PENGARUH TEKNIK KLASIFIKASI PADA PESAN BENCANA BANJIR DI TWITTER DENGAN METODE MULTICLASS- SVM

Mera Kartika Delimayanti[1)], Risna Sari[1)], Mauldy Laya[1)], M. Reza Faisal[2)], Pahrul[2)]

[1]Jurusan Teknik Informatika dan Komputer, Politeknik Negeri Jakarta,
Jl. Prof GA Siwabessy Kampus UI, Depok, 16425.
[2]Program Studi Ilmu Komputer, Universitas Lambung Mangkurat,
Jl. A. Yani KM 36, Banjarbaru, 70714
E-mail: mera.kartika@tik.pnj.ac.id

Abstract

Eyewitness message classification on flood disaster on twitter aims to find out the resulting classification performance. In the classification of flood disaster messages, preprocessing data is required before the classification process is carried out. Preprocessing affects the resulting level of accuracy in the classification process. Support Vector Machine (SVM) is used as a method for classification by weighting words using Term Frequency-Inverse Document Frequency (TF-IDF). The data taken from Twitter is 3000 data with 1000 data each for each label. The OVO and OVA approaches have been carried out in this experiment by applying the multiclass Support Vector Machine method. Based on the experiments that have been conducted, the OVA approach with the RBF kernel provides the highest accuracy value for the classification of flood disaster messages on Twitter at 87.03% compared to previous research and methods which reached 77.87%.

Keywords: *classification, twitter, flood disaster, multiclass, support vector machine*

Abstrak

Klasifikasi pesan saksi mata pada bencana banjir di twitter bertujuan untuk mengetahui kinerja klasifikasi dihasilkan. Pada klasifikasi pesan bencana banjir dibutuhkan *preprocessing* data sebelum dilakukannya proses klasifikasi. Preprocessing berpengaruh terhadap tingkat akurasi yang dihasilkan pada proses klasifikasi. Support Vector Machine (SVM) digunakan sebagai metode untuk klasifikasi dengan pembobotan kata menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). Data yang diambil dari twitter berjumlah 3000 data dengan masing-masing 1000 data setiap labelnya. Pendekatan OVO dan OVA telah dilakukan pada eksperimen ini dengan menerapkan metode *multiclass Support Vector Machine*. Berdasarkan eksperimen yang telah dilakukan, pendekatan OVA dengan kernel RBF memberikan nilai akurasi yang paling tinggi untuk klasifikasi pesan bencana banjir di Twitter sebesar 87.03% dibandingkan riset dan metode sebelumnya yang mencapai 77.87%.

Kata Kunci: *klasifikasi, twitter, bencana banjir, multiclass, support vector machine*

PENDAHULUAN

Bencana banjir menjadi salah satu bencana rutin tahunan, menurut Badan Nasional Penanggulangan Bencana tercatat telah terjadi 111 kejadian banjir di Indonesia sepanjang bulan januari 2020(BNPB, 2008). Banjir juga dapat menimbulkan dampak terhadap masyarakat misalnya pada banjir bandang, jalan, jembatan, pertanian, rumah

dan mobil banyak yang hancur. Akibat dari dampak bencana banjir maka orang mencari informasi kapan dan dimana terjadinya bencana banjir tersebut. Pada beberapa kasus bencana alam, sosial media menjadi media informasi yang lebih cepat daripada banyak informasi yang bersumber dari berita seperti televisi dan radio.

