OPEN ACCESS
CENTIVE 2019
ISBN : 978-602-53004-0-0

Conference on Electrical Engineering, Telematics, Industrial Technology, and Creative Media 2019

Analisa Relevansi Tweet terhadap Hashtag dengan Metode Logistic Regression

Muhammad Yusril Aldean[1], Muhammad David Hilmawan[2], Rini Indriyati[3], Jerry Lasama[4], Apri Junaidi[5]

Fakultas Teknologi Industri dan Informatika, Institut Teknologi Telkom Purwokerto
Jl. DI Panjaitan No 128 Purwokerto 53147 Indonesia

[1] 118102062@ittelkom-pwt.ac.id
[2]18102023@ittelkom-pwt.ac.id
[3]19102060@ittelkom-pwt.ac.id
[4]18102018@ittelkom-pwt.ac.id
[5]apri@ittelkom-pwt.ac.id

Abstrak

Twitter sering kali digunakan sebagai sumber data untuk penelitian natural language processing, namun ada banyak sekali tweet yang tidak relevan pada topik yang dibicarakan, tweet yang tidak relevan itu seringkali membuat set data menjadi terkontaminasi sehingga dapat mempengaruhi kualitas dari hasil penelitian, dan perlu dibersihkan secara manual. Model yang diusulkan ini menggunakan tweet sebagai input untuk mengklasifikasikan tweet yang relevan atau tidak dengan topik yang sedang dibicarakan, metode yang digunakan adalah mengubah tweet tersebut dari sebuah kalimat menjadi sebuah bentuk data yang berisikan angka yang kemudian dimasukan kedalam sebuah bentuk matriks dan diproses dengan menggunakan metode logistic regression, dari hasil prediksi tersebut menunjukkan bahwa hasil akurasi model yang telah dibuat ini berada diatas angka 70%.

Kata kunci: *twitter, klasifikasi, logistic regression,prediksi*

I. PENDAHULUAN

TWITTERadalah salah satu media sosial yang populer dan banyak digunakan pada saat ini. Twitter menempati peringkat kedua sebagai media sosial teraktif di Indonesia[1].

Twitter merupakan media sosial *microblog* yang memungkinkan pengguna untuk mengirimkan pesan yang dibatasi hingga 280 karakter[2]. Twitter digunakan oleh semua orang untuk melakukan penilaian dan mengeluarkan opini mengenai segala sesuatu, melakukan posting dan rating dengan opini yang berbeda–beda[3]. Dalam sistem Twitter, tanda # atau *hashtag* menunjukkan topik-topik khusus yang sedang dibahas. Fungsi *hashtag* dalam Twitter antara lain sebagai media pencarian dan menampilkan informasi lebih mudah, dan sebagai penanda topik yang sedang ramai atau *trend*[4].

Kemudahan dan cepatnya perolehan informasi membuat topik pada *tweet* sering kali disalahgunakan dan membuat *tweet* menjadi tidak relevan dengan *hashtag* yang dimaksud, penelitian ini bertujuan membuat sebuah model untuk mengklasifikasikan relevansi antara *tweet* terhadap *hashtag* sehingga data yang didapat menjadi lebihbersih.

II. METODE PENELITIAN

Dalam penelitian ini, digunakan metode seperti Gambar1 yang terdiri dari : *Twitter scraping,labeling, TF-IDF Tokenizer* dan diprediksimenggunakan algoritme *logistic regression.*

Gambar 1. Alur metode penelitian

A. *Twitter Scraping*

Scraping merupakan teknik yang seringdigunakan untuk mengambil data atau konten dari web. pengembang Facebook dan Twitter menyediakan API (*Application Program Interface)* yang menyanggupkan situs web untuk mengakses informasi dari situs web media sosial[5]. Program yang digunakan untuk *scraping* data dari Twitter disebut dengan *Twitter Scraper*.

