitur akan dikumpulkan seperti fitur akun dan fitur konten dengan penggabungan sebanyak 200 aktivitas penggunanya. Hasil lengkap akan dilakukan proses normalisasi data dengan skala minimum dan maksimum untuk mengubah data berada di rentang 0 sampai 1.

Label *retweet* yang dianggap berkembang luas akan diberikan dengan metode yang digunakan oleh Cheng[9] yaitu menggunakan klasifikasi biner untuk memprediksi apakah ukuran akhir *tweet* setidaknya melebihi median dari dataset. Pada dataset kali ini median pembagian ulang adalah 1 sehingga jika *tweet* lebih besar dari satu akan diberi label 1 sedangkan dibawah 2 adalah 0. Pada metode pelatihan machine learning, benchmark dari tebakan acak memiliki akurasi 50%.

C. Deskripsi Fitur

Faktor-faktor yang menjadi prediktor dalam prediksi dibagi menjadi tiga kelas: fitur konten , fitur akun dan terakhir adalah fitur linimasa.

TABEL II
FITUR YANG DIGUNAKAN PADA AKUN

No.	Fitur	Deskripsi	Jenis Data
1	Umur_akun	Waktu semenjak penyiar pertama mendaftar Twitter	Numerik
2	Rasio_follower_friend	Perbandingan antara jumlah follower dengan following	Numerik
3	Tweet_frekuensi_umur_akun_per_hari	Frekuensi tweet penyiar per hari sepanjang umur_akun	Numerik
4	Verified	Verifikasi dari twitter untuk memberitahu pengguna bahwa akun bersangkutan asli	Boolean
5	Jumlah_favorit	Jumlah likes(disukai) sepanjang umur_akun	Numerik
6	Jumlah_listed	List adalah grup kurasi pada linimasa pengguna. Jumlah listed menunjukan popularitas pengguna selain follower.	Numerik
7	friends_count	Jumlah akun yang diikuti pengguna	Numerik
8	followers_count	Jumlah akun yang mengikuti pengguna	Numerik

Faktor pertama yang berkontribusi pada kemampuan informasi untuk menyebar adalah konten itu sendiri dengan detil pada Tabel II. Di Twitter, konten *tweet* dan khususnya, *hashtag*, *mention*, panjang *tweet* digunakan untuk

menghasilkan fitur konten[4]. Konten pada saat ini juga diperluas dengan fitur media seperti Gambar dan video yang juga dimasukan pada fitur.

Fitur akun pengguna ditemukan pada riset Petrovic [12] terutama jumlah *follower*, *listed*, dan *friend* sebagai prediktor konten yang di *retweet* lebih signifikan dibanding kontennya sendiri. Salganik[13] dan Dow[14] mengungkapkan, walaupun dengan isi konten yang sama, bagaimana menyebarnya sebuah informasi dapat berbeda bentuk penyebarannya. Fitur tambahan yaitu *verified* dan *rasio_follower_friend* juga dimanfaatkan. Akun *verified* ditemukan perbedaan bentuk *graph* dengan pengguna umum dan *rasio_follower_friend* yang rendah menjadi indikasi bot dan lebih sering melakukan *retweet*[15].

TABEL III
FITUR YANG DIGUNAKAN PADA KONTEN

No.	Fitur	Deskripsi	Jenis Data
1	Foto	Apakah konten terdapat foto?	Boolean
2	Video	Apakah konten terdapat foto?	Boolean
3	Panjang_text	Panjang huruf dalam satu *tweet*	Numerik
4	Hashtag	Apakah konten terdapat tagar	Boolean
5	User_Mention	Apakah konten terdapat @ dan menyebut akun orang lain	Boolean
6	Time_from_origin	Berapa lama *tweet* disebar ulang dari pertama kali disiarkan	Numerik

Fitur konten pada penelitian sebelumnya memiliki akurasi lebih rendah dibanding fitur lain[9][8][12]. Salah satu fitur penting lainya adalah fitur temporal, dan salah satu fitur temporal yang didapatkan akan dimasuki ke kelas konten karena cap waktu mulai penerbitan didapat pada konten pada Tabel III. Salah satu indikator yang biasa digunakan dan berkorelasi adalah kecepatan jumlah *retweet* seiring waktu. Namun dataset yang didapatkan dengan mayoritas median *Retweet* hanya 1 sehingga tidak dimasukan dalam fitur. *Time_from_origin* digunakan untuk mengukur sudah berapa lama sebuah konten dikirim, Ma Zhongyang[16] pada penelitian Twitter menemukan semakin lama sebuah konten maka akan semakin turun jumlah *retweet*.