Data berupa informasi diambil dari media sosial berupa tweet yang berasal dari Twitter. Data dari media sosial juga menjadi sumber informasi yang berguna untuk menanggapi bencana banjir. Pada penelitian(Wu, 2016) mengusulkan konsep baru untuk analisis media sosial yang disebut Sosial Sensor. Media sosial menjadi sumber informasi yang berasal dari banyak warga lokal, para pengamat dan saksi mata. Informasi dari laporan saksi mata lebih disukai dari sumber informasi lain (misalnya orang di luar daerah bencana). Lembaga penegak hukum dan responden pertama selalu mencari tangan pertama (saksi mata) dan informasi yang dapat dipercaya. Pada penelitian Zahra (Zahra et al., 2020) dilakukan untuk memahami berbagai jenis laporan saksi mata yang terdiri dari tiga kelas (i) Saksi mata, (ii) non-saksi mata, dan (iii) Tidak tahu.

Penelitian tersebut melakukan klasifikasi pesan sanksi mata di twitter berdasarkan kelas yang sudah ditentukan sebanyak 3 kelas. Algoritma yang digunakan pada penelitian ini adalah Support Vector Machine (SVM). Support Vector Machine (SVM) adalah suatu teknik yang baik dalam kasus klasifikasi maupun regresi (Pratama & Murfi, 2014). Pada penelitian (Luqyana et al., 2018), Support Vector Machine (SVM) digunakan untuk klasifikasi sentimen pada komentar instagram, digunakan ekstraksi fitur TF-IDF dan algoritma klasifikasi Support Vector Machine (SVM). Dari pengujian tersebut didapatkan hasil akurasi tertinggi sebesar 90%.Dari berbagai referensi penelitian yang telah dilakukan, Algoritma Support Vector Machine (SVM) menjadi salah satu algoritma klasifikasi yang dapaat digunakan untuk klasifikasi pesan saksi mata pada saat bencana banjir. Penelitian ini sangat berguna untuk memprediksi kejadian bencana alam dan memberikan sistem peringatan dini dari klasifikasi data dari pesan twitter khususnya untuk bencana banjir yang terjadi di Indonesia.

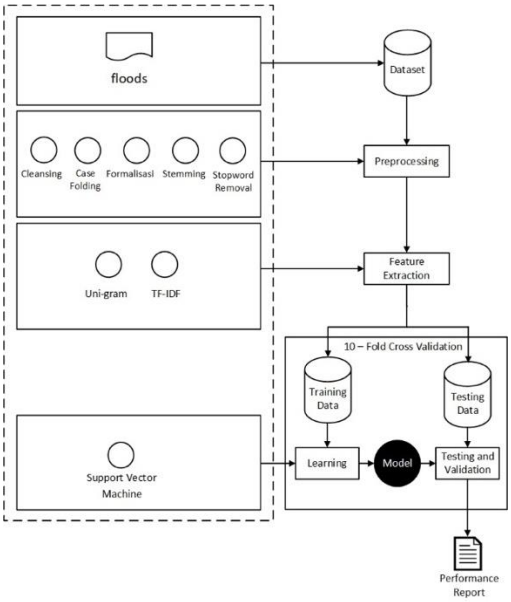
METODE PENELITIAN

Social Network Sensor (SNS) adalah sebuah konsep baru yang berasal dari sensor fisik. Sensor fisik mengukur dan menghasilkan data fisik atau kimia, seperti cahaya, panas, suhu, dan kelembaban. Sementara sensor sosial mengumpulkan data dari media

sosial berupa tweet, pengguna, lokasi, dan lain-lain. (Wu, 2016). Media sosial telah terbukti memberikan dampak dalam mempercepat penyebaran kabar atau berita. Pengguna sosial media mengabarkan apa yang dilihatnya seketika setelah melihat kejadian. Hal ini membuat banyak riset yang dilakukan memanfaatkan data media sosial sebagai *Social Network Sensor* atau media sosial sebagai sensor (Christakis & Fowler, 2010)(Kryvasheyeu et al., 2015) dan pemanfaatan media sosial sebagai sensor untuk peringatan dini dan monitoring setelah terjadi bencana alam (Hernandez-Suarez, Sanchez-Perez, Toscano-Medina, Perez-Meana, Portillo-Portillo, Sanchez, et al., 2019) . Bencana alam yang dapat ditangani oleh *social network sensor* diantaranya adalah gempa, banjir, angin putting beliung dan kebakaran hutan (Zahra et al., 2020). Teknik pengolahan data dari media sosial sebagai sensor adalah menggunakan teknik *text mining*.

