Perbandingan Metode Support Vector Machine dan Logistic Regression untuk Klasifikasi Bencana Alam

Kharisma Wiati Gusti[1]
[1] Prodi S1 Informatika, Fakultas Ilmu Komputer,
[1] Universitas Pembangunan Nasional Veteran Jakarta
kharismawiatigusti@upnvj.ac.id[1]

Abstrak. Media sosial, terutama Twitter, telah menjadi sumber penting untuk memantau dan merespons bencana alam. Klasifikasi teks dapat membantu mengidentifikasi pesan yang terkait dengan bencana alam di twitter. Penelitian ini membandingkan kinerja dari dua metode klasifikasi: *Support Vector Machine*, dan *Logistic Regression*. Dataset yang digunakan berisi sejumlah tweet yang dikategorikan secara manual menjadi tiga kelas, yaitu darurat, non darurat dan tidak relevan. Pra-pemrosesan data dilakukan untuk membersihkan teks, menghapus tautan dan karakter tertentu, tokenisasi dan normalisasi. Selanjutnya, representasi vektor *Word2vec* digunakan untuk mengekstraksi fitur yang terkait. Guna mengatasi ketidakseimbangan kelas pada dataset dilakukan teknik untuk mensintesis sampel baru menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Model klasifikasi dilatih menggunakan dataset yang telah diolah dengan menggunakan *Support Vector Machine*, dan *Logistic Regression*. Kinerja kedua metode dievaluasi menggunakan metrik seperti akurasi, presisi, *recall*, dan *F1-score*. Hasil eksperimen menunjukkan bahwa SVM memiliki akurasi 80,41% dengan atau tanpa menggunakan metode SMOTE. Sedangkan metode *Logistic Regression* memiliki akurasi 63,36% tanpa SMOTE dan 70,74% dengan menggunakan SMOTE. Sehingga didapatkan bahwa metode *Support Vector Machine* memberikan hasil terbaik yaitu akurasi sebesar 80.41%. Penggunaan metode SMOTE tidak terlalu berpengaruh ketika menggunakan metode SVM, sedangkan dalam *Logistic Regression* pengguanaan SMOTE cukup berpengaruh, dengan memberikan kenaikan akurasi sebesar 28,26%. Penelitian ini dapat membantu dalam pengambilan keputusan darurat dan pemantauan bencana alam di media sosial.

Kata Kunci: Klasifikasi, Twitter, Bencana, 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.

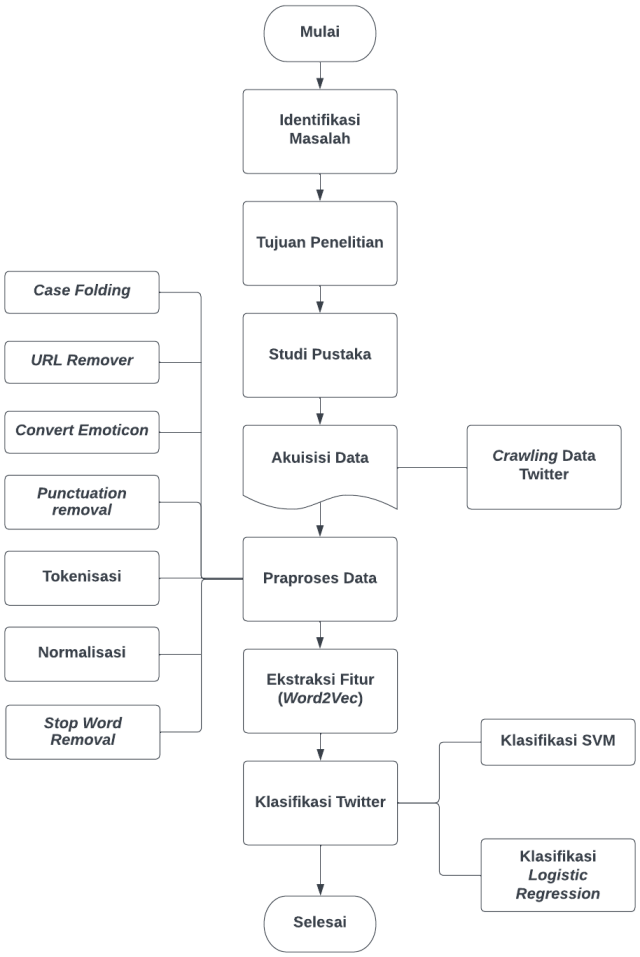
Penelitian sebelumnya mengenai bencana banyak melakukan analisis sentiment terhadap penanggulangan bencana di Indonesia dilakukan oleh [1], pada penelitian ini dilakukan klasifikasi terhadap twitter dan menghasilkan klasifikasi positif, netral atau negatif menggunakan *TextBob*. Sementara penelitian lain mengenai analisis sentiment juga dilakukan untuk mengetahui respon masyarakat megenai penanganan banjir di Jawa Barat menggunakan Jaringan Saraf Tiruan (JST) dengan model *Multi-Layer Perceptron* (MLP), dengan hasil akurasi 73,83% [2]. Analisis sentiment juga dilakukan dengan menggunakan Metode Naïve Bayes untuk melihat opini publik dari Twitter mengenai bencana alam di Kalimantan Selatan [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 menggunakan metode naïve bayes berdasarkan data lokasi tweet 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 mendapatkan hasil performa 87.03% menggunakan algoritma SVM dengan kernel RBF [7]. Pada penelitian lain Algoritma *Random Forest* digunakan untuk klasifikasi Buzzer atau Bot di twitter [8], penelitian ini menghasilkan nilai akurasi sebesar 98%. Perbandingan anatara metode *Decision Tree*, *Random Forest*, dan SVM dilakukan untuk klasifikasi tweet yang mengandung informasi gempa atau tidak. Hasilnya akurasi SVM dengan recall 86.3% dan presicion 88.7%, secara keseluruhan lebih baik dibandingkan *Decision tree* dan *Random Forest* [9]. Penelitian lain mencoba membandingkan metode SVM dan *Naïve Bayes* untuk mengklasifikasikan tweet *cyberbullying*. Hasilnya klasifikasi menggunakan SVM mendapatkan akurasi 99,60%, sedangkan dengan metode *Naïve bayes* mendapatkan akurasi sebesar 97.99% [10].