TABEL IV
FITUR YANG DIGUNAKAN PADA LINIMASA

No.	Fitur	Deskripsi	Jenis Data
1	*In_reply_to*	Berapa kali pengguna melakukan fiture menjawab?	Numerik
2	*Quoted_UT*	Berapa kali pengguna menggunakan fitur quote?	Numerik
3	*Retweet*_UT	Berapa kali pengguna menggunakan fitur *retweet*?	Numerik
4	*Retweet*__count_UT	Berapa post pengguna mendapatkan *retweet*	Numerik
5	*viral100_UT*	Berapa post pengguna mendapatkan *retweet* lebih dari 100?	Numerik
6	*viral1000_UT*	Berapa post pengguna mendapatkan *retweet* lebih dari 1000?	Numerik
7	Foto_UT	Berapa banyak post terdapat foto?	Numerik
8	Video_UT	Berapa banyak post terdapat video?	Numerik
9	User_Mention_UT	Berapa banyak post terdapat @ dan menyebut akun orang lain?	Numerik
10	Hashtag_UT	Berapa banyak post terdapat hashtag?	Numerik
11	Favourite_UT	Berapa post pengguna mendapatkan liked	Numerik
12	average_*tweet*_per_day	Berapa kali pengguna rata-rata menyiarkan konten dalam 1 hari	Numerik

Fitur linimasa merupakan sebuah fitur baru dalam eksperimen kali ini. Aktivitas linimasa telah dimanfaatkan untuk mendeteksi apakah pengguna adalah bot atau asli[17], lalu tingkat depresi seseorang[18], apakah mengandung ujaran kebencian[19] seperti pada Tabel IV. Korelasi antara aktivitas pengguna dalam 200 aktivitas terakhirnya diuji apakah akan berpengaruh pada akurasi prediksi penyebaran konten yang dipublikasikannya.

Strategi keterlibatan seperti user_mention_UT, *in_reply_to*, merupakan salah satu strategi agar informasi dibagikan ulang. Keterkaitan dalam aktivitas konten yang dibagikan seperti *foto_UT, video_UT, In_reply_to, quoted_UT, average_tweet_per_day ,Retweet_UT, dan Hashtag_UT* akan diperiksa. Popularitas dan keberhasilan persebaran konten yang tercatat pada *Retweet_count_UT, Retweet_100, Retweet_1000, dan Favourite_UT* akan dimanfaatkan sebagai fitur prediksi.

D. Model Pembelajaran Mesin

Permasalahan prediksi akan direpresentasikan dengan jumlah *retweet* lebih dari 1 oleh sekumpulan fitur dan kemudian digunakan pengklasifikasi pembelajaran mesin untuk memprediksi pertumbuhan *retweet*. Metode pembelajaran, termasuk regresi logistik, SVM, dan *random forest* dengan library sklearn akan digunakan. Sebelum

dilakukan pelatihan data dilakukan normalisasi dengan *Min Max Scaler*.

Hiperparameter yang ditetapkan dalam konstruktor penduga menentukan perilaku algoritma pembelajaran dan oleh karena itu kinerja model yang dihasilkan pada data yang tidak terlihat. Masalah pemilihan model, oleh karena itu, untuk menemukan, dalam beberapa ruang hyper-parameter, kombinasi terbaik dari *hiperparameter*, sehubungan dengan beberapa kriteria yang ditentukan pengguna. *Pada random forest* dengan nilai yang terlalu kecil untuk parameter kedalaman pohon maksimal akan cenderung kurang pas, sementara nilai yang terlalu besar akan membuatnya terlalu pas.

Pemilihan model hiper parameter akan digunakan metode GridSearchCV. Variabel akan dimasukan pada *estimator* (dasar atau komposit), dengan *hiperparameter* telah dioptimalkan menjadi satu set pengaturan *hiperparameter*. Kumpulan ini direpresentasikan sebagai pemetaan nama parameter ke sekumpulan pilihan diskrit dalam kasus penelusuran, yang secara lengkap disebut *grid* produk dari kombinasi parameter lengkap. Hasil pencarian *hiperparameter* tertuang pada Tabel V.

TABEL V
HIPERPARAMETER TERBAIK SETELAH PENYELARASAN DENGAN GRID SEARCH.

No	Model	Pemasangan Hiperparameter
1	Logistic Regression	Penalty =l2 | C=3792.690190732246
2	Support Vector Machine	C= 1000 | gamma= 0.01 | kernel= 'sigmoid'
3	Random Forest	n_estimators= 1000 | min_samples_split= 5 | min_samples_leaf= 4 | max_features= 'auto | max_depth= 100 | bootstrap= True

Kinerja dilaporkan pengklasifikasi *random forest* terutama dalam perbandingan antara fitur linimasa dengan fitur konten dan fitur akun. Dalam banyak percobaan yang dilaporkan, kinerja pengklasifikasi nonlinear *random forest* biasanya berkinerja lebih baik daripada pengklasifikasi linear seperti regresi logistik. Digunakan metode *grid search* untuk setiap metode pelatihan yang tercatat pada Tabel V. untuk mendapatkan hyperparameter yang memiliki performa akurasi paling baik. Dalam semua kasus, dilakukan validasi silang 10 kali lipat dan melaporkan akurasi klasifikasi, nilai *recall*, skor F1, dan area di bawah kurva AUC.

E. Pencarian Fitur Penting

Untuk memahami kontribusi keseluruhan dari setiap fitur individu untuk akurasi prediksi, akan digunakan metode *recursive feature elimination* (RFE) pada regresi logistik, *mean decrease impurity (*MDI) dan *permutation importance* pada *random forest*.

RFE adalah metode pemilihan fitur yang sesuai dengan model dan menghilangkan fitur terlemah (atau fitur) hingga jumlah fitur yang ditentukan tercapai. Fitur diberi peringkat berdasarkan atribut koefisien model, dan dengan secara rekursif menghilangkan sejumlah kecil fitur pada setiap putaran, RFE mencoba menghilangkan dependensi dan collinearity yang mungkin ada dalam model.