Twitter Scraper digunakan untuk pengambilanserta pengumpulan data dari Twitter[6]. *Hashtag* yang digunakan pada penelitian adalah: #jokowimundur, #kkndesapenari, #ambilpositifnya,#novemberrain,#reformasidikorupsi,#festivalsekaten2019, #stmbergerak, #TangkapAdeArmando, dan #bitcoin. *Hashtag* yang digunakan berasal dari pengeposan tahun 2010 hingga tahun 2019 untuk memaksimalkan variasi dari data agar hasil prediksi tidak terlalu condong ke salah satu *hashtag* saja. Dari hasil *scraping* yang dilakukan, didapatkan total 2883 data dari *hashtag* yang sudah ditentukan.

B. *Labeling*

Data yang sudah diperoleh dari proses *scraping* dilakukan labelisasi dengan dibuatkan sebuah kolom baru pada data untuk relevansi data dimana kolom tersebut diisi dengan nilai 0 atau 1 secara manual. Nilai 0 menandakan bahwa *tweet* tidak relevan dengan *hashtag* pada *tweet* tersebut, sedangkan nilai 1 menandakan yang sebaliknya dimana *tweet* relevan dengan *hashtag* pada *tweet* tersebut.

C. *TF-IDF Tokenizer*

Proses tokenisasi merupakan sebuah proses untuk memisahkan setiap kata yang tersusun dalam sebuah dokumen. Bagian yang dihilangkan adalahkarakter yang selain huruf alfabet, dikarenakan karakter tersebut tidak berpengaruh dalam premrosesan suatu teks[7]. Token kemudian dimasukan kedalam sebuah variabel

dan diubah menjadi sebuah *Tensor* berisikan angka yang diproses melalui sebuah algortime. Algoritme yang digunakan pada *Tokenizer* adalah TF-IDF.

Metode TF-IDF menghitung bobot setiap kata yang digunakan pada pengambilan informasi. Metode ini terkenal efisien, sederhana, dan memiliki akurasi yang tinggi[8]. TF-IDF merupakan statistik numerik yang mencerminkan betapa pentingnya sebuah kata bagi sebuah dokumen dalam koleksi atau korpus[9].

TF-IDF digunakan untuk menyingkirkan istilah yang berbobot rendah dari suatu dokumen danmembantu untuk meninggikan efektivitas pengambilan data. peningkatan dari nilai TF-IDF berbanding lurus dengan banyaknya suatu kata yang muncul di dalam dokumen, tetapi dinetralisasi oleh frekuensi dari kata yang ada didalam korpus, dimana hal itu membantu menyeimbangkan kata mana yang lebih sering muncul secara umum[10]. Nilai pembobotan TF-IDF ($W_{t,d}$) dilakukan dengan munggunakan persamaan :

$$W_{t,d} = (TF_{t,d}) \times log_{10}\left(\frac{N}{DF_t}\right) \quad (1)$$

dimana$TF_{t,d}$ mengacu pada frekuensi istilah *t* muncul dalam dokumen *d*, *N* adalah banyaknya dokumen yang ada di set data dan DF_tmerupakan banyaknya dokumen yang mengandung istilah *i* [11].

D. Logistic Regression

Logistic regression merupakan salah satualgoritme yang paling sering digunakan untuk klasifikasi[12]. *Logistic regression* adalah metode standar untuk memperkirakan rasio kemungkinan yang telah disesuaikan[13]. Model *logisticregression* banyak digunakan untuk menyelidikipengaruh independen suatu variabel terhadap hasil binominal dalam literatur medis[14].

Logistic regression mencari model yang palingsesuai dan hemat untuk mendeskripsikan hubungan antara hasil dari fungsi logit dan sebuah himpunan variabel penjelas independen (*p*)[15]:

$$logit(Y) = ln\left[\frac{\pi}{1-\pi}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad (2)$$

dimana fungsi logit adalah adalah logaritma natural dari peluang dari *Y*, peluang tersebut adalah rasio dari probabilitas (π) *Y* yang terjadi ke probabilitas (1 – π) *Y* tidak terjadi, *α* adalah intersep *Y*, *X* adalah variabelprediktor input, dan *β* merupakan koefisien regresi. Mengambil antilog dari (2), persamaan untuk memprediksi probabilitas terjadinya hasil p adalah[16] :

$$\pi(Y|X) = \frac{\exp(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}{1+\exp(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)} \quad (3)$$