Text mining (penambangan teks) adalah penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit, yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda(Feldman & Sanger, 2007). Text Mining cenderung mengarah pada bidang penelitian data mining. Oleh karena itu, tidak mengherankan bahwa Text Mining dan Data Mining berada pada tingkat arsitektur yang sama (Falahah & Dwiki Adriadi Nur, 2015). Langkah-langkah penelitian seperti yang ditunjukkan dalam gambar 1 yang meliputi data preprocessing yakni berupa pembersihan data /*data cleansing, case folding*, formalisasi, stemming, dan stopword removal. Berikutnya adalah *feature selection* dan dilanjutkan dengan model data dan melakukan klasifikasi dengan algoritma SVM dengan berbagai parameter yang diujicobakan.

Dataset yang digunakan pada penelitian ini adalah data tweets tentang bencana banjir yang diambil dari Twitter. Proses pengumpulan data tweets yang diperlukan menggunakan package Twitter Scraper yang merupakan salah satu *package* Python. Agar dapat melakukan scraping data dari Twitter. Jumlah data yang digunakan ada 3000 data. Tabel 1 menunjukkan contoh data dari dataset yang digunakan.



Gambar 1. Tahapan Proses Penelitian

Tabel 1.Contoh data pesan twitter dan label data

Text	Label
Depan rumah dh banjir. Huhu pic.twitter.com/KV1Bz2R977	Saksi mata
Takutnya banjir!!!!	Saksi mata
Semoga para korban banjir Diberikan ketabahan #WowoNangisJakartaBanjir	Non-saksi mata

Pada tahap penelitian diawali dengan *pre-processing* untuk mengubah data sesuai dengan format yang dibutuhkan. Tahapan dalam *pre-processing* adalah sebagai berikut :

A. Pelabelan

Pelabelan dilakukan pada data yang sudah di-scraping dari twitter. Data dibagi menjadi 3 kelas yaitu (i) Saksi mata, (ii) non-saksi mata, dan (iii) Tidak tahu. **Remove Duplicate** Proses penghapusan pada data yang sama.

B. Cleansing

Cleansing merupakan proses menghapus karakter-karakter yang tidak berkontribusi pada sentiment analysis sehingga hanya menyisakan karakter alfabet.

C. Case folding

Hal ini dilakukan dengan mengubah kata menjadi lower case atau huruf kecil.

D. Formalisasi

Proses mendeteksi dan memperbaiki atau menghapus data yang rusak atau data yang tidak akurat agar dapat menambah akurasi proses klasifikas. Sebagai contoh penggunaan kalimat alay atau bahasa gaul mengakibatkan penggunaan Bahasa Indonesia tidak baku.

E. Stemming

Stemming bertujuan untuk mentransformasikan kata menjadi kata dasarnya (*root word*) dengan menghilangkan semua imbuhan kata. Tabel 2 adalah hasil stemming data.

Tabel 2. Hasil dari Proses Stemming Data

Text	**Label**
depan rumah sudah banjir huhu	Saksi mata
takut banjir	Saksi mata
moga para korban banjir berik ketabahan	Non-saksi mata

E. Stopword Removal

Stopword Removal merupakan proses penyaringan kata yang muncul dalam jumlah besar/umum atau kata yang tidak baku dan tidak memiliki makna (*stopword*).

Tabel 3. Kamus Slang yang digunakan

Kamus 1	Kamus id.stopwords.02.01.2016.txt
Kamus 2	Kamus yang dibuat berdasarkan data banjir yang digunakan

Setelah tahap pre-processing adalah tahap pembobotan TF-IDF, 10-Cross Validation and perhitungan akurasi dengan algoritma Support Vector Machine untuk multiclass dengan parameter One Versus One (OVO) dan One Versus All (OVA).