RFE memerlukan sejumlah fitur tertentu untuk dipertahankan, namun sering kali tidak diketahui sebelumnya berapa banyak fitur yang valid. Untuk menemukan jumlah optimal fitur, validasi silang digunakan dengan RFE untuk menilai subset fitur yang berbeda dan memilih koleksi fitur yang paling baik. Dipilih 10 fitur yang terbaik setelah dipangkas fitur lainya.

Permutation importance dan MDI, dapat digunakan untuk dibangun peringkat kepentingan variabel dan pemilihan variabel. *Permutation importance* mengukur pentingnya variabel dengan mengukur perubahan dalam akurasi prediksi, ketika nilai variabel secara acak dibandingkan dengan pengamatan asli.

Pada permutation importance model m akan dijalankan dengan dataset D. Skor validasi akan dihitung pada model klasifikasi m pada setiap fitur j pada kolom pada D dengan pengulangan sebanyak K. Pada setiap penghitungan data secara acak akan menghasilkan data Dkj yang melakukan komputasi ckj. Dengan ij sebagai permutation importance maka akan dihitung dengan rumus:

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \quad [1]$$

MDI adalah jumlah dari semua penurunan *impurity* Gini karena variabel tertentu untuk membentuk perpecahan di *random forest*, yang akan dinormalisasi dengan jumlah pohon. MDI berasal dari teori informasi dan mengukur seberapa banyak variabel acak X informatif tentang variabel acak lain Y. Hal ini berkaitan erat dengan konsep entropi. Entropi variabel acak X, yang secara tradisional dilambangkan dengan H (X), mengukur tingkat ketidakpastian dalam variabel X. Ini dihitung sebagai rumus:

$$H(X) = -\sum_{x} P_X(x)\log P_X(x) \quad [2]$$

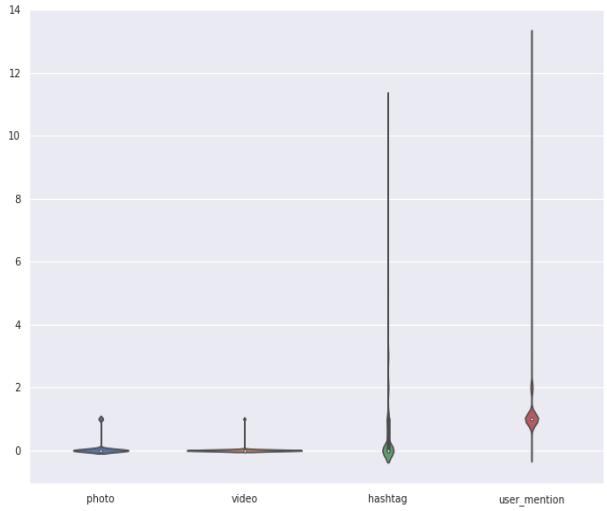
di mana PX (x) adalah distribusi probabilitas dari X. Entropi bersyarat H (X | Y) mengukur rata-rata ketidakpastian di X dengan memperhatikan variabel Y. Kemudian MI (X, Y) didefinisikan sebagai penurunan ketidakpastian tentang X setelah mengamati Y.

IV. HASIL DAN PEMBAHASAN

A. Data Eksplorasi

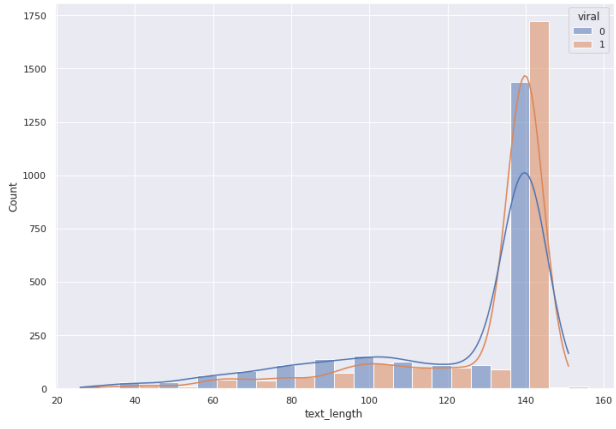
Data eksplorasi merupakan tahap awal dalam melakukan pengolahan data dimana pada tahap ini dilakukan eksplorasi

mendalam terhadap penyebaran informasi di Twitter untuk mendapatkan informasi baru apa saja yang dapat digunakan untuk tahap selanjutnya. Analisis data eksploratif merupakan metode eksplorasi data dengan menggunakan teknik aritmatika sederhana dan teknik grafis dalam meringkas data pengamatan. Pada tahap ini akan dibagi dengan tiga fitur utama yaitu diamati dalam fitur konten, fitur akun dan fitur linimasa.