E. Confusion Matrix

Confusion Matrix meringkas kinerja modeldalam mengklasifikasi suatu data uji. *ConfusionMatrix* adalah matriks dua dimensi, yang diindeks kedalam satu dimensi oleh nilai benar kelas sebenarnya dari suatu objek yang di prediksi[17]. Berikut merupakan contoh dari *confusion matrix* 2 dimensi:

TABEL I
CONTOH CONFUSION MATRIX

Confusion Matrix	*Predicted Positive*	*Predicted Negative*
Actual Positive	TP	FN

Actual Negative	FP	TN

F. Precision and Recall F1-Score

Precision dan *recall* didefinisikan seperti didalam *machine learning* dimana *precision* adalah rasio antara *true prositives* dan *predicted positives* dan *recall* adalah rasio antara *true positives* dan*actual positives*. F1-Score parametermenggabungkan *precision* dan *recall* sebagai berikut[18]:

$$F1 - Score = 2.\frac{Precision\ .Recall}{Precision + Recall} \quad (4)$$

III. HASIL PENELITIAN

Berikut merupakan parameter yang digunakan dalam penelitian:

TABEL II
PARAMETER

Parameter	Nilai
Penalty	L2
Tol	0.0001
C	1.0
Max_iter	100

Berdasarkan model yang telah dibuat, didapatkan hasil seperti berikut:

TABEL III
CONFUSION MATRIX

Kelas	Terklasifikasi Tidak Relevan	Terklasifikasi Relevan
Tidak Relevan	98	154
Relevan	213	976

Setelah mendapatkan data tabel hasil, maka didapatkan akurasi dari model sebesar 74,53%. Laluhasil dari *Precision, Recall* dan F1-*Score* adalah seperti berikut:

TABEL IV
PRECISION, RECALL DAN F1 SCORE

Kelas	*Precision*	*Recall*	F1	*Samples*
TidakRelevan	0.3151	0.3888	0.3481	252
Relevan	0.8637	0.8208	0.8417	1189
Rerata	0.5891	0.6048	0.5949	720.5

IV. PEMBAHASAN

Hal pertama yang dilakukan adalah pengumpulan data menggunakan *twitter scraper*, data terkumpul sebanyak 2883 data.

TABEL IV
SET DATA PERTAMA

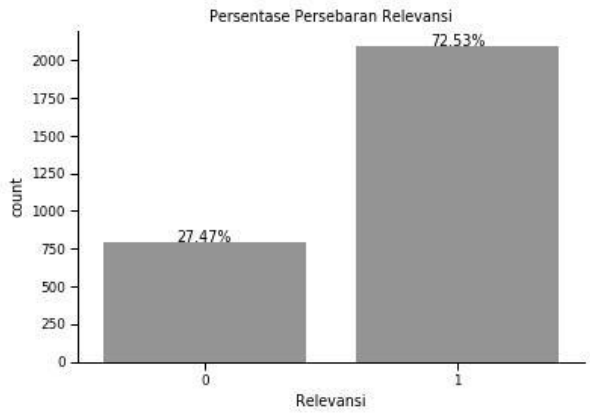
No	Timestamp	Tanggal	Bulan	Tahun	Tweet
1	15/09/2019	15	Oct	2019	Apa nih mau nyaingin #KKNDesaPenari
2	25/08/2015	25	Aug	2019	#jokowimundur
3	26/10/2019	26	Oct	2019	Jika menang nanti November benar hujan
4	04/10/2019	04	Oct	2019	Kalo rakyat sudah turun
5	31/10/2019	31	Oct	2019	Hey hey hey…

Kemudian, pada tabel set data dibuatkan sebuah kolom baru pada data untuk relevansi data yang diberi nama relevansi dimana kolom tersebut diisi dengan nilai 0 atau 1 secara manual. Nilai 0 menandakan bahwa *tweet* tidak relevan dengan *hashtag* pada *tweet* tersebut, sedangkan nilai 1menandakan yang sebaliknya dimana *tweet* relevan dengan *hashtag* pada *tweet* tersebut. Seperti tabel berikut :