Term Frequency-Inverse Document Frequency (TF-IVDF) adalah metode yang digunakan untuk menghitung boboVt setiap kata yang telah diekstrak. Model pembobotan TF-IDF merupakan metode yang mengintegrasikan model *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). *Term Frequency* (TF) merupakan proses untuk menghitung jumlah kemunculan term dalam satu dokumen dan Inverse Document Frequency (IDF) digunakan untuk menghitung term yang muncul di berbagai dokumen (komentar) yang dianggap sebagai term umum, yang dinilai tidak penting (Akbari et al., 2017). Tahapan pembobotan dengan TF-IDF adalah:

1. Hitung Term Frequency (tft,d)

2. Hitung Weighting Term Frequency (Wtf t,d)

$$\text{Wtf}_{t,d} = \begin{cases} 1 + \log 10\ \text{tf}_{t,d} & , \text{if tf}_{t,d} > 0 \\ 0 & , \text{if tf}_{t,d} = 0 \end{cases} \quad (1)$$

3. Hitung document frequency (df)
4. Hitung bobot Inverse Document Frequency (idf)

$$idft = log10\ \frac{\text{N}}{\text{df}} \quad (2)$$

5. Hitung nilai bobot TF-IDF

$$Wt,d = Wtft,d \times idf\text{t} \quad (3)$$

Ekstraksi fitur digunakan untuk mengubah data menjadi data yang terstrukur sehingga dapat dilakukan proses text mining. Uni-gram digunakan untuk ekstraksi fitur. Uni-gram merupakan ekstraksi fitur kata tunggal, pada proses ini data akan diekstrak menajadi kata tunggal. Pada tabel 9 dapat dilihat cara kerja uni-gram untuk memisahkan setiap dokument teks menjadi kata tunggal. Hasil ekstraksi fitur unigram dapat dilihat di tabel 4.

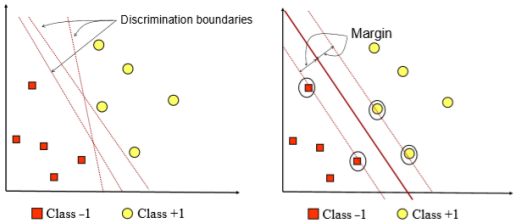
Tabel 4. Hasil dari Proses Ekstraksi Fitur Unigram

depan rumah sudah banjir	"depan","rumah","sudah", "banjir"
banjir dimanaa macet dimana	"banjir","dimanaa","macet", "dimana"
takut banjir	"takut","banjir"
moga para korban banjir berik ketabahan	"moga","para","korban", "banjir", "berik","ketabahan"

Metode k-fold cross validation membagi sampel actual menjadi k sebagai sampel yang berukuran sama. Setiap subsample diambil sebagai data validasi untuk menguji model klasifikasi dan ulangi proses sebanyak k kali. Keuntungan metode ini adalah lebih dari pengulangan sampel acak sebagai pelatihan dan validasi untuk masing-masing untuk validasi setidaknya sekali. (Raju et al., 2018).

Support Vector Machine (SVM) adalah suatu teknik yang baik dalam kasus klasifikasi maupun regresi. Support Vector Machine (SVM) memiliki prinsip dasar linear classifier yaitu kasus klasifikasi yang secara linier dapat dipisahkan, namun Support Vector Machine (SVM) telah dikembangkan agar dapat bekerja pada problem non-linear dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi. Pada ruang

berdimensi tinggi, akan dicari hyperplane yang dapat memaksimalkan jarak (margin) antara kelas data (Santosa, 2007).