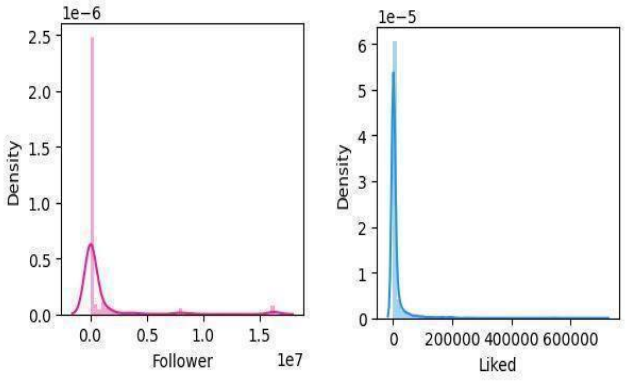
Gambar 3. Eksplorasi fitur konten dengan melihat boxplot pada foto,video, *hashtag* dan *user_mention*

Fitur konten seperti yang tertera pada Gambar 3 ditelusuri sebagai Pada Gambar x melihat frekuensi pengguna dalam menggunakan foto, video, *hashtag* dan *user_mention.* Penggunaan foto dan video ditemukan lebih jarang digunakan dan hal ini menunjukan bahwa text atau tulisan lebih sering digunakan dibandingkan fitur media. Sedangkan fitur yang paling sering digunakan adalah *user_mention* dan dengan rata-rata setiap post terdapat satu. Sedangkan penggunaan hashtag atau tagar juga belum selalu digunakan.



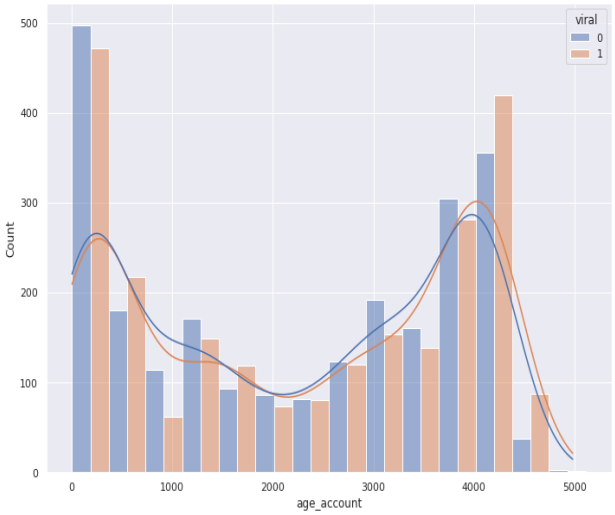
Gambar 4. Melihat korelasi antara panjang konten teks dengan perluasan penyebaran konten

Berikutnya pada Gambar 4, dapat dilihat penggunaan teks pada twitter kebanyakan pada panjang teks 140 karakter. Teks pada karakter 40 menunjukan kecenderungan untuk dapat menyebarkan informasi dibandingkan dengan panjang text yang lebih sedikit.



Gambar 5. Distribusi dari Jumlah disukai dengan *follower*

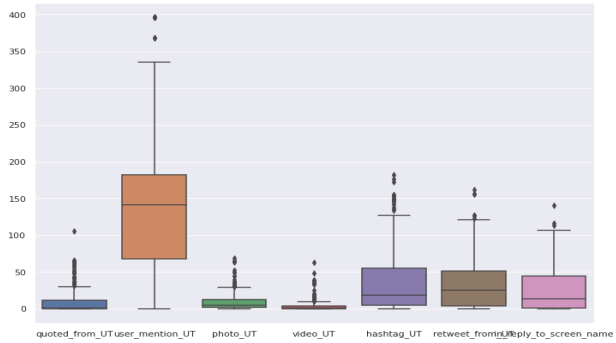
Eksplorasi pada fitur akun pada Gambar 5 menunjukan bagaimana aktivitas seorang pengguna dan status pertemanannya. Hal yang berlaku pada pada banyak jaringan media sosial adalah terbentuk dari jaringan *scale-free*, di mana seseorang yang memiliki pengikut lebih banyak akan menjadi lebih menarik untuk diikuti. Hal ini mengakibatkan bentuk pertemanan seperti pada Gambar yang diperlihatkan bahwa kebanyakan orang memiliki *follower* dan *like* yang sedikit sedangkan segelintir orang akan mendapatkan jumlah *follower* dan *like* yang sangat banyak.



Gambar 6. Korelasi antara umur akun dengan penyebaran informasi pada fitur akun

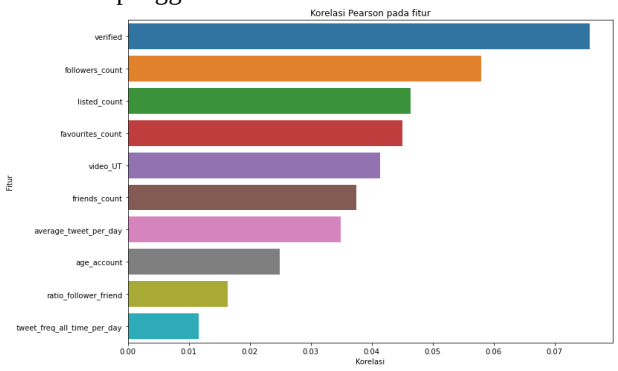
Pengguna yang terdapat pada dataset kali ini memiliki dua puncak seperti ilustrasi pada Gambar 6 dan dapat dikategorikan sebagai pengguna baru mulai pada tahun pertamanya dan pengguna lama yang telah menggunakan Twitter hingga 10 tahun. Pada data eksplorasi ditemukan bahwa pengguna lama lebih memiliki kecenderungan untuk dapat melakukan propagasi informasi.

