KLASIFIKASI BENCANA ALAM PADA TWITTER MENGGUNAKAN *NAÏVE BAYES, SUPPORT VECTOR MACHINE* DAN *LOGISTIC REGRESSION*

Kharisma Wiati Gusti
Prodi S1 Informatika, Fakultas Ilmu Komputer,
Universitas Pembangunan Nasional Veteran Jakarta
email: kharismawiatigusti@upnvj.ac.id

Informasi Artikel:
Submit: 16-06-2023; Accepted: 07-10-2023; Published: 10-10-2023
Doi : http://dx.doi.org/10.31602/tji.v14i4.11614

Abstrak

Penggunaan hashtag pada pelaporan bencana dapat membantu pihak yang berkepentingan seperti pemerintah atau lembaga penanggulangan bencana dalam menangani bencana. Akan tetapi pemberian hashtag oleh pengguna seringkali tidak sesuai dengan isi tweet. Sehingga perlu untuk dilakukan klasifikasi twitter ke dalam kategori darurat, non darurat, dan tidak relevan. Hal tersebut dilakukan untuk mempermudah lembaga terkait untuk melakukan koordinasi, pemantauan dan tanggap darurat dalam penanggulangan bencana. Penelitian ini melakukan perbandingan klasifikasi menggunakan metode Naïve Bayes, Support Vector Machine, dan Logistic Regression. Ekstraksi fitur menggunakan pembobotan term TF IDF. Guna mengatasi ketidakseimbangan kelas pada dataset dilakukan teknik untuk mensintesis sampel baru menggunakan metode Synthetic Minority Over-sampling Technique (SMOTE). Hasil penelitian menunjukkan klasifikasi menggunakan metode Logistic Regression memberikan hasil terbaik yaitu akurasi sebesar 92,4%.

Keywords: *Klasifikasi, Twitter, Bencana, Naïve Bayes, Support Vector Machine, Logistic Regression*

1. PENDAHULUAN

Media sosial khususnya twitter telah menjadi *platform* yang sering digunakan untuk berbagi informasi, terutama mengenai bencana. Pelaporan bencana oleh masyarakat di media sosial dilakukan untuk berbagi informasi mengenai bencana yang terjadi di sekitar mereka. Informasi ini digunakan untuk melakukan penanganan dan tanggap darurat terhadap bencana yang terjadi. Salah satu cara untuk memanfaatkan dan mengorganisir informasi dari twitter adalah dengan melakukan analisis hashtag yang diberikan pada setiap unggahan di twitter. Akan tetapi sering kali hashtag yang digunakan tidak relevan dengan isinya. Sehingga penting untuk melakukan klasifikasi mengenai hashtag yang relevan dan tidak relevan dengan bencana. Klasifkasi hashtag dapat membantu pihak terkait seperti pemerintah atau lembaga penanganan bencana untuk mendapatkan informasi yang lebih akurat. Dari informasi tersebut selanjutnya dapat dilakukan pemantauan secara *realtime*, koordinasi dan pengiriman bantuan yang dibutuhkan oleh masyarakat yang terkena dampak bencana.

Penelitian sebelumnya mengenai bencana banyak melakukan analisis sentiment terhadap penanggulangan bencana di Indonesia dilakukan oleh [1], pada penelitian ini dilakukan klasifikasi terhadap twitter dan menghasilkan klasifikasi positif, netral atau negatif menggunakan *TextBob*. Sementara penelitian lain mengenai analisis sentiment juga dilakukan untuk mengetahui respon masyarakat megenai penanganan banjir di Jawa Barat menggunakan Jaringan Saraf Tiruan (JST) model *Multi Layer Perceptron* (MLP), dengan hasil akurasi 73,83% [2]. Analisis sentiment juga dilakukan untuk melihat opini publik dari Twitter mengenai bencana alam di

p-ISSN: 2086-6917
e-ISSN: 2656-8047

Kalimantan Selatan menggunakan Metode *Naïve Bayes* [3]. Terdapat juga beberapa sistem yang dibuat untuk penanganan bencana alam diantaranya pembuatan sistem informasi monitoring bencana [4], sistem menampilkan peta wilayah Indonesia dan titik terjadinya bencana berdasarkan *geolocation* pada data tweet. Penelitian ini menggunakan metode *naïve bayes* dan menghasilkan akurasi sebesar 75%. Sistem peringatan *realtime* untuk kebakaran telah dibuat oleh [5] dengan menampilkan peta geografis kebakaran di kota Jakarta menggunakan SVM dengan akurasi 89%. Penelitian lain [6] dibuat untuk pembangkitan hashtag otomatis menggunakan standar OCHA untuk memudahkan pengguna memberikan laporan bencana dengan hashtag yang sesuai dengan standar (*Office for Coordination of Humanitarian Affairs (OCHA)*. Pembangkitan hashtag mendapatkan hasil dengan rata-rata *recall* 61.2%, *precission* 87.4% dan *f-measure* 66.9%.