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 *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 Metodologi Penelitian

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



Gambar. 1. Metode Penelitian

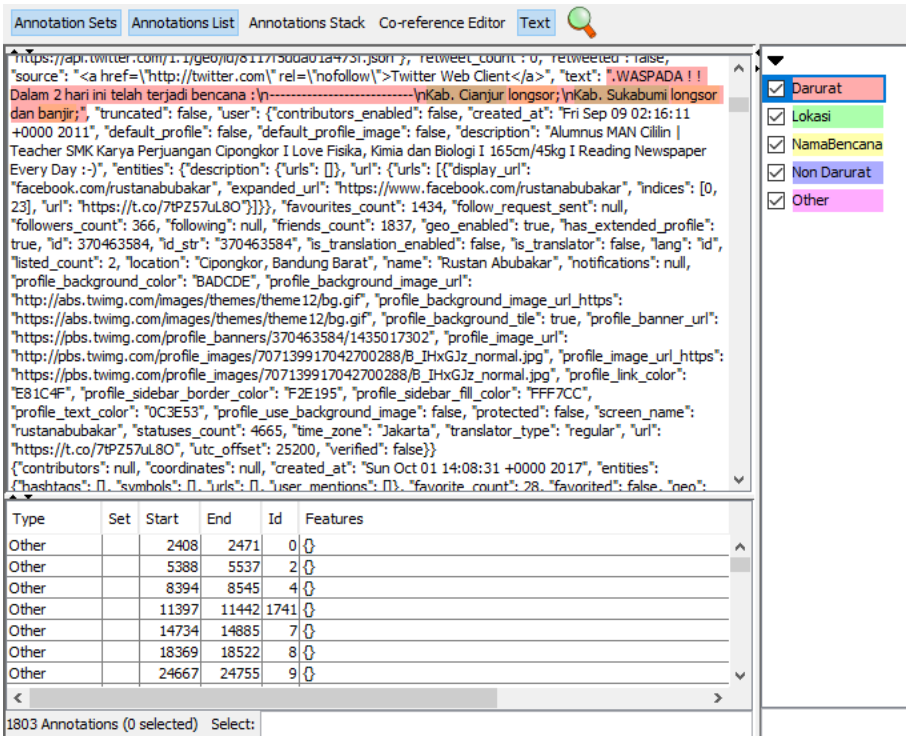
3 Hasil dan Pembahasan

3.1 Pengumpulan Dataset

Proses pertama yang dilakukan proses pengumpulan dataset yang akan menjadi bahan penelitian. Data yang diambil adalah data tweet yang terdapat dalam Twitter dengan 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.



Gambar. 2. Pelabelan *Tweet*

3.3 Pelabelan

Berikut contoh sampel data sebelum dilakukan praproses:

Tabel 1. Sampel Data Sebelum Praproses

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

Pada tahap ini dilakukan beberapa proses:

1. *Case folding* digunakan untuk menyeragamkan karakter dalam kata dengan mengubah seluruh teks menjadi huruf kecil.

Tabel 2. Setelah *Case Folding*

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

2. *URL removal* dilakukan untuk menghilangkan URL.

Tabel 3. Setelah *URL Removal*

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 ..

3. *Convert emoticon* untuk mengubah simbol *emoticon* ke dalam text.
4. *Convert Number* untuk menghapus angka yang tidak dibutuhkan atua mengubah angka menjadi kata.

Tabel 4. Setelah *Convert Number*

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 ..