p-ISSN : 2443-2210
e-ISSN : 2443-2229



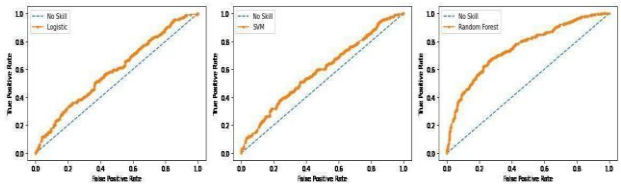
Gambar 7. Plotbox dalam eksplorasi fitur linimasa

Pada fitur linimasa Gambaran perilaku pengguna dapat terGambar di Gambar 7. Penggunaan *user_mention_UT* pada pengguna menunjukan bahwa dalam text yang dituliskan juga mencantumkan pengguna lain untuk menambahkan keterlibatan. Strategi keterlibatan dengan mention bahkan dapat meningkatkan peningkatan popularitas suatu konten. Penggunaan fitur tagar, *retweet*, *in_reply_to screenname* adalah tiga fitur lainya yang cukup sering digunakan sedangkan foto dan video jarang dicantumkan pengguna.



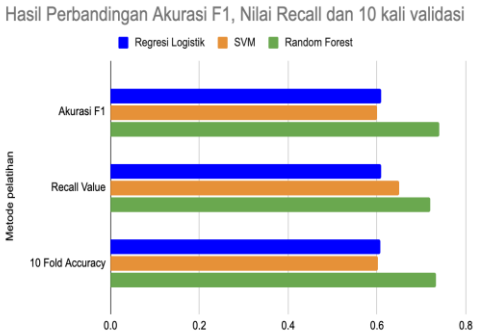
Gambar 8. Peringkat korelasi tertinggi dengan pengaruh penyebaran informasi dengan nilai pearson

Berikut adalah hasil korelasi antar poin data. Korelasi tersebut akan dibandingkan dengan prediksi viralitas pada Gambar 8 dan 9 . Ditemukan korelasi secara statistik belum bermakna signifikan dengan nilai *pearson* karena nilainya masih dibawah 0.1.



Gambar 9. Kurva AUC pada pelatihan dengan regresi logistik, SVM dan random forest

Pembelajaran mesin baik random forest, regresi logistik dan SVM dapat memprediksi penyebaran informasi. Dengan menggunakan model yang tertera pada Tabel V didapatkan hasil pada Tabel VI. Kita dapat melihat bahwa performa *random forest* lebih tinggi dari metode lain seperti yang tercantum pada gambar 10.

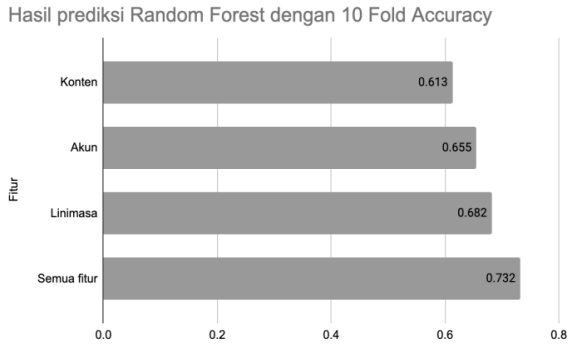


Gambar 10. Perbandingan nilai akurasi f1, nilai recall dan validasi sepuluh kali dengan pelatihan regresi logistik, SVM dan random forest

Persentase akurasi F1, recall value dan dengan validasi 10 kali berardi di angka 70%. Sedangkan dengan metode lainya hanya berada di sekitar 60%. Hal ini juga berkesesuaian dengan penelitian Bunyamin[8] menemukan tingkat akurasi lebih tinggi dengan *random forest*.

Pada grafik *area under curve* (AUC) dapat dilihat performa dari SVM dan Regresi Logistik berada didekat dengan kontrol hal ini menunjukan performa yang kurang akurat. Sedangkan dengan *random forest* AUC memiliki performa AUC yang terlihat pada Gambar 9 yang lebih baik dibandingkan kedua algoritma lainnya.

Random forest membangun sekitar 1000 pohon *decision tree*. Model akan memprediksi nilai dengan mempelajari keputusan yang disimpulkan dari fitur data. Setiap keputusan pada sebuah node dibuat dengan klasifikasi menggunakan fitur tunggal. Dengan membangun *estimator* hingga 1000, akan divisualisasikan dengan satu pohon seperti yang tertera pada Gambar 12 untuk memberikan intuisi model pembelajaran.



Gambar 11. Perbandingan nilai akurasi f1, nilai recall dan validasi sepuluh kali dengan pelatihan regresi logistik, svm dan random forest

Agregasi Bootstrap, adalah metode ensemble yang sederhana dan sangat kuat. Bagging adalah penerapan prosedur Bootstrap ke algoritma pembelajaran mesin varian

tinggi, pada random forest digunakan pohon *decision tree* , seperti salah satu contoh pohon pada Gambar 13.