(a) (b)

Gambar 2. Diagram proses SVM

Pada awalnya Support Vector Machine (SVM) dikembangkan untuk persoalan klasifikasi dua kelas, kemudian dikembangkan kembali untuk klasifikasi multiclass (Santosa, 2007). Terdapat dua teknik multi class yang sering digunakan pada SVM yaitu One Versus One (OVO) dan One Versus All (OVA). OVO adalah teknik klasifikasi yang membandingkan satu dengan satu kelas lainnya, sedangkan OVA yaitu membandingkan satu dengan semua selain dirinya yang dianggap menjadi satu kesatuan. Pada model ini digunakan SVM OVA untuk Multi class. ada penelitian ini digunakan model klasifikasi dengan package Kernlab dengan menggunakan parameter kernel Radial Basis Function (RBF). *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi dengan membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya.

HASIL DAN PEMBAHASAN

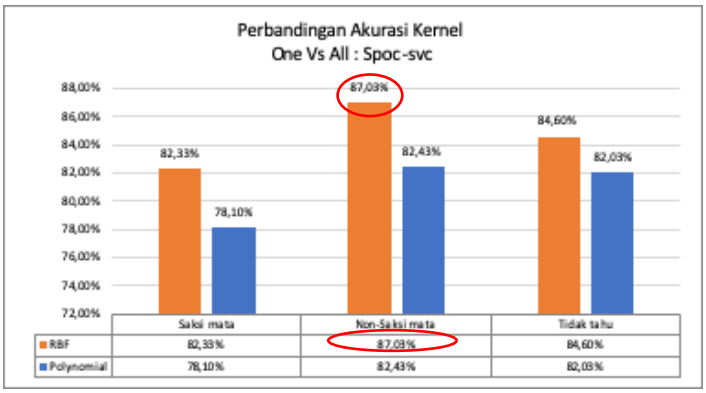
Klasifikasi pesan bencana banjir berdasarkan data tweet bencana banjir dengan menggunakan support vector machine dengan kernel Radial Basis Function (RBF). Data yang digunakan berjumlah 3000 data dengan sejumlah fitur data adalah 3230 fitur. Untuk hasil pemrosesan menggunakan algoritma SVM dengan metode One Versus One (OVO) dapat dilihat hasilnya di tabel 5 dengan nilai akurasi tertinggi mencapai 77.90% pada type Spoc-svc dengan kernel RBF.

Tabel 5.Hasil Akurasi untuk metode OVO

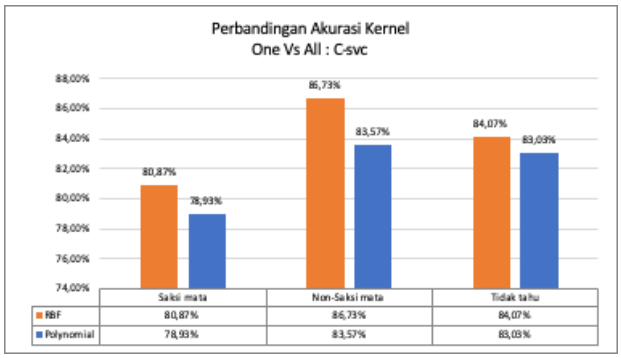
One Versus One (OVO) : Kernlab_Spoc-svc					
Label	**Type**	**Kernel**	**Parameter**	**Akurasi**	**Sensitifity**
Saksi mata	Spoc-svc	RBF	default	77,90%	79,40%
Non-Saksi mata					82,70%

Tidak tahu					71,60%
Saksi mata	Spoc-svc	Polynomial	default	72,77%	76,50%
Non-Saksi mata					75,30%
Tidak tahu					66,50%

Sedangkan untuk percobaan dengan algoritma SVM dan metode multiclass OVA didapatkan nilai akurasi yang lebih tinggi dengan berbagai tipe kernel dan tipe parameter yang digunakan. Perbandingan nilai akurasi kernel dengan metode OVA dan Spoc-svc adalah ditunjukkan pada gambar 3 dan 4.