Beberapa metode juga telah digunakan untuk melakukan klasifikasi seperti Metode *Multiclass SVM* untuk klasifikasi pesan bencana banjir, Hasil eksperimen telah menunjukkan pendekatan OVA pada algoritma SVM dengan kernel RBF menghasilkan nilai performa paling tinggi sebesar 87.03% [7]. Pada penelitian lain Algoritma *Random Forest* digunakan untuk klasifikasi *Buzzer* atau Bot di twitter [8], penelitian ini menghasilkan nilai akurasi sebesar 98%.

Berdasarkan hal tersebut maka perlu dilakukan klasifikasi twitter bencana berdasarkan hashtag yang digunakan. Twitter diklasifikasikan ke dalam kategori darurat, non darurat dan tidak relevan. Penelitian membandingkan klasifikasi hashtag menggunakan metode *Naïve Bayes*, *Support Vector Machine* dan *Linear Regression*. Penelitian ini dibuat untuk memudahkan tim tanggap darurat untuk mendapatkan informasi bencana di Indonesia berdasarkan laporan di twitter dari masyarakat. *Tweet* yang masuk ke dalam kategori darurat akan menjadi prioritas penanganan tanggap darurat bencana, sehingga dapat dilakukan pemantauan dan penanganan yang cepat terhadap bencana yang terjadi.

2. METODE PENELITIAN

Dalam melakukan penelitian, berikut alur penelitian yang dilakukan secara bertahap seperti yang digambarkan dalam gambar berikut:

Gambar 1. Metode Penelitian

3. HASIL DAN PEMBAHASAN

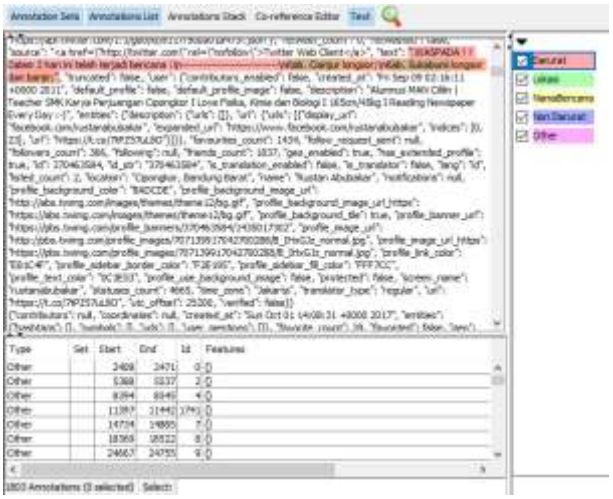
3.1 Pengumpulan Dataset

Proses pertama yang dilakukan proses pengumpulan dataset yang akan menjadi bahan penelitian. Data yang diambil merupakan data tweet yang terdapat dalam Twitter menggunakan koneksi untuk mengakses API Twitter. Dataset dikumpulkan dengan melakukan *crawling* data di twitter menggunakan program. Hasil *crawling* didapatkan dataset dengan jumlah 2.685 tweet dengan kata kunci dari hasthtag dan berbagai nama bencana seperti banjir, longsor, tsunami, gempa dan kebakaran.

3.2 Pelabelan

Tahap selanjutnya, data *tweet* yang didapatkan dari hasil *crawling* dilakukan antotasi manual oleh *anotator* menggunakan aplikasi GATE. Setiap *tweet* direpresentasikan sebagai satu dokumen, dan dilakukan anotasi per *tweet*. Sebanyak 2.685 tweet dilakukan anotasi dengan hasil terdiri dari 183 tweet darurat, 346 tweet non darurat, dan 2156 tweet yang tidak relevan dengan bencana.

p-ISSN: 2086-6917
e-ISSN: 2656-8047



Gambar 2. Pelabelan *Tweet*

3.3 Praproses

Pada tahap ini dilakukan beberapa proses:

1. *Case folding* digunakan untuk menyeragamkan karakter dalam kata dengan cara mengubah semua teks menjadi huruf kecil.
2. *URL removal* dilakukan untuk menghilangkan URL.
3. *Convert emoticon* untuk mengubah simbol *emoticon* ke dalam text.
4. *Convert Number* untuk menghapus angka yang tidak dibutuhkan atua mengubah angka menjadi kata.
5. *Punctuation removal* untuk menghilangkan tanda baca dan simbol dalam dataset, sehingga semua karakter non alphabet dihapus.
6. Tokenisasi untuk memisahkan kalimat menjadi kata pertoken atau bagian tertentu, dalam hal ini kalimat dari tweet dibagi menjadi bagian kata. Yang menjadi acuan pemisahan adalah spasi dan tanda baca.
7. Normalisasi untuk merubah kata tidak baku menjadi kata baku sehingga terdapat keseragaman kata. Terdapat kamus bahasa yang digunakan untuk melakukan proses normalisasi.
8. *Stopword removal* digunakan untuk menghapus kata-kata yang bersifat umum dan tidak terlalu penting dalam ekstraksi informasi berdasarkan hasil tokenisasi. Proses *filtering* untuk memilih *tweet* yang lengkap dan sesuai dengan kategori.

Sampel Data Sebelum Praproses
Longsor di Desa #Blongko , Jalur Trans Sulawesi Tertutup Tak Bisa Dilewati . #ongsor sepanjang 400 kilometer Sabtu (1? https://t.co/JwC9D8xnKe
Jalan Raya Porong Banjir Lagi ... Banjir Lagi ..https://t.co/96J7HC1p01 https://t.co/9pxfNtHh6L

Gambar 3 Sampel Data Sebelum Praproses

Sampel Data Setelah *Case Folding*
longsor di desa #blongko , jalur trans sulawesi tertutup tak bisa dilewati . #ongsor sepanjang 400 kilometer sabtu (1? https://t.co/jwc9d8xnke
jalan raya porong banjir lagi ... banjir lagi .. https://t.co/96j7hc1p01 https://t.co/9pxfnthh6l

Gambar 4 Setelah *Case Folding*

Sampel Data Setelah *URL Removal*
longsor di desa #blongko , jalur trans sulawesi tertutup tak bisa dilewati . #ongsor sepanjang 400 kilometer sabtu (1?
jalan raya porong banjir lagi ... banjir lagi ..

Gambar 5 Setelah *URL Removal*

Sampel Data Setelah *Convert Number*
longsor di desa #blongko , jalur trans sulawesi tertutup tak bisa dilewati . #ongsor sepanjang empat ratus kilometer sabtu (?
jalan raya porong banjir lagi ... banjir lagi ..

Gambar 6 Setelah *Convert Number*

Sampel Data Setelah *Punctuation Removal*
longsor di desa blongko jalur trans sulawesi tertutup tak bisa dilewati ongsor sepanjang empat ratus kilometer sabtu
jalan raya porong banjir lagi banjir lagi

Gambar 7 Setelah *Puctuation*

Sampel Data Setelah Tokenisasi
“longsor” “di” “desa” “blongko” “jalur” “trans” “sulawesi” “tertutup” “tak” “bisa”

"dilewati" "ongsor" "sepanjang" "empat" "ratus" "kilometer" "sabtu"
"jalan" "raya" "porong" "banjir" "lagi" "banjir" "lagi"

Gambar 8 Setelah Tokenisasi

Sampel Data Setelah Normalisasi
"longsor" "di" "desa" "blongko" "jalur" "trans" "sulawesi" "tutup" "tidak" "bisa" "lewat" "sepanjang" "empat" "ratus" "kilometer" "sabtu"
"jalan" "raya" "porong" "banjir" "lagi"

Gambar 9 Setelah Normalisasi

Sampel Data Setelah *Stopword Removal*
"longsor" "blongko" "trans" "sulawesi"
"porong" "banjir"

Gambar 10 Setelah *Stopword Removal*

Setelah dilakukan praproses dan *filtering*, dari 2.685 *tweet* didapatkan dataset *clean* sebanyak 1309 *tweet* unik.

3.4 Ekstraksi Fitur

Tahap selanjutnya adalah ekstraksi fitur untuk tahap klasifikasi tweet menggunakan proses pembobotan TF-IDF. Pembobotan TF-IDF digunakan untuk mengevaluasi seberapa penting dan sering sebuah kata muncul dalam dokumen atau dalam sekelompok kata.

3.5 Klasifikasi Tweet

Dari 1309 tweet bencana, didapatkan 89 tweet darurat, 168 tweet non darurat, dan 1052 tweet yang tidak relevan. Data dibagi menjadi data latih 70% dan data uji 30%.

Tabel 1. Data Latih dan Uji

Kategori	**Data Training**	**Data Testing**
Darurat	62	27
Non Darurat	118	50
Tidak Relevan	736	316
Total	**916**	**393**

Untuk mengatasi *imbalanced dataset* digunakan metode SMOTE *(Synthetic Minority Over-sampling Technique)*. Teknik ini menambahkan data yang dibangkitkan sebagai sampel baru dari kelas minoritas, untuk menyeimbangkan dataset dengan melakukan sampling ulang terhadap sampel kelas minoritas. Berikut adalah hasil dataset setelah dilakukan metode SMOTE:

Tabel 2. Data setelah SMOTE

Kategori	**Data Training**	**Data Testing**
Darurat	725	27
Non Darurat	731	50
Tidak Relevan	736	316
Total	**2.192**	**393**

Tahap selanjutnya melakukan klasifikasi *tweet* dibagi ke dalam tiga kategori:

1. Darurat
2. Non Darurat
3. Tidak Relevan

Klasifikasi dilakukan dengan membandingkan metode *Naïve Bayes*, *Support Vector Machine,* dan *Logistic Regression.* Berikut adalah hasil klasifikasi tweet bencana menggunakan beberapa metode:

Tabel 3. Hasil Penelitian

No	***Classifier***	***TF IDF***	***SMOTE TF IDF***
1.	***Naïve Bayes***	**76.547 %**	**91.8549 %**
	Darurat	0.414	0.955
	Non Darurat	0.407	0.912
	Tidak Relevan	0.865	0.893
2.	***SVM***	**81.8946 %**	**91.8231 %**
	Darurat	0.000	0.956
	Non Darurat	0.212	0.913
	Tidak Relevan	0.899	0.890
3.	***Logistic Regression***	**75.6303 %**	**92.4276 %**
	Darurat	0.326	0.957
	Non Darurat	0.434	0.915
	Tidak Relevan	0.856	0.901

Berdasarkan klasifikasi yang telah dilakukan menggunakan tiga metode didapatkan hasil bahwa tanpa menggunakan algoritma SMOTE didapatkan *imbalanced dataset.* Sedangkan eksperimen dengan metode SMOTE dapat mengatasi *imbalance dataset* dan mendapatkan akurasi 91,85

% dengan menggunakan *Naïve Bayes*, 91,82% hasil akurasi dengan metode SVM, dan hasil klasifikasi menggunakan *Logistic Regression* mendapatkan akurasi sebesar 92,4%.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, maka didapatkan kesimpulan sebagai berikut:

1. *Imbalanace dataset* dapat diatasi meggunakan metode SMOTE *(Synthetic Minority Over-sampling Technique)*
2. Berdasarkan perbandingkan beberapa metode yang dilakukan, didapatkan hasil terbaik dari klasifikasi menggunakan metode *Logistic Regression* yang mendapatkan akurasi sebesar 92,4%.

5. REFERENSI

[1] E. Nofiyanti dan E. M. Oki Nur Haryanto, "Analisis Sentimen terhadap Penanggulangan Bencana di Indonesia," *Jurnal Ilmiah SINUS*, vol. 19, no. 2, hlm. 17, Jul 2021, doi: 10.30646/sinus.v19i2.563.

[2] A. Layalia Safara Az-Zahra Gunawan dan K. Muslim Lhaksamana, "Analisis Sentimen pada Media Sosial Twitter terhadap Penanganan Bencana Banjir di Jawa Barat dengan Metode Jaringan Saraf Tiruan Sentiment Analysis On Twitter Social Media On Flood Disaster Management In West Java With Neural Network Method."

[3] A. M. Maksun, Y. A. Sari, dan B. Rahayudi, "Analisis Sentimen pada Twitter Bencana Alam di Kalimantan Selatan menggunakan Metode Naïve Bayes," 2021. [Daring]. Tersedia pada: http://j-ptiik.ub.ac.id

[4] E. Ananda Tasya, R. E. Saputra, dan C. Setianingsih, "SISTEM INFORMASI MONITORING BENCANA ALAM MENGGUNAKAN DATA MEDIA SOSIAL DENGAN ALGORITMA NAÏVE BAYES NATURAL DISASTER MONITORING INFORMATION SYSTEM FROM SOCIAL MEDIA DATA USING NAÏVE BAYES ALGORITHM." [Daring]. Tersedia pada: https://t.co/GhZJxNUUmT

[5] F. W. Budhi dan M. Y. G. Prasidana, "SISTEM PERINGATAN REAL-TIME BERBASIS TWITTER UNTUK BENCANA KEBAKARAN DI KOTA JAKARTA," *Jurnal Riset Jakarta*, vol. 13, no. 2, Des 2020, doi: 10.37439/jurnaldrd.v13i2.41.

[6] K. Wiati Gusti dan R. Mandala, *The 2 nd International Conference on Informatics for Development 2018 Generating of Automatic Disaster Hashtag Based on OCHA Standard*.

[7] M. Kartika Delimayanti, R. Sari, M. Laya, M. Reza Faisal, dan dan Pahrul, "Edu Komputika Journal Pemanfaatan Metode Multiclass-SVM pada Model Klasifikasi Pesan Bencana Banjir di Twitter," 2021. [Daring]. Tersedia pada: http://journal.unnes.ac.id/sju/index.php/edukom

[8] F. N. Yudianto, "Klasifikasi Hashtag Buzzer/Bot Menggunakan Algoritma Random Forest dengan Atribut Komunitas untuk Mengurangi Disinformasi Pada Twitter." [Daring]. Tersedia pada: www.trends24.in/indonesia/.