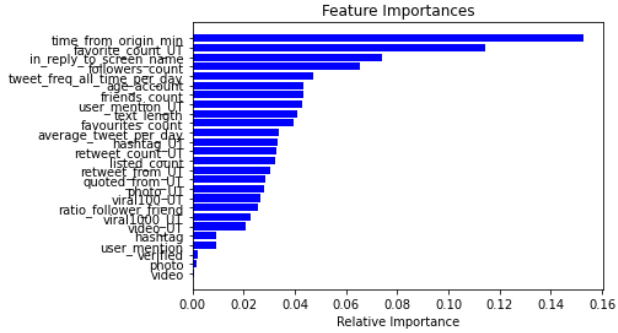
Untuk setiap sampel bootstrap yang diambil dari data pelatihan, akan ada sampel tertinggal yang tidak disertakan. Sampel ini disebut sampel *Out-Of-Bag* atau OOB. Performa setiap model pada sampel yang tersisa saat di rata-ratakan dapat memberikan perkiraan akurasi model yang dikantongi. Estimasi kinerja ini sering disebut dengan estimasi kinerja OOB. Ukuran kinerja ini adalah estimasi kesalahan pengujian yang andal dan berkorelasi baik dengan estimasi validasi silang. Hal ini yang menjadi kelebihan *random forest* pada data yang non linier yang ditemukan pada Twitter.

Pemanfaatan fitur konten, akun dan linimasa akan dibandingkan performanya. Hasil ini memberi indikasi bahwa fitur linimasa dapat menjadi indikasi apakah konten dapat menjadi prediktor yang baik. Hasil akurasi dapat mencapai 0.682 dengan menggunakan fitur linimasa dan lebih baik dibandingkan dengan hanya menggunakan fitur konten ataupun akun.

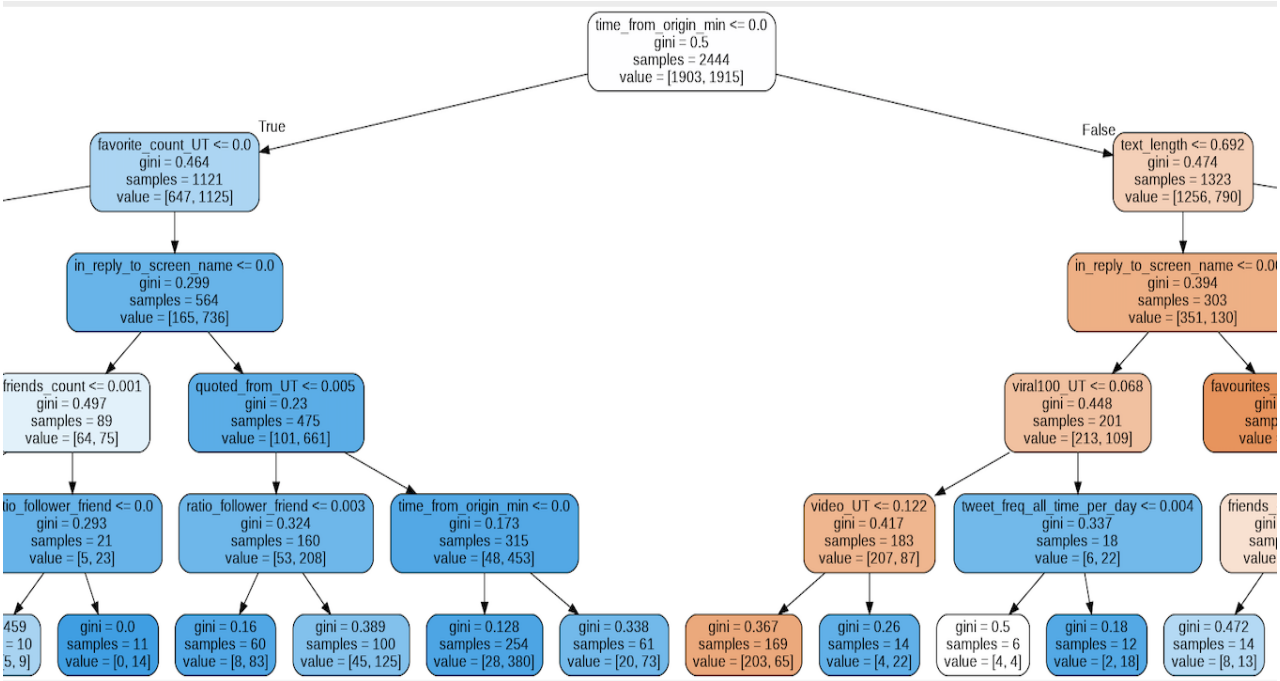
Gambar 11 menunjukan pada pemanfaatan fitur konten, akun dan linimasa sebagai prediktor penyebaran informasi akan dibandingkan performanya. Hasil ini memberi indikasi bahwa fitur linimasa dapat menjadi indikasi apakah konten dapat menjadi prediktor yang baik. Hasil akurasi dapat mencapai 0.682 dengan menggunakan fitur linimasa dan lebih baik dibandingkan dengan hanya menggunakan fitur konten ataupun akun dengan angka fitur konten 0.613 dan fitur akun 0.655.

B. Pencarian fitur penting

Kalkulasi dengan MDI menunjukkan nilai seperti pada Gambar 12. Pada test fitur yang dilakukan ditemukan 10 fitur utama yang dihitung melalui MDI dan *permutation importance* pada Tabel VI , sama dengan urutan yang berbeda. Salah satu fitur yang paling dominan adalah dengan fitur konten yaitu berapa menit waktu setelah penyiar pertama ke penyiaran ulang dengan *retweet*. Korelasinya adalah semakin lama sebuah konten beredar di Twitter maka semakin rendah juga kemampuan replikasi. Hal ini juga ditemukan di penelitian Ma[16], umur optimum *tweet* adalah pada jam ke 24 dan berangsur angsur menurun.