Gambar 3. Perbandingan Nilai Akurasi Kernel untuk tiga class untuk type Spoc-svc



Gambar 4. Perbandingan Nilai Akurasi Kernel untuk tiga class untuk type C-svc

Terlihat dari gambar 3, nilai akurasi tertinggi adalah 87.03% untuk class non-saksi mata dengan algoritma SVM untuk metode kernel RBF dengan tipe Spoc-svc. Hal ini terlihat jelas bahwa metode multiclass OVA pada algoritma SVM dapat memperbaiki nilai akurasi dari informasi dari twitter untuk bencana banjir. Hal tersebut telah memperbaiki hasil riset sebelumnya yang mana menggunakan teknik stopword dan metode untuk dua class dengan nilai akurasi 77.87%.

SIMPULAN

Berdasarkan hasil eksperimen penggunaan data media sosial twitter untuk pesan bencana banjir dapat disimpulkan bahwa pada metode multiclass SVM dengan pendekata OVA dapat meningkatkan nilai akurasi. Nilai akurasi tertinggi mencapai 87.03% untuk klasifikasi pesan bencana banjir melalui twitter. Hal ini dapat terlihat dari class data yang digunakan dalam eksperimen ini adalah tiga class.

DAFTAR PUSTAKA

Akbari, M., Novianty, A., & Setianingsih, C. (2017, August). *Analisis Sentimen Menggunakan Metode Learning Vector Quantization Sentiment Analysis Using Learning Vector Quantization Method*.

BNPB. (2008). *Tanggap Tangkas Tangguh Menghadapi Bencana, Pedoman Penyusunan Rencana Penanggulangan Bencana.* BNPB.

Christakis, N. A., & Fowler, J. H. (2010). Social Network Sensors for Early Detection of Contagious Outbreaks. *PLoS ONE*, *5*(9), e12948. https://doi.org/10.1371/journal.pone.0012948

Falahah, & Dwiki Adriadi Nur, D. (2015). PENGEMBANGAN APLIKASI SENTIMENT ANALYSIS MENGGUNAKAN METODE NAÏVE BAYES (Studi Kasus Sentiment Analysis dari media Twitter). *Seminar Nasional Sistem Informasi Indonesia*, *November*, 2–3.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press. http://www.books24x7.com/marc.asp?bookid=23164

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., And Luis, V. S., & Javier Garcia Villalba, L. (2019). Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation. *Sensors (Basel, Switzerland)*, *19*(7). https://doi.org/10.3390/s19071746

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., & García Villalba, L. (2019). Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation. *Sensors*, *19*(7), 1746. https://doi.org/10.3390/s19071746

Kryvasheyeu, Y., Chen, H., Moro, E., Van Hentenryck, P., & Cebrian, M. (2015). Performance of Social Network Sensors during Hurricane Sandy. *PLOS ONE*, *10*(2), e0117288. https://doi.org/10.1371/journal.pone.0117288

Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(11), 4704–4713.

Pratama, M. L., & Murfi, H. (2014). *STUDI KOMPARASI METODE MULTICLASS SUPPORT VECTOR MACHINE UNTUK MASALAH ANALISIS SENTIMEN PADA TWITTER*. 20.

Raju, K. S., Murty, M. R., Rao, M. V., & Satapathy, S. C. (2018). Support Vector Machine with K-fold Cross Validation Model for Software Fault Prediction. *International Journal of Pure and Applied Mathematics*, *118*(20), 321–334.

Santosa, B. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*.

Wu, C.-H. (2016). SOCIAL SENSOR: AN ANALYSIS TOOL FOR SOCIAL MEDIA. *International Journal of Electronic Commerce Studies*, *7*(1), 77–94. https://doi.org/10.7903/ijecs.1411

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information Processing and Management*, *57*(1), 102107. https://doi.org/10.1016/j.ipm.2019.102107