TABEL IV
SET DATA RELEVANSI

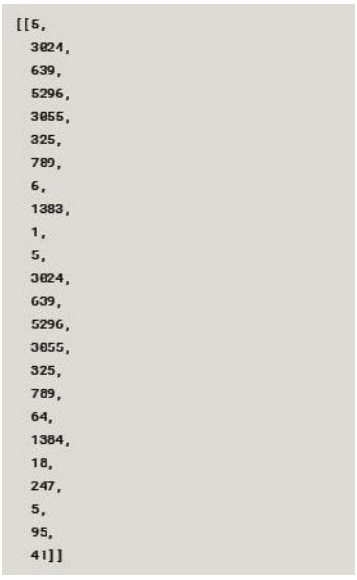
No	Timestamp	Tanggal	Bulan	Tahun	Relevansi	Tweet
1	15/09/2019	15	09	2019	0	Apa nih mau nyaingin #KKNDesaPenari
2	25/08/2015	25	08	2019	1	#jokowimundur
3	26/10/2019	26	10	2019	1	Jika menang nanti November benar hujan
4	04/10/2019	04	10	2019	1	Kalo rakyat sudah turun
5	31/10/2019	31	10	2019	0	Hey hey hey…

Setelah memasukan data pada kolom relevansi didapatkan hasil seperti berikut :



Gambar 2. Presentase Penyebaran Relevansi

Didapatkan data sesuai diagram diatas, menyatakan bahwa 72,53% dinyatakan relevan dan 27,47% tidak relevan. Data terlihat tidak seimbang dimana data yang relevan jauh lebih tinggi dibandingkan dengan data yang tidak relevan. Data yang tersedia diatas tidak lupa untuk dibagi menjadi 2, 50% data digunakan sebagai *Data Train* dan 50% digunakan untuk *Data Test*. Data yang sudah didapatkan lalu dibentuk menjadi sebuah *sequences* menggunakan *tokenizer* dan berubah menjadi seperti berikut :

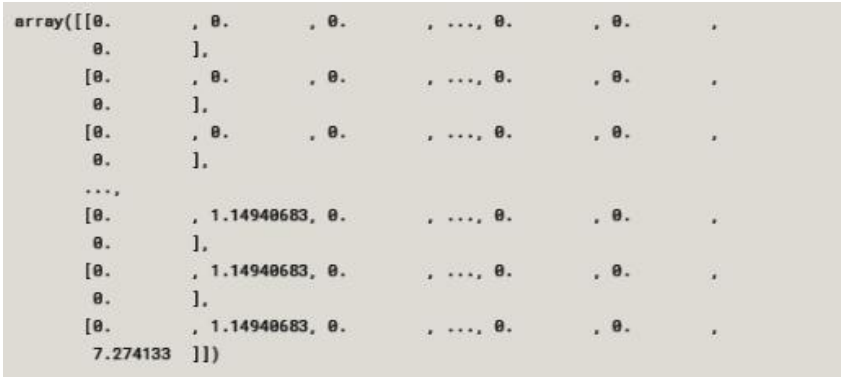


Gambar 3. Data dalam bentuk *sequence*

Setelah data tersebut berbentuk menjadi sebuah *sequence*, selanjutnya dibuat menjadi sebuahbentukan data berbentuk menjadi sebuah *array* menggunakan perhitungan TF-IDF seperti yang sudah dijelaskan pada bagian bab sebelumnya, ketika *tokenizer* menggunakan TF-IDF untuk mengubah sebuah data menjadi sebuah *array* yang berisikan angka seperti berikut:

Gambar 4. Data yang diubah menjadi sebuah *array*

Data yang sudah berbentuk *array* digunakan sebagai sebuah input untuk algoritme TF-IDF dan diubah menjadi sebuah matriks:



Gambar 5. Data yang sudah menjadi matriks

Data yang sudah dibuat akan dimasukan kedalam Algoritme *Logistic Regression* untuk diprediksi.

V. PENUTUP

A. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, model yang dibuat dengan algoritme *LogisticRegression* dapat memprediksi relevansi *tweet* dengan *hashtag* yang dimilikinya dengan tingkat keakuratan diatas 70%.

B. Saran

Untuk penelitian selanjutnya diharapkan dapat menggunakan sumber data yang memiliki lebih dari 3000 data dan dibuat dengan lebih dari dua bahasa agar dapat melihat perbedaannya dan mengukur sebagaimana sebuah fitur *hashtag* dapat digunakan dengan sesuai, dan kemungkinan perkiraan akan semakin akurat.

DAFTAR PUSTAKA

[1] V. N. Aini and A. Alamsyah, "Analisis pada peringkat top brand menggunakan jejaring sosial percakapan dengan social network analysis (Studi kasus pada smartphone Samsung , Blackberry , Nokia , Iphone di Indonesia)," *e-Proceeding Manag.*, vol. 3, no. 1, pp. 77–85,2016.
[2] M. N. Ardhiansyah, R. Umar, and Sunardi, "Analisis sentimen pada Twitter menggunakan metode support vector machine," *Semin. Nas.Teknol. Fak. Tek. Univ. Krisnadwipayana*, vol. 1,no. 1, pp. 739–742, 2019.
[3] Hartanto, "Text mining dan sentimen analisis Twitter pada gerakan LGBT," *Intuisi J. Psikol.Ilm.*, vol. 9, no. 1, pp. 18–25, 2017.
[4] A. A. Budiman and S. Widiksono, "Aplikasi pengolahan data untuk menganalisa penggunaan hashtag pada Twitter," *J. Gerbang*, vol. 8, no. 2, 2018.
[5] L. C. Dewi, Meiliana, and A. Chandra, "Social media web scraping using social media developers API and Regex," *Procedia Comput.Sci.*, vol. 157, pp. 444–449, 2019.
[6] I. Van Der Schalk, Z. A. Koesoemahardja, and S. Jansen, "The usefulness of Twitter for open source developers as a feedback tool for the success of their projects," in *IWSECO@ ICIS*, 2016, pp. 25–38.
[7] Suprianto, Sunardi, and A. Fadlil, "Aplikasi sistem temu kembali angket mahasiswa menggunakan application of information retrieval for opinion student," *J. Teknol. Inf. danIlmu Komput.*, vol. 6, no. 1, pp. 33–40, 2019.
[8] P.KharismaditaandF.Rahutomo, "Implementasi tokenizing plus pada sistem pendeteksi keiripan jurnal skripsi," *J. Inform.Polinema*, vol. 2, no. 1, pp. 24–28, 2015.
[9] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using Term Frequency-Inverse Document Frequency (TF-IDF)," *Procedia Comput. Sci.*, vol. 7, no. 4, p. 285, 2016.
[10] M. Kumari, A. Jain, and A. Bhatia, "Synonyms based term weighting scheme: An extension to TF.IDF," *Procedia Comput. Sci.*, vol. 89, pp. 555–561, 2016.Z. Ye, A. P. Tafti, K. Y. He, K. Wang, and M. M.
[11] He, "SparkText: Biomedical text mining on big data framework," *PLoS One*, vol. 11, no. 9, p. e0162721, 2016.
[12] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn, "Distributionally robust logistic regression," *Adv. Neural Inf. Process. Syst.*, pp. 1576–1584, 2015.
[13] M. A. Mansournia, A. Geroldinger, S. Greenland, and G. Heinze, "Separation in Logistic Regression: Causes, Consequences, and Control," *Am. J. Epidemiol.*, vol. 187, no. 4, pp. 864–870, 2018.
[14] Z. Zhang, "Model building strategy for logistic regression: Purposeful selection," *Ann. Transl.Med.*, vol. 4, no. 6, pp. 4–10, 2016.
[15] L. Lombardo, M. Cama, C. Conoscenti, M. Märker, and E. Rotigliano, "Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy)," *Nat. Hazards*, vol. 79, no. 3, pp. 1621–1648, 2015.
[16] A. Safitri, Sudarmin, and M. Nusrang, "Model regresi logistik biner pada tingkat pengangguran terbuka di Provinsi Sulawesi Barat tahun 2017," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 1,no. 1, 2019.
[17] C. Sammut and G. I. Webb, Eds., *Encyclopediaof Machine Learning and Data Mining*, 2nd ed.New York: Springer Publishing Company Incorporated, 2017.
[18] S. Santos *et al.*, "The Mendeleev-Meyer force project," *Nanoscale*, vol. 8, no. 40, pp. 17400– 17406, 2016.