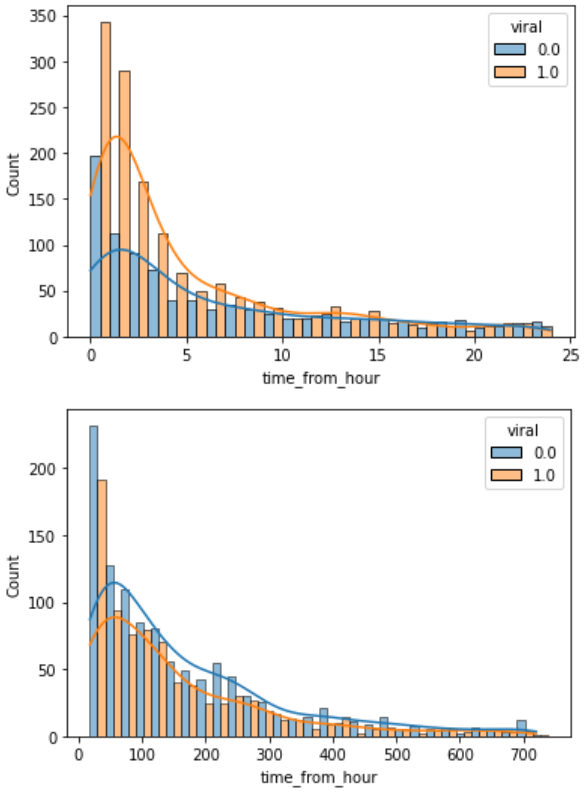


Gambar 12. Tabel perhitungan *feature importance* dengan kalkulasi MDI



Gambar 13. Salah satu contoh gambar *decision tree* pada pelatihan random forest dengan 1000 *estimat*

p-ISSN : 2443-2210
e-ISSN : 2443-2229



Gambar 14. Perbandingan histogram pada Gambar diatas dimana kecenderungan penyiaran lebih berkembang dibandingkan 24 jam berikutnya pada Gambar di bawah.

Seperti yang diGambarkan histogram frekuensi Gambar 14, penyiaran pada 15 jam pertama lebih cenderung menyebar luas dibandingkan jam berikutnya. Lalu setelah 24 jam pertama sebuah konten disiarkan maka memiliki kecenderungan *tweet* tidak berkembang secara luas.

Hasil pada urutan fitur penting pada MDI (Tabel VI) pada urutan 4 dan 7 dan permutation importance pada urutan 5 dan 9 pada sesuai dengan Petrovic[12] mereka bahwa jumlah follower dan friend memiliki hubungan yang sangat kuat dengan prediktor penyebaran informasi tersebut. Jumlah *tweet* yang disukai ditemukan juga sebagai prediktor yang baik dan juga ditemukan hal yang sama pada penelitiannya. Fitur frekuensi *tweet* juga ditemukan sebagai prediktor yang baik bersama dengan itur umur akun dan panjang text[8].

TABEL VI
HASIL KALKULASI MDI DAN FEATURE IMPORTANCE PADA PELATIHAN RANDOM FOREST

Urutan	Fitur Penting MDI	Kelas Fitur	Fitur Feature Importance	Kelas Fitur
1	Time from Origin	Konten	Time from Origin	Konten
2	favourites_countUT	Linimasa	favourites_countUT	Linimasa
3	In_reply_to_screen	Linimasa	In_reply_to_screen	Linimasa
4	Follower_Count	Akun	Text_Length	Konten
5	Tweet_Freq	Akun	Follower_Count	Akun
6	User_Mention_UT	Linimasa	Tweet_Freq	Akun
7	Friend_Count	Akun	Age_account	Akun
8	Age_account	Akun	User_Mention_UT	Linimasa
9	Favourite_count	Akun	Friend_Count	Akun
10	Text_Length	Konten	Favourite_count	Akun

Pada fitur linimasa ditemukan dari bagaimana popularitas pengguna, isi konten pengguna dan aktivitas engagement pengguna. Pada fitur *viral_100UT*, *viral_1000UT*, dan jumlah favorit menunjukan bagaimana popularitas konten pengguna sebagai prediktor yang dapat digunakan untuk penyiaran konten pada masa depan. Penggunaan video pada fitur konten belum ditemukan korelasi namun jika dilihat pada linimasa, penggunaan konten video pada linimasa pengguna ditemukan sebagai fitur yang dipertahankan pada RFE di Tabel VII.

TABEL VII
HASIL KALKULASI RFE PADA PELATIHAN REGRESI LOGISTIK

Fitur Penting	Kelas Fitur
time_from_origin_min	Konten
video_UT	Linimasa
viral100_UT	Linimasa
viral1000_UT	Linimasa
user_mention	Konten
friends_count	Akun
listed_count	Akun
favourites_count	Akun
*tweet*_freq_all_time_per_day	Akun
ratio_follower_friend	Akun

Pengguna yang rajin keterlibatan yang juga menjadi strategi *engagement* dapat dikatakan mempengaruhi penyebarluasan informasi[5][6]. Dalam Tabel I dan II menunjukan aktivitas menjawab *tweet* orang lain dan melakukan menyebut akun lain menjadi prediktor penyebaran ulang konten di masa depan.

V. SIMPULAN

Dalam mencari metode mesin pembelajaran yang diuji dengan tiga metode pelatihan yaitu regresi logistik, SVM dan *random forest* ditemukan memiliki performa yang paling baik dalam memprediksi konten yang tersebar luas adalah *random forest*. Kedua, fitur linimasa dapat dimanfaatkan sebagai prediktor yang baik, dalam perbandingan dengan fitur konten dan fitur akun yang telah diujikan menunjukan performa akurasi yang lebih baik. Terakhir dalam pencarian fitur paling penting ditemukan bahwa waktu pertama kali disiarkan menjadi fitur prediksi terbaik dalam *time_from_origin*. Pada fitur linimasa ditemukan tercatat 6 fitur terpenting yaitu *favourites_countUT* ,*video_UT*, *viral100_UT*, *viral1000_UT*, *in_reply_to_screen* dan *user_mention_UT* dan pada fitur akun, *follower_count, age_account, favourites_count, tweet_freq_all_time_per_day, ratio_follower_friends_ friends_count* dan pada konten yaitu *user_mention dan text_length*.

Pada penelitian selanjutnya, kita akan melakukan ekstraksi konten pada linimasa dan membandingkan dengan konten terbaru dengan melakukan *topic modelling* untuk meningkatkan akurasi pada prediksi.

VI. DAFTAR PUSTAKA

[1] Broersma, Marcel, and Todd Graham.,"Twitter as a news source: How Dutch and British newspapers used *tweet*s in their news coverage, 2007–2011," *Journalism practice* 7,4: 446-464, 2013.

[2] Tsur, Oren, and Ari Rappoport, "What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities," *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012.

[3] Pramanik, Soumajit, et al, "Modeling cascade formation in Twitter amidst mentions and *retweet*s," *Social Network Analysis and Mining* 7,1: 41. 2017

[4] Suh, B, Hong, L, Pirolli, P, and Chi, E. H., "Want to be *retweet*ed? Large scale analytics on factors impacting *retweet* in twitter network," *In Social Computing (SocialCom),IEEE Second International Conference on*, 177–184, 2010.

[5] Sundstrom, Beth, and Abbey Blake Levenshus, "The art of engagement: Dialogic strategies on Twitter," *Journal of Communication Management*, 2017.

[6] Ashley, C, & Tuten, T, "Creative Strategies in Social Media Marketing: An Exploratory Study of Branded Social Content and Consumer Engagement" .*Psychology & Marketing, 32(1), 15-27* 2017.

[7] Xu, Zhiheng, and Qing Yang "Analyzing user *retweet* behavior on twitter" *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 2012.

[8] Bunyamin, Hendra, and Tomas Tunys "A comparison of *retweet* prediction approaches: the superiority of random forest learning method" *Telkonika (Telecommun Comput Electron Control)* 14.3: 1052-1058 2016.

[9] Cheng, Justin, et al "Can cascades be predicted?" *Proceedings of the 23rd international conference on World wide web* 2014.

[10] Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn "Virality prediction and community structure in social networks" *Scientific reports 3: 2522* 2013.

[11] Yang, Zi, et al "Understanding *retweet*ing behaviors in social networks" *Proceedings of the 19th ACM international conference on Information and knowledge management* 2010.

[12] Petrovic, Sasa, Miles Osborne, and Victor Lavrenko "Rt to win! predicting message propagation in twitter" Proceedings of the *International AAAI Conference on Web and Social Media* Vol. 5. No. 1 2011.

[13] M. Salganik, P. Dodds, and D. Watts "Experimental study of inequality and unpredictability in an artificial cultural market," *science, 311(5762), 854-856*, 2006.

[14] Dow, P. Alex, Lada Adamic, and Adrien Friggeri, "The anatomy of large facebook cascades," Proceedings of the International *AAAI Conference on Web and Social Media*, Vol. 7. No. 1, 2013.

[15] Aparicio, Sofía, Javier Villazón-Terrazas, and Gonzalo Álvarez, "A model for scale-free networks: application to twitter," *Entropy* 17,8: 5848-5867,2015.

[16] Ma, Zongyang, Aixin Sun, and Gao Cong, "On predicting the popularity of newly emerging hashtags in Twitter," *Journal of the American Society for Information Science and Technology* 64.7: 1399-1410, 2013.

[17] Minnich, Amanda, et al, "BotWalk: Efficient adaptive exploration of Twitter bot networks," *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* , 2017.

[18] Reece, Andrew G., et al, "Forecasting the onset and course of mental illness with Twitter data," *Scientific reports* 7.1: 1-11, 2017.

[19] Chaudhry, Prateek, and Matthew Lease, "You Are What You *Tweet*: Profiling Users by Past *Tweet*s to Improve Hate Speech Detection," *arXiv preprint arXiv*:2012.09090,2020.

[20] Hoang, Thi Bich Ngoc, and Josiane Mothe, "Predicting information diffusion on Twitter–Analysis of predictive features," *Journal of computational science* 28: 257-264, 2018.