5. *Punctuation removal* untuk menghilangkan tanda baca dan simbol dalam dataset, sehingga semua karakter non alphabet dihapus.

Tabel 5. Setelah *Puctuation Removal*

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

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.

Tabel 6. Setelah Tokenisasi

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"

7. Normalisasi untuk merubah kata tidak baku menjadi kata baku sehingga terdapat keseragaman kata. Terdapat kamus bahasa yang digunakan untuk melakukan proses normalisasi.

Tabel 7. Setelah Normalisasi

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"

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.

Tabel 8. Setelah *Stopword Removal*

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

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 *Word2Vec* dengan pendekatan *Skip-gram*. *Word2Vec* adalah teknik pemrosesan bahasa alami (*Natural Language Processing/NLP*) yang digunakan untuk membuat representasi vektor kata. Representasi ini berguna sebagai fitur ekstraksi dalazm berbagai tugas pemrosesan bahasa alami, termasuk klasifikasi teks.

Metode *Skip-gram* menggunakan model *Word2Vec* untuk memprediksi kata-kata konteks berdasarkan kata target. Langkah-langkahnya mirip dengan CBOW, tetapi fungsi kata target dan kata konteks berbeda. Pelatihan model akan menekankan pemahaman kata-kata konteks yang berbeda yang mungkin muncul untuk setiap kata target.

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%.

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

Tabel 9. 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 10. 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 Support Vector Machine, dan Logistic Regression. Berikut adalah hasil klasifikasi tweet bencana menggunakan beberapa metode:

Tabel 11. Hasil Penelitian

Non SMOTE		SMOTE	
SVM	**80.4071 %**	**SVM**	**80.4071 %**
Logistic Regression	**63.3588 %**	***Logistic Regression***	**70.7379 %**

Berdasarkan klasifikasi yang telah dilakukan menggunakan dua metode didapatkan hasil: penelitian menggunakan dataset asli memiliki akurasi sebesar 80.41% menggunakan metode SVM dan 63.36% menggunakan Logistic Regression. Sedangkan eksperimen dengan metode SMOTE mendapatkan akurasi 80.41% dengan metode SVM, dan hasil klasifikasi menggunakan Logistic Regression mendapatkan akurasi sebesar 70.74%.

4 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, maka didapatkan kesimpulan sebagai berikut:
1. Hasil eksperimen menunjukkan bahwa SVM memiliki akurasi 80,41% dengan atau tanpa menggunakan metode SMOTE. Sedangkan metode Logistic Regression memiliki akurasi 63,36% tanpa SMOTE dan 70,74% dengan menggunakan SMOTE. Sehingga didapatkan kesimpulan bahwa metode Support Vector Machine memberikan hasil terbaik yaitu akurasi sebesar 80.41%.
2. Pengguanan metode SMOTE tidak terlalu berpengaruh ketika menggunakan metode SVM, sedangkan dalam Logistic Regression penggunaan SMOTE cukup berpengaruh, dengan memberikan kenaikan sebesar 28,26%.

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

Referensi

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

[2] A. Layalia Safara Az-Zahra Gunawan and 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, and B. Rahayudi, "Analisis Sentimen pada Twitter Bencana Alam di Kalimantan Selatan menggunakan Metode Naïve Bayes," 2021. [Online]. Available: http://j-ptiik.ub.ac.id

[4] E. Ananda Tasya, R. E. Saputra, and 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." [Online]. Available: https://t.co/GhZJxNUUmT

[5] F. W. Budhi and M. Y. G. Prasidana, "Sistem Peringatan Real-Time Berbasis Twitter Untuk Bencana Kebakaran Di Kota Jakarta," *Jurnal Riset Jakarta*, vol. 13, no. 2, Dec. 2020, doi: 10.37439/jurnaldrd.v13i2.41.

[6] K. W. Gusti dan R. Mandala, "Generating of Automatic Disaster Hashtag Based on Ocha Standard," dalam ICID Proceedings 2018, hal. C1 04, Informatics Departemen, UIN Sunan Kalijaga Yogyakarta, 10-12 November 2018, ISBN 978-602-53524-0-9.

[7] M. Kartika Delimayanti, R. Sari, M. Laya, M. Reza Faisal, and dan Pahrul, "Edu Komputika Journal Pemanfaatan Metode Multiclass-SVM pada Model Klasifikasi Pesan Bencana Banjir di Twitter," 2021. [Online]. Available: 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." [Online]. Available: www.trends24.in/indonesia/.

[9] Y. Imanuela Claudy, R. S. Perdana, dan M. A. Fauzi, "Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN)," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 8, hlm. 2761-2765, Agustus 2018. [Online]. Tersedia: http://j-ptiik.ub.ac.id

[10] N. Chamidah and R. Sahawaly, "Comparison Support Vector Machine and Naive Bayes Methods for Classifying Cyberbullying in Twitter," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 7, no. 2, p. 338, Sep. 2021, doi: 10.26555/jiteki.v7i2